The Author, Not the Tale

Memory, Narrative, and the Self

By

Philipp Rau

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

University of Sheffield

Department of Philosophy

June 2016

Abstract

There is a confusing diversity of conceptions of 'the self' in philosophical, psychological, psychiatric, and neuroscientific discourse. To remedy this, I propose and defend a naturalistic view of the self: *the system view*. The self is here conceived of as the complex and dynamic system of our higher-level self-monitoring functions, including our capacities for self-representation over time. These are grounded in more basic self-representational capacities that are widespread among different species. On the system view, the self is not to be confounded with the attributes of personhood, as it often has been in philosophical discourse. Nor is the self over time a product of memory, as philosophers in Locke's tradition, and some popular intuitions, seem to take it to be. I discuss the complex nature of autobiographical memory and argue that, given that much of our autobiographical remembering is already a reconstructive process, the self is not produced by our memories, but is the system that produces them. The system view is also opposed to currently fashionable views of the self as 'narrative'. Narrative constructionism about the self has an authorship problem: it does not account for the processes that enable and subserve narration about oneself in the first place. I argue that it is in these processes, rather than in their productions, that we should conceptually locate the self. Neither should we take narrative capacities to be essential for a self. To illustrate the advantages of the system view, I discuss autism spectrum conditions and other defects and disorders such as dementia, dissociative disorders, and schizophrenia. In these conditions, particular self-representational capacities are differently configured, impaired, or absent, but this does not entail a wholesale loss or lack of self. Instead, such conditions are better characterized as specific system malfunctions. I conclude by suggesting directions for future research.

Contents

List of figures

Acknowledgements

First and foremost, I am grateful to my supervisors, George Botterill and Dominic Gregory. George has accompanied my research on the self for over five years—from advising me on my first PhD proposal up to helping me polish the final draft of this thesis. Our countless conversations—covering anything from matters of antiquarian interest in Locke scholarship to current debates in cognitive science—have been of invaluable assistance: in developing my research strategy, unclouding my thinking about the issues involved, drawing the right distinctions and preventing me from blindly endorsing the wrong ones—and indeed in our occasional quibbling over the right turn of phrase. He encouraged me when things weren't going well, and gently but firmly nudged me onwards when they were. Dominic, too, has been tremendously helpful, offering a fresh and rigorous look at my drafts, issuing initially puzzling challenges on seemingly minor points that on closer inspection proved important, and making useful suggestions for clarifying and structuring my arguments. I could not have wished for a better pair of supervisors for my thesis. Such faults as remain are mine.

I thank Shaun Nichols and Mike Bruno for their ready and friendly assistance with my replication of one of their experimental studies. For their useful feedback and for our helpful and cheerful exchanges at conferences and the like, I thank Pat Churchland, John Doris, Robert Foley, Albert Newen, Galen Strawson, and Şerife Tekin.

As I prepare to hand in this thesis, my 'last piece of student work' (as my department's handbook for postgraduates reassuringly puts it), I cannot help but think back to my *first* pieces of student work—and express my gratitude to the institution that allowed me to re-enter, and advance in, higher education after a couple of false starts: the Open University. I am grateful to my OU tutor, John Shand, for his thorough and considered comments on my undergraduate work and for encouraging me to pursue postgraduate studies in philosophy, and to Keith Frankish for his advice in this respect. And I thank my fellow OU student Mike Sherwood for our correspondence, now in its eleventh year, that followed our meeting at summer school.

At Sheffield, besides my doctoral supervisors, I have benefited from the advice, support, and teaching of many other academic staff, past and present, in various formal and informal roles: Luca Barlassina, Chris Bennett, Paul Faulkner, Chris Hookway, Rob Hopkins, Rosanna Keefe, Nils Kürbis, Steve Laurence, Jimmy Lenman, Steve Makin, Eric Olson,

Komarine Romdenh-Romluc, Jenny Saul, Robin Scaife, Yonatan Shemmer, and Bob Stern. I am grateful to my department for the opportunity to teach as well as conduct research—allowing me to obey the imperatives on both pages of the open book depicted in our university's coat of arms: '*Disce. Doce*.' And I thank my students, undergraduates and MA students alike, for their queries, contributions, and challenges in lectures, seminars, and tutorials.

The academic camaraderie in the postgraduate philosophy community at Sheffield—with some welcome sprinklings of psychology—has been a great source of support and enjoyment to me. My thanks go to all my postgraduate colleagues who have made up this community during my nearly seven years here—including, but not limited to, Charlotte Alderwick, Jess Begon, Josh Black, Pete Caven, Charlie Crerar, Harry Cusworth, Ed Donnellan, Lily FitzGibbon, Josh Forstenzer, Carl Fox, Max Gattie, Trystan Goetze, Richard Hassall, Rich Healey, Stephen Ingram, Katharine Jenkins, Armin Khameh, Lizzy Kirkham, Damiano La Manna, Tash McKeever, Neri Marsili, Ashley Pennington, Angie Pepper, Kathy Puddifoot, Jack Wadham, Neil Williams, and Steve Wright. In particular, I thank Ryan Doran for our empirical philosophical kinship, our conversations about our research, his helpful observations regarding mine, and his enthusiasm—and forbearance—in our collaborations.

I gratefully acknowledge the funding I received from the University of Sheffield, the Arts and Humanities Research Council, and the Jacobsen Trust (via the Royal Institute of Philosophy).

Neither of my parents lived to see me complete this final piece of student work, but without their support, pecuniary and otherwise, my philosophical studies might not have been possible. I think they both would be pleased that a concoction of their genes had produced this doctoral thesis. I thank my brother, Matthias Rau, for our late-night telephone conversations swapping anecdotes about academic work, and much else besides. For taking an interest in my research and, more important, for their friendship and encouragement during my years of philosophical doctoring, I thank Dominic Clifton, Pierre Fickinger, Kay Karpinsky, Kathryn Kirton, Frank Speidel, Malcolm Stuart, Rosemary Stuart, and Lucy Watt. Most especially, I thank Michael J. Savage: for many dotty and instructive distractions from philosophical toil, for keeping me sane, for putting up my bookshelves—and for putting up with me all these years.

Sheffield, May 2016 P. R.

Typographical note

Throughout this thesis, omissions in quotations are indicated by three spaced dots, thus: . . . This will allow the reader to distinguish between the quoted author's own ellipses (marked in the usual way by unspaced dots: …) and omissions in the quotation, without resorting to ungainly square-bracketed ellipses […] which would impede the flow of reading.

Chapter One

The self

1.1 Introduction

'The self' can be conceptualized in various ways—as a subject of experience, as the locus of 'personal identity', as a fictional construct, as a moral and social agent, etc. (Gallagher, 2011a). So diverse are these conceptions that Eric Olson (1998) has argued that, given the absence of an 'agreed use of the term 'self"', we should abandon talk of selves altogether.

Over the past two decades, however, publications on the self—by philosophers, psychologists, psychiatrists, and neuroscientists—have proliferated; some aimed at academic, some at lay audiences. Ignoring for the moment the countless articles on the self published in academic journals catering to various disciplines, already a brief (and, no doubt, incomplete) survey of books with 'self' (or 'ego') in the title or subtitle published in the last twenty years reveals the diversity of their concerns. The philosopher Marya Schechtman, in *The Constitution of Selves* (1996), defends a narrative constructionist view of the self,[1] as does Valerie Hardcastle in *Constructing the Self* (2008). The neurophilosopher Thomas Metzinger, in his monumental *Being No One: The self-model theory of subjectivity* (2003) and in *The Ego Tunnel* (2009), takes the view that the self as traditionally conceived is an illusion; and so does the psychologist Bruce Hood in *The Self Illusion* (2012). Richard Sorabji's *Self* (2006) covers the philosophy of self, identity, self-awareness, and mortality from the ancients to the present. Dan Zahavi (2005) and J. J. Valberg (2007) take phenomenological approaches to the self. Galen Strawson offers a 'revisionary metaphysics' of the self in *Selves* (2009). Meanwhile, the neuroscientist Antonio Damasio revisits and revises his earlier (1999) writings on the neurological basis of the self in *Self Comes*

[1] See my chs. 5–6.

to Mind (2010). A companion piece to Damasio is Patricia Churchland's *Touching a Nerve* (2013), whose ontologically bold subtitle *The Self as Brain* changes to a more non-committal *Our Brains, Our Selves* in the second edition. Then, John Doris's *Talking to Our Selves* (2015) takes a psychologically informed social constructionist approach, not only to the self, but to agency. At the other end of the ontological spectrum between simple and constructionist conceptions of the self, Geoffrey Madell endorses a 'simple' view of personal identity in *The Essence of the Self* (2015). Towering over all these is *The Oxford Handbook of the Self* edited by Shaun Gallagher (2011b), whose thirty-seven contributors are philosophers, psychologists, psychiatrists, and neuroscientists, covering the self in the history of Western philosophy, bodily selves, the phenomenology and metaphysics of self, personal identity, narrative identity, self-knowledge, actions and morality, self pathologies, and the self in pragmatist thought and social constructionism.

This diversity of titles and topics seems to support Olson's complaint that there is little accord on what the self is—indeed, a compendium like the *Oxford Handbook* seems almost to celebrate the variety of conceptions and definitions of 'the self' in current academic discourse. On the other hand, the sheer volume of recent treatments of 'the self' under that label makes it apparent that his suggestion that we abandon the term has not, so far, met (and looks, for the foreseeable future, unlikely to meet) with much agreement—quite the reverse. Yet, not only philosophers, but also psychologists frequently fail to explain what they mean by 'the self' (Klein, 2012a). The challenge, then, is to bring some order to the current conceptual disarray about the self. That is the object of this thesis.

I here offer a view of the self that is both empirically informed and (I hope) empirically fruitful, and that seeks to avoid the confusions and misconceptions in the recent (and not so recent) philosophical and psychological literature on the self. My view of the self is that it is a functional system, both complex and dynamic. In this chapter, I will motivate and explain this view, highlight some of its advantages, and preview the arguments against rival views that I develop in subsequent chapters.

1.2 Main claims

In brief, my main claims are these. First, there *is* a self in a meaningful sense, grounded in empirical considerations—but not as a simple, indivisible and invariable entity (§ 1.3). Rather, the self is a complex and dynamic functional **system** (§ 1.4). Secondly, this complex and dynamic self is **not a construct** from memory and/or narratives but, instead, engages in their construction (§ 1.5). The bulk of my thesis (especially chs. 4–6) is a protracted argument in support of this second claim.

My system view of the self (§ 1.4) entails three subsidiary claims, two of which I'll revisit in later chapters. First, **persistence** conditions of the self depend on those of the material entity in which it is realized, i.e. the brain (§ 1.6.1). Secondly, and relatedly, questions about the self are separable from, and prior to, questions about 'personal identity' (§ 1.6.2 and § 2.4). Thirdly, there is **one** self per brain, synchronically and diachronically (§ 1.7 and § 8.3).

1.3 No self?

Despite the plethora of recent works on some notion or other of 'the self' cited in the introduction, it remains a recurring challenge for an apprentice philosopher writing a thesis on the self—both in conversations with other philosophers and in studying the relevant literature—to defend the view that we should entertain any such notion at all. For, alongside those expounding different conceptions of the self, there are not a few sceptics about, and outright deniers of, the self in the history of Western philosophy, most famously David Hume (of whom more shortly). Among the prominent self deniers in 20th-century analytic philosophy, Richard Sorabji (2006) names Ludwig Wittgenstein, his disciples Elizabeth Anscombe—whose (1975) startling conclusion that *'I' does not refer to anything* is perhaps one of the absurdest outcrops of 'analytic' philosophy—and Norman Malcolm; further, Anthony Kenny and Daniel Dennett.

With the exception of Dennett (who is not actually a self *denier*, but a fictionalist—see § 5.3.3), I will not discuss their views here. Rather, I attempt to show in this and the next section, by discussing Hume's scepti-

cism and some more recent contributions to the literature, that there *is* a case for talking about 'the self'—but under a conceptualization that takes seriously some of the difficulties that made Hume forbear to discuss the self in all but his earliest work. It is worth recalling Hume's difficulties, in § I.iv.6 of his *Treatise* (1739/1978), at some length:

> If any impression gives rise to the idea of self, that impression must continue invariably the same, thro' the whole course of our lives; since self is suppos'd to exist after that manner. But there is no impression constant and invariable. Pain and pleasure, grief and joy, passions and sensations succeed each other, and never all exist at the same time. It cannot, therefore, be from any of these impressions, or from any other, that the idea of self is deriv'd; and consequently there is no such idea.
>
> . . . For my part, when I enter most intimately into what I call *myself*, I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never can catch *myself* at any time without a perception, and never can observe any thing but the perception. . . . If any one upon serious and unprejudic'd reflection, thinks he has a different notion of *himself*, I must confess I can reason no longer with him. All I can allow him is, that he may be in the right as well as I, and that we are essentially different in this particular. He may, perhaps, perceive something simple and continu'd, which he calls *himself*; tho' I am certain there is no such principle in me.
>
> But setting aside some metaphysicians of this kind, I may venture to affirm of the rest of mankind, that they are nothing but a bundle or collection of different perceptions, which succeed each other with an inconceivable rapidity, and are in a perpetual flux and movement. . . . (pp. 251–2)

Hume here makes two connected assumptions that, on but the slightest consideration, seem dubious. While these two assumptions explain Hume's denial of the self in this passage, questioning them shows that such denial is rather precipitate. Hume's first assumption is that the self should be *introspectible* as such—or, more precisely, that the *idea* of self can only result from a corresponding *impression*, which in this case would have to be the result of introspection. Leaving aside a lengthy consideration of the Empiricist doctrine of ideas being always derived from impressions, it will suffice to note that even Hume's system allows that ideas can be *complex* and thereby not necessarily a reflection of a single impression.[2] If, then, the idea

[2] 'I observe, that many of our complex ideas never had impressions, that corresponded to them' (Hume, 1739/1978, I.i.1, p. 3).

of self were a complex one, it would be unsurprising for it not to be underwritten by a single impression and therefore not made directly apparent in introspection.

But the self Hume fails to locate amid his impressions is conceived as something *simple* and *invariable*, 'since self is suppos'd to exist after that manner'. Hume here takes this supposed conception of the self as read, without telling us why, or by whom, it is so supposed.[3] Whatever its antecedents, this conception of self as simple and invariable is Hume's second dubious assumption, and the one that leads him to deny that there is such a thing: the 'perpetual flux and movement' of perceptions rules out such a self. Yet, as Hume goes on,

> we may observe, that the true idea of the human mind, is to consider it as a system of different perceptions or different existences, which are link'd together by the relation of cause and effect, and mutually produce, destroy, influence, and modify each other. Our impressions give rise to their correspondent ideas; and these ideas in their turn produce other impressions. One thought chaces another, and draws after it a third, by which it is expell'd in its turn. In this respect, I cannot compare the soul more properly to any thing than to a republic or commonwealth, in which the several members are united by the reciprocal ties of government and subordination, and give rise to other persons, who propagate the same republic in the incessant changes of its parts. And as the same individual republic may not only change its members, but also its laws and constitutions; in like manner the same person may vary his character and disposition, as well as his impressions and ideas, without losing his identity. (p. 261)

Hume here speaks of the 'mind' and 'soul' and attributes to them qualities of compositionality and changeability. And he points out that such mental mutability is no bar to the preservation of an individual's identity. Might not therefore an individual's *self*, appropriately construed, have these same qualities of compositionality and changeability? William James (1890) takes such a view, as do those philosophers and psychologists after him (e.g. Mead, Gergen, Bruner, Schechtman) who espouse some form of constructionist theory of the self (of which more anon).

---

[3] Writing within a century of Descartes's (1641) *Meditations*—and while in France—Hume may have had something like a Cartesian ego in mind: a simple, indivisible, and immaterial substance.

Notwithstanding these 20th-century accounts of the self, Thomas Metzinger begins his provocatively titled *Being No One* (2003) by contending that 'Nobody ever *was* or *had* a self' (p. 1). But he then, on the basis of both our neurobiology and our phenomenology, posits a *phenomenal self-model* (PSM) whose contents are an integration of 'your current bodily sensations, your present emotional situation, plus all the contents of your phenomenally experienced cognitive processing' (p. 299). And it is the activity of the PSM that gives us a sense of self.[4] So, on the one hand, Metzinger presents an intricate and elaborate theory of a neurophenomenomal system that one might conceptualize as 'the self'—on the other hand, he maintains that there is no such thing as a self. This may well apply to 'the self' understood *a priori* as an ontologically simple—indivisible, unchanging—entity. But we could still allow that a self could be conceived of as something more complex and dynamic.

Why does Metzinger, like Hume, seem to take the view of the self as a simple and invariable entity as the default conceptualization? Whereas Hume may have had the excuse that, in his time, the self was perhaps widely—if not universally—'suppos'd to exist after that manner', the same cannot be said for Metzinger, before whom quite a number of philosophers and psychologists have ventured to give accounts of the self under more complex descriptions.

Is, perhaps, the conception of self as a simple invariable entity more attuned to our ordinary use of the term? No. As Patricia Churchland (2002a) observes, popular discourse concerning the self is highly unsystematic and often metaphorical, sometimes equating 'self' with 'body', sometimes distinguishing 'self' from 'body'; sometimes conceiving of 'self' as an object and sometimes a subject; sometimes taking 'self' to mean 'person', sometimes just *some* of a person's attributes and dispositions.

> What the *nonsystematic* character of metaphorical language suggests is that *the self* is not a thoroughly coherent, single, unified representational scheme about which we have thoroughly coherent, unified beliefs. Rather, the self is something like a squadron of capacities flying in loose formation. (p. 63)

[4] See § 9.2.1.

In other words, far from committing to a Cartesian-ego-like view of the self, ordinary discourse supports conceiving of the self as a composite, rather than a single and invariable entity.

As it happens, Hume himself had second thoughts about what the self might be, as attested in the Appendix to his *Treatise* (1739/1978):

> When I turn my reflexion on *myself*, I never can perceive this *self* without some one or more perceptions; nor can I ever perceive any thing but the perceptions. *'Tis the composition of these, therefore, which forms the self.*
>
> (p. 634, third emphasis added)

Here, while repeating his earlier insistence that his self is not introspectible on its own, Hume now admits the notion of the self as a composite. (This still leaves him with the puzzle as to how we are able to perceive the connection between our distinct perceptions and thus how they form the self, but here perhaps Hume is too empiricist for his own good. For one thing, it is not necessary for a composite self that we should by mere introspection be able to perceive all the connections between its parts. For another, Hume's own account of the compositional nature of the 'mind' or 'soul' quoted above can readily be transposed to the self.)

In sum, I agree with Hume (and Metzinger) that there is no self conceived as something both *simple* and *invariable*. This, however, does not rule out a conception of the self as both *complex* and *dynamic*, and it is such a conception of the self that I now propose we should adopt.

1.4 A complex dynamic system

There should be nothing particularly startling about the suggestion that the self is something complex. It often happens that, in the light of empirical observations, ideas that initially seemed to denote something simple turn out to be about complex phenomena. Take the example of air.[5] The Ancients thought air to be one of only four elements, the four simplest substances in nature. We now know air to be a complex gaseous compound made up of (mostly) molecular nitrogen, molecular oxygen, argon, and small amounts of other atomic or molecular gases in varying quantities. Air does not cease

[5] I thank Dominic Gregory for suggesting this analogy.

to be air when the quantities of its components, or its pressure, vary somewhat, or when other components are added to it: polluted air is still air. And although our concept of air has evolved with scientific progress, its basic referent has not changed: what we call 'air' is the same thing that the ancient Greeks called ἀήρ—it is what they breathed, and what we breathe. But since scientific exploration has greatly added to our understanding of what air is, we would not want to return to the ancient folk concept.

Now, something analogous seems to apply to the self. Here, too, is an idea that seemed, to some at least (Descartes, for one), to refer to something simple. As discussed in the previous section, that particular conception of the self has run into difficulties. But then, why not follow Hume in dismissing the idea of a self altogether? Because there are rather important—but complex—processes at work in the natural world that match at least some of our uses of 'the self'. For the ability to distinguish 'self' from 'non-self' is one of the most useful evolutionary adaptations around. This is the capacity of an organism to represent itself *as* itself *to* itself, so as to distinguish itself from its environment (and thus to avoid harming, or indeed eating, itself) and to be able to take care of itself (to avoid being eaten by others). In more sophisticated organisms like ourselves, there is also the ability to represent oneself over time.

In characterizing the self through these abilities, I take my lead from Pat Churchland (2002a), who suggests 'recasting Hume's problem in terms of *self-representational capacities*'[6] (p. 63):

> Self-representational capacities include representing the internal milieu and viscera via chemical and neural pathways aimed largely at the brainstem and hypothalamus; representing musculoskeletal structures via the somatic sensory system; representing autobiographical events via medial temporal lobe structures; deferring gratification and controlling impulses via prefrontal lobe and limbic structures; and representing the sequence of actions to take next, as well as representing where one is in space-time and the social order. (Churchland, 2002b, p. 309)

---

[6] Here, and throughout this thesis, the hyphenated prefix 'self-' has its ordinary reflexive meaning. Thus, 'self-representation' is an instance of something representing *itself*, 'self-awareness' of something being aware of *itself*, 'self-monitoring' of something monitoring *itself*, etc. The prefix should not be understood as referring to *the* self—this obviously would make explaining the self in terms of self-representations and/or self-monitoring viciously circular.

This is still quite a large bag of tricks, ranging from homeostasis through sensorimotor integration and proprioception to higher-level self-representations such as autobiographical memory, impulse control, and representing and evaluating one's situation in the social context. It will be useful, therefore, to distinguish among them. Homeostatic functions and basic sensorimotor co-ordination are, of course, vitally important, but they occur below the threshold of conscious awareness. They amount to what Antonio Damasio (1999; 2010) calls the '*proto-self*' because, even at that level, different processes are already *integrated*, allowing an organism to regulate its internal milieu and to interact with its environment. Moreover, as Churchland (2002b) puts it, '[t]his level of integration, shared across many species, is the nonconscious neurobiological platform for higher levels of self-representation' (p. 310).

It is these *higher-level* self-representational processes that generally concern us when we speak of 'the self' in human beings: memory of oneself, planning for oneself, impulse control, assessing one's social situation, and so on. Developing Churchland's suggestion, I propose, therefore, that we should understand the self as those self-representational processes that are potentially conscious (*potentially* conscious because they are not always consciously accessible—indeed, the larger part of e.g. our memories is not, fortunately, permanently conscious—but, by their potential of *becoming* consciously accessible, they are distinguishable from the lower-level, always *unconscious* self-representational processes involved in homeostasis and sensorimotor integration). These higher-level self-representational capacities result in both awareness of oneself as a *distinct individual* within one's natural and social environment, *and* awareness of oneself as the *same individual over time*.[7] In accomplishing this, I suggest, our higher-level self-representational capacities form a more-or-less coherent *self-monitoring system*—in short, 'the self'. Call this view of the self **the system view**.

In speaking of a 'system', I mean nothing more grandiose than a bundle or cluster of different processes serving a common function (cf. the visual system, the sensorimotor system, the immune system…)—in this case, the function of self-monitoring. The self conceived of in this way is

---

[7] Thus, they satisfy what Richard Sorabji (2006), in his characterization of the self, rather nicely calls one's 'need to see [oneself] as *me* and *me again*' (p. 31). But we can recast this as one's *capacity* to see oneself as '*me* and *me* again'.

defined by what it does, rather than by how or where precisely it is instantiated. Thus, I am not suggesting that the self is a localized neurological system. Rather, the self *qua* functional system recruits many different neural networks and subsystems (Gillihan & Farah, 2005; Vogeley & Gallagher, 2011). This also means that localized neural damage or deterioration affecting some self-representational capacity may leave others intact (see my discussion of defects and disorders of the self in ch. 8). But it should also be clear that the self as a system depends upon the brain, which provides the neural architecture in which it is realized.[8]

This system is clearly *complex*, as it involves a number of different self-representational capacities. This complexity will be a recurring theme in the rest of my thesis. The self as a system is also *dynamic*: as its self-representational inputs change over time, so the system changes. It adapts to new goals and takes account of new experiences. And at any one time, different self-monitoring processes take precedence over others. For example, in writing up a thesis, monitoring goal-directed behaviour is a more useful self-representational capacity to deploy than, say, revelling in autobiographical recollections. But all these self-monitoring activities and self-representing capacities form part of the system.

The system view is not, at this stage, a fully-fledged theory of the self.[9] For the purposes of this thesis, it is a view with which to contrast the other views of the self that I will discuss in subsequent chapters. But beyond that, it is something of a conceptual recommendation for philosophers, psychologists and neuroscientists alike, which—without being an exercise in conceptual analysis—captures many of the usages of 'the self' in both ordinary discourse and the recent scientific literature, while avoiding the shortcomings of other conceptions of the self. If (*pace* Olson), we *are* going to contin-

---

[8] Conceivably, such a system could be instantiated in inorganic machines. Mobile robots, for instance, require something of a proto-self and sometimes more sophisticated *unconscious* self-models, in order to co-ordinate their movements and adapt to their environment (Metzinger, 2007). I'll revisit this topic briefly in my conclusions (§ 9.3), but otherwise, such putative artificial selves are not my concern in this thesis (excepting a thought-experimental robot in § 6.2.1). For the avoidance of doubt, unless otherwise specified, where I speak of 'the self', I here refer to the *human* self.

[9] I will return to the question of quite how to characterize the self system in my concluding chapter (§ 9.2).

ue speaking of 'the self', I contend—and the rest of this thesis will show—that the system view is a more useful conceptualization of the self than others.

Among our higher-level self-representational processes that make up our self-monitoring system, my focus in this thesis will be on those that are engaged in self-monitoring *over time*. These are the processes by which we construct autobiographical memories and self narratives, which on other views of the self are often mistaken for what the self *is*. In the following section, I will sketch my disagreements with these views, which I develop in Chapters 4–8.

1.5 Memory and narrative

In his neurological theory of the self, Antonio Damasio (1999; 2010) notes that in forming, processing, and retrieving memories, our brains extend our consciousness—our *self*-consciousness—beyond the present. Self-monitoring does not stop at representing an organism in its current activities and interactions with the world (the 'core self', in Damasio's terms), but involves learning from current interactions, recalling past interactions, and planning future interactions (and in so doing the self becomes an 'autobiographical self'). There are obvious adaptive advantages to self-monitoring over time. An organism that is able to remember *itself* encountering danger in a particular location will in future try to avoid that spot, or if unavoidable proceed to it only with great caution. Remembering *oneself* successfully overcoming some obstacle by a particular method will enable one to deploy similar strategies in overcoming similar obstacles. It might be suggested that the mere learning of what situations are dangerous or what strategies are successful, without implicating oneself in them, would be just as efficient. But there is an added advantage of having a memory representing *oneself* in these interactions. When faced with the need to cross running water, an instinct to look for a ford or stepping stones is useful, but having a recollection o*f one's own* previous successful (or unsuccessful) brook-crossings is more useful: it allows one to replicate the very motor actions that previously got one to the other side (and avoid repeating the ones that did not).

In humans, of course, self-monitoring over time goes far beyond such matters of bare survival. Most of our interactions with the world are interactions with other human beings. Consequently, it is in *social* contexts that we form most of our personal memories, from parent–child interactions through the experiences of schooling to our various friendships, sports teams, amorous liaisons, professional interactions, etc. And here the *recall* of one's memories, though important in guiding us to avoid *faux-pas* and to pursue successful strategies in navigating our complex social world, is only part of the story. Much of one's (mostly unconscious) self-monitoring over time consists in *organizing* one's memories. Such organization is somewhat chronological but mostly thematic. Particular memories are tagged according to what period of one's life they belong to, whether they concern one's educational experiences, one's leisured social interactions, or one's professional life, what pursuit or goal or achievement, embarrassment or failure they relate to, what other people and what places they involve, and so on. In this way, the self builds up an 'autobiographical knowledge base' (Conway & Rubin, 1993—see my § 4.3.1).

This autobiographical knowledge is not, generally, a completely private matter. While there may be elements to it that one likes to keep to oneself, the social context frequently demands that one recount particular selections of one's experiences to others in greater or lesser detail. Making new acquaintances, reunions with old acquaintances, job interviews, small talk at parties or academic conferences, meeting friends in the pub after work (and, returning home, responding to such queries as 'Good day at the office?' and/or 'Why are you so late?'), phone calls and visits from and to parents and other relatives, judicial enquiries, the Catholic sacrament of confession, being a 'castaway' on Radio 4's *Desert Island Discs*…—all these are situations where we are called upon to produce shorter or longer, more or less detailed *self narratives*, selective (and not always entirely truthful) accounts of ourselves and our experiences. Moreover, the rehearsal of our experiences in such narrative practices shapes and consolidates our autobiographical memories (see ch. 5).

Self-monitoring over time, then, is accomplished by one's remembering and narrating one's experiences—these activities of the self are important for one's self-awareness of 'me again'.

It is perhaps understandable, then, that our autobiographical memories and self narratives could plausibly be mistaken for *being* the self. But this is the view—or rather, a whole family of views, which I will call **historical-constructionist views** of the self—that I wish to oppose in this thesis. They include both the view that the self is constructed from our autobiographical *memories*, as many readings of Locke (see ch. 2) have it, and the currently widespread view that the self is a *narrative* construct of some sort (see ch. 5). The feature that they have in common, and which distinguishes them from my system view, is that the self is taken to be a *construct* or *construction*, whose raw materials, as it were, are our autobiographical memories, our self narratives, or both—that is, the personal *histories* we do indeed construct of ourselves. But *who or what* is doing this construction? My claim is that historical-constructionist theories of the self, in taking the self to be a construct from our memories and stories, put the cart before the horses: they confuse the agent—the self *qua* self-monitoring system—with some of its characteristic activities—the construction of autobiographical memories and narratives as a means of self-monitoring over time. For the self here is the constructor, not the construction. Our memories and narratives are not the self, but its productions; they are, as it were, the 'print-outs' of our self-monitoring processes. Let me explain.

Memory, to begin with, is not fixed. We may have what are sometimes called memory traces or 'engrams' of our experiences, but these are constantly updated, recombined, sorted, and organized in the light of subsequent experiences and the circumstantial demands of each particular act of recollection. As Conway and Rubin (1993) note, there is 'no specific type of knowledge which can be easily singled out as being *a memory*. Rather, memories are compilations, constructions, or compositions of knowledge' (p. 104).

> Remembering is not the re-excitation of innumerable fixed, lifeless and fragmentary traces. It is an imaginative reconstruction, or construction, built out of the relation of our attitude towards a whole active mass of organised past reactions and experience . . . (Bartlett, 1932, p. 213)

This is particularly so for so-called *episodic* memories, the recollection of particular scenes and events from one's life:

> Episodic remembering bears a close family resemblance to other higher-order mental achievements (e.g., introspecting, daydreaming, anticipating)

> that are not typically considered to be acts of memory.
>
> (Wheeler, 2000, p. 606)

Thus, our memories—and particularly our episodic recollections—are *constructions* or *reconstructions*. And where our memories are autobiographical, they are an instance of constructive self-representation, *at the time of recollection*, of a previous state of oneself. Their reconstruction is an activity of the self-monitoring system which I have identified as the self (see § 4.5).

Therefore, if we, as we must, take a constructionist view of autobiographical memory, we cannot at the same time subscribe to a constructionist view of the self as assembled from memories, since we cannot take memories to be a fixed 'raw material'. Worse, taking both these views would lead to a circularity of the self and its memories constructing each other. And it would leave us without an agent or engine to accomplish the construction of either. One could say that the constructor of both memory and self is the brain—but it is precisely those activities of the brain that are concerned with self-representation and self-monitoring that I take to be the self. And those are the brain activities that are behind, *and prior to*, our autobiographical memories.

The activity of constructing self *narratives* is continuous with that of constructing autobiographical memories. For one thing, narrative practices serve to rehearse and structure autobiographical memory. And the normally unconscious organizational and editing processes that compile our autobiographical knowledge base are here made more explicit; for the purpose of telling others about ourselves, we again select and recombine memories and, additionally, subsume them under a narrative arc. Usually, our narrative practices are instances of self-*presentation* as much as self-representation.

Self narratives, therefore, are the public face of the self, or better: its public *faces*—for our self narratives are generally tailored and adapted to the relevant audience: different self narratives are deployed with friends in the pub, at job interviews, and in a court of law. (Where self narratives fail to take account of our interlocutors' expectations—by being too detailed, too long, too forthright, or in some cases perhaps too cagey—they amount to social solecisms. But even then, they are exercises in self-presentation, however misjudged.) Our self narratives reveal how we want to come

across in different social situations, how we see ourselves, whom we take ourselves to be.

There is some plausibility, therefore, to the notion that our self narratives *are* the self. But this can only be a metaphor. Our self narratives are partial, selective, context-dependent. Sometimes they are untruthful. But, most important, if the self were a narrative construct, who or what would then be the author of those narratives? I will discuss this **authorship problem** in detail in Chapter 6. Its solution is the system view: it is the self *qua* self-representational brain activity that, *inter alia*, produces our self narratives—and not our self narratives that produce the self. To confuse the self with its narrative productions is somewhat analogous to mistaking a recording of the Berlin Philharmonic for the orchestra itself. For, while it is a perfectly appropriate shorthand usage to say 'That *is* the Berlin Philharmonic' when listening to one of its recordings, it is understood that what is meant here is that the sound waves emanating from the loudspeakers, the information encoded on the CD, are a recording *of* the Berlin Philharmonic. The reason why the shorthand is appropriate is that it is, of course, an orchestra's *function* to produce music. And it is a function of the self—but only *one* of its functions—to produce self-representations in narrative form.

The analogy, then, does not carry over fully to the self. Even speaking metaphorically, someone's autobiography, however detailed and truthful it might be, *isn't* all there is to its author in the way that the Berlin Philharmonic's performances and recordings (and choices of conductors) may satisfactorily characterize that orchestra. Our self narratives—or for that matter our autobiographical memories—do not exhaust the functions of our self-monitoring system like making music exhausts the function of an orchestra. Our self-representations also include implicit memories, character traits, attitudes, affective and behavioural dispositions. This becomes especially clear when explicit self-representational activities—like narrating and remembering—are impaired by neural damage or degeneration (see ch. 8), or in individuals who may have no narrative capacities to begin with, as may be the case in autism (ch. 7). In such cases, other, implicit self-representational functions remain, suggesting the presence and persistence of the self in the absence of its usual overt indicators. This brings me to the issue of the persistence of self more generally, which I'll now address—along with the relations between self and questions of identity.

1.6 Persistence, identity, and personhood

1.6.1 Persistence

Complexity in response to certain philosophical questions can entail simplicity in response to others. In taking the view that the self is a complex system realized in the brain, I find that the answers to some of the questions that have befogged the philosophical literature on the self become pleasingly straightforward. The foremost of these is the question as to the *persistence conditions* of the self, and the answer to it is indeed relatively simple. Since the self is a system of brain activity, it persists for as long as the brain is active, i.e. living. When the brain is dead, so is the self. But we can be a little more precise about what brain activity is necessary for the self to persist. We've taken the self to be those higher-level self-representational activities that are potentially conscious (§ 1.4). These require activity in the forebrain, especially the cerebral cortex. Therefore, in a brain that had completely ceased its cortical and most subcortical activity but whose brainstem was still active, maintaining homeostasis and thus technically keeping its organism alive, we should say that the self no longer persists. The organism may still have a 'proto-self' in its brainstem's self-representational activity, but no higher-level self-monitoring, no self.

Notice, however, that localized damage to or even progressive global deterioration of cortical self-monitoring activities (as in amnesias and dementia) does not yet entail the death of the self. In these conditions, as we shall see further on (§ 8.2), some self-representational capacities remain active. Nor does a self-monitoring system that produces odd or erroneous readings (as in schizophrenia, see § 8.3.4) amount to a loss of self. A damaged or malfunctioning system is still *that* system; thus, the self as a system is still the self, even when damaged or malfunctioning.

Could a self *qua* functional self-monitoring system be transferred from one brain to another or to some other, perhaps inorganic, platform? Very likely not. Such scenarios assume the brain to work like a serial computer with distinguishable hardware and software. But brains are not like that. Information processing in the brain is not just the exchange of electrical po-

tentials between neurons, but involves the very configuration of neural pathways and synaptic connections, which is itself dynamic and subject to change over time in response to new inputs, across the whole life-span of the organism (the phenomenon known as neuroplasticity). So, since our cerebral processing architecture is not separable into hardware and software, it is not at all clear how we could conceive of one brain's processing systems being transferred to another or to a different medium, for it is not clear quite what, in neurophysiological terms, is supposed to be transferred here. For we cannot separate the *function* of self-monitoring from its 'realization' in the brain—the realization is the very neural architecture that fulfils the function.

What about transplanting the whole brain? If live brain transplants were possible, surely such a transplant would entail the transplant of the brain's self as well? Again, while a *logical* possibility, that scenario seems *empirically* highly unlikely, once we consider the details of what would be involved in such a transplant. The brain as the operations centre of the central nervous system is not easily isolable from the multitude of efferent and afferent neural pathways which connect it to all parts of the organism. But, supposing that a transplanted brain could be successfully attached to all the neural pathways of the recipient organism (all this while keeping it supplied with blood and thereby with oxygen, without which widespread tissue death would occur very rapidly), the result would be a very confused brain, whose brainstem and hypothalamus would now be faced with regulating a strange organism and whose sensorimotor systems would be confronted with receiving sensory input from, and generating motor output to, new and unaccustomed limbs and sense organs. Most likely, the transplanted brain would be unable to adjust to such a sudden and overpowering demand on its plasticity. (Our brains do of course adjust to somatic changes, particularly in childhood and adolescence. Such changes, however, are gradual, rather than involving the wholesale replacement of an organism with another.) But, supposing a transplanted brain *could* adapt to all that, it would then be generating self-representations of a different organism. Even if some traces of higher-level self-representations from its original organism remained, the self-monitoring system would have been reconfigured from the bottom up. In such circumstances, the question

whether the transplanted brain's self-monitoring system would be the *same* persistent self is rather moot.

Leaving aside the neuroscience fiction, there is another, connected issue to which the system view yields a simple answer. This concerns a distinction that is sometimes made between questions about the self considered at a given time (*synchronic*) and questions about the self over time (*diachronic*) (Zahavi, 2011). It is often supposed that the diachronic *unity* of the self poses problems and requires conditions in addition to those for its synchronic unity. What is it, the question goes, that unifies the self over time? The historical-constructionist answer is that memory, psychological continuity, or narrative structure provide the diachronic unity of self (Grice, 1941; Shoemaker, 1984; 2011; Schechtman, 1996). But this strategy is again the result of confusing the self with its productions, or of mistaking *evidence* of a self with *criteria* for a self. As discussed in the previous section, some of the key functions of a self-monitoring system are to produce autobiographical memories and self narratives. Their *contents*, by and large, may indeed be temporally continuous. But it is not their continuity that ensures the continuity of self. Rather, it is the other way around: the continuously operating self-monitoring system is what allows for the production of such temporally coherent autobiographical memories and narratives. And even when the self produces temporally *discontinuous* self-representations (which it does far more frequently than the historical-constructionist views allow—see §§ 4.5.1; 6.3.2), it is the same system, the same self persisting. Thus, on the system view, there is no problem of diachronic unity. The continuous activity of the system, rather than the continuity of its productions, is what makes the self persist over time.

1.6.2 Identity and personhood

Since Locke (1690/1706—see my ch. 2), the question of the persistence of self over time has often been conflated with what the philosophical literature calls the question(s) of 'personal identity' over time. At least, discussions of personal identity still often use 'the self' as a convenient label for their topic (e.g. Williams, 1970; Madell, 2015). Colloquially speaking, this may seem plausible: are not questions about the identity of my person questions about my self? But in defining these questions, it soon becomes

necessary to clarify what we mean by 'identity', 'person', and 'self', and this is where unwarranted confusions and conflations abound. I have already sketched what I take the self to be. What about the other two terms, 'identity' and 'person'?

To begin with, *identity*—in the philosophical sense of the term—is a straightforward logical relation, the relation a thing bears to itself (not to its *self*—the ambiguities of language tend to add to the confusions here[10]) and nothing else. The question of something's identity over time is thus the question whether it is the *same* thing, or a different thing, at different times. This is, or ought to be, a straightforward yes-or-no question. But with complex entities, questions of identity may be anything but straightforward. For it now matters just what sort of an entity it is whose identity over time we're interested in, and what its criteria for sameness are:

> [T]he question of personal identity is a difficult one, especially when some theorists tend to look for identity in only the biological or embodied existence of persons, and others tend to look for it exclusively in psychological existence. (Gallagher, 2011a, p. 13)

Thus, the difficulties with and disagreements over questions of personal identity are not difficulties and disagreements about 'identity', but about what a *person* is. Is it a biological organism—a 'human animal' (Olson, 1997)—which happens to have certain qualities of personhood? Or is it a bundle of psychological traits independent of their physiological substrate? These are the two classic and apparently irreconcilable positions in the personal identity literature.

Psychological-continuity theories of personal identity resort to a variety of otherworldly thought experiments—disembodiment, fusions, fissions—to tease out the identity conditions of persons (see ch. 3). One approach resulting from this is a view of persons as four-dimensional objects whose various 'time-slices' are causally connected and thus held together in their temporal dimension (Perry, 1972; Lewis, 1976). This approach

---

[10] One way of teasing apart 'identity' and 'self' is to make use of the difference between Latin *idem* 'the same' and *ipse* 'oneself', as suggested by Paul Ricœur (1990), leading to a distinction between 'sameness' (*mêmeté*) and 'selfhood' (*ipséité*). While the distinction is useful, Ricœur's own notion of selfhood is not, for it (as far as I understand it) takes the historical-constructionist approach that I critique in this thesis.

avoids transitivity objections (see § 2.3.1) but faces the unsatisfactory consequence that multiple persons can occupy the same biological organism.

The opposite position is the biological-continuity theory of personal identity, defended notably by Eric Olson (1997). According to this view, whether one is the *same* at different times boils down to one's being the same biological organism, the same animal. Psychological continuity plays no part in settling the question of personal identity over time. As an account of one's diachronic *identity*, this seems right. But one might wonder whether it accounts for all the characteristics we associate with *persons*, such as, in Schechtman's (1996) terms, 'moral responsibility, self-interested concern, and compensation', and probably others besides. Such characteristics may indeed require some kind of psychological continuity, *in addition to* (not instead of) the biological continuity of the organism. But then it seems that *persons* are insufficiently basic sortals to be the subject of questions of diachronic identity.

As we'll see in the next chapter (§ 2.4.1), the question of what makes a person is an important and controversial philosophical issue—but it is not a question we need to answer when discussing the *self*. I'll argue that, while personhood requires higher-level self-monitoring functions, our self-representational capacities can operate with or without the trappings and trimmings of personhood. Having a self is thus a necessary but not sufficient condition for being a person.

The concern of this thesis, then, is not 'personal identity'—it is *neither* identity *nor* personhood. As far as questions of identity are about persistence through time, these have already been answered in the preceding subsection: the persistence conditions of the self *qua* self-monitoring system are the persistence conditions of the relevant brain activity. As to personhood, a self may or may not have the attributes of a person—but what these are is a separate and rather vexed question. I will sometimes have to engage with the personal identity literature (especially in chs. 2–3). But, insofar as my thesis is a contribution to that literature, it is a foundational one: we need an account of the self before we can go on to questions of personhood.

1.7 One self per brain

Another consequence of the system view is numerical: with the self as a system of brain activity, it makes little sense to impute more than one self to any one brain, either consecutively or simultaneously.

First, could there be distinct *successive* selves instantiated in the same persistent brain at different times? This seems to have been the contention of historical-constructionist theories of the self since Locke (1690/1706):

> But if it be possible for the same man to have distinct incommunicable consciousness at different times, it is past doubt the same man would at different times make different persons . . . (II.xxvii.20)

—and thereby (according to Locke's person–self equation[11]) at different times have different *selves*. Now, if the self were merely a construct, such a view might (just) be plausible. But if we take the system view, we must ask what conditions would have to obtain, empirically, for one organism at different times to have different selves. (I shall here leave aside the scenario of the organism's receiving a transplanted brain, such wildly fictional cases having already been discussed in § 1.6.1.) The brain's self-monitoring activities would have to be entirely reset. At the very least, its remembering and self-narrating capacities (as the outwardly most important aspects of self-monitoring) would have to be put in a position of restarting from, as it were, a blank slate. This would require *complete amnesia*—a condition which, in discussions of the self, has been a favourite philosophers' trope since Locke. But it is not a condition that has ever been reported as actually having occurred: as I note in my discussion of amnesias and dementias (§ 8.2), conditions of memory loss always seem to leave some aspects of memory (basic language skills, procedural memory, implicit memory of e.g. character traits, musical memory…) intact. As already observed, brains are not like computers in which one may simply wipe all the software and leave the hardware intact. A brain that had ceased to maintain even the most basic self-representational processes would be so devastatingly damaged that it would be unlikely to be able to start those processes anew. Thus, it is, empirically speaking, no more sensible to assume that there can be distinct self-monitoring systems in one brain at different times than it

[11] See § 2.4.2.

would be to speak of one's having distinct successive visual systems or motor systems at different times in one's life.

Secondly, what about the possibility of one's having distinct self-monitoring systems simultaneously, operating *in parallel*? It might be suggested that since our higher-level self-monitoring functions are a matter of cortical activity, and since the cortex is arranged in two anatomically symmetrical hemispheres, there could be a possibility of two self-monitoring systems operating independently in each cerebral hemisphere. That suggestion might at first seem to be supported by commissurotomy or 'split-brain' patients, whose corpus callosum—the central commissure of nerve fibres linking the two cerebral hemispheres—has been resected (as a treatment for severe epilepsy). I will discuss this condition and its implications more fully later (§§ 6.3.1, 8.3.1), but here is a brief preview.

Split-brain cases have indeed shown that the sensorimotor systems of each hemisphere, along with higher-level cognitive functions (e.g. language in the left hemisphere, perceptual distinctions in the right hemisphere), can operate independently of the other hemisphere (Gazzaniga, 2000). However, split-brain patients retain a good deal of bilateral co-ordination of movement, so it is not the case that their brains have been *entirely* bisected. Furthermore, there does not seem to be an overall bisection of 'their general cognitive awareness, affect and sense of self' (p. 1309).

Of course, the general cognitive awareness and sense of self of split-brain patients is faulty to the extent that their self-monitoring system is interrupted by the severed corpus callosum: these patients cannot verbally report on, and seem to have no conscious awareness of, stimuli received only by the right cerebral hemisphere (as happens mostly in tightly controlled experimental conditions). But this does not suggest that each hemisphere operates a segregated self-monitoring system. Rather, it means that the self-monitoring system serving the whole organism is deficient in that it does not receive input from the right hemisphere for verbal processing. Thus, rather than producing two selves, the split-brain condition results in a somewhat deficient self.

If an acquired neural bisection like commissurotomy does not, on the system view, produce more than one self, it is even less likely that conditions without a discernible neural disconnection should produce multiple selves. This concerns particularly the condition known variously as mul-

tiple personality disorder (MPD) and dissociative identity disorder (DID). This disorder too will be the subject of a lengthier discussion later (§ 8.3.3). For now, the supposition presents itself that, in so far as those afflicted with MPD/DID can be said to have various distinct 'personalities' or 'identities', these might well be different interpretative artefacts by the self-monitoring system, which for whatever reason produces multiple partial histories of the self. But those different personal histories are still outputs of the same self-monitoring system serving the same single organism—i.e. of a single self. A complex system like the self can develop faults and fractures. This does not entail that in so doing it duplicates or multiplies.

1.8 Synopsis of following chapters

To defend the system view of the self that I have sketched here, it will be useful to consider its main rival: historical-constructionist accounts. In *Chapter 2*, I discuss Locke's (1690/1706) account of the self and of personal identity and its neo-Lockean variations. I also note that there is an important distinction to be drawn between *person* and *self*, a distinction that is inchoate but not developed in Locke, and has been much overlooked since. *Chapter 3* considers how different thought experiments trigger diverging intuitions about the self, but how recent experimental philosophy studies seem to suggest that quasi-Lockean intuitions are particularly robust.

Thought experiments and experimental philosophy vignettes triggering Lockean intuitions usually appeal to the loss or preservation of memory. Thus, *Chapter 4* is devoted to autobiographical memory, which I show to be complex and subject to various constraints upon its availability and accessibility, and its reliability. Given the reconstructive nature of episodic recollection in particular, I argue that the self is the constructor of, not a construct from, autobiographical memories. In *Chapter 5*, I consider how narrative practices contribute to autobiographical memory. I then review a number of narrative accounts of the self, which divide into four positions: strong and weak narrative constructionism, and essential and simple narrative-capacity views. In *Chapter 6*, I argue first against *strong* narrative constructionism. Identifying the authorship problem of even weak narrative constructionism, I then present an empirical and a metaphysical argument

against *all* narrative constructionist views. Finally, I argue against *essential* narrative-capacity views (leaving us with a simple narrative-capacity view).

The topic of *Chapter 7* is autism spectrum conditions: why they are relevant to discussions of the self, and how the self in autism is affected (or unaffected) by the specific deficits that have been identified with autism spectrum conditions. In *Chapter 8* I look at other defects and disorders, of memory (lack of episodic memory, amnesias, dementia) and of the self (split brains, bipolar disorder, multiple personality, and schizophrenia), and suggest that we should regards these defects and disorders as specific malfunctions of the self-representational system.

Chapter 9 contains a brief review of the thesis, a further discussion of how the self system may be characterized, and suggested directions for future research.

Chapter Two

Self and person in Locke

2.1 Introduction

In treating of the self, one cannot avoid John Locke. He is one of the first to use 'self' as a noun, with a specific metaphysical meaning, in the English language.[1] And his account of 'personal identity' in the *Essay* (1690/1706/1997, II.xxvii) is the classic example of an historical-constructionist view of the self: it locates the criterion for selfhood in the continuity of consciousness over time, regardless of how or where such consciousness is instantiated. There is something both right and wrong with this approach. Locke is right in taking the self to be a *functional entity*, defined by what it does, rather than following his precursors, like Descartes (1641), in assuming it to be something ontologically basic, like a substance in its own right. Where Locke goes wrong is in the specific criteria of what makes something the same self over time, for he takes *evidential* criteria for the continuity of the self ('same consciousness') to be *constitutive* of what a self is ('consciousness . . . constitutes . . . self', II.xxvii.17). It is from this unfortunate confusion that historical-constructionist accounts of the self originate, which take the self to be constituted by, or *constructed* from, one's consciousness of one's own past: one's autobiographical memories. In §§ 2.2 and 2.3, I will discuss Locke's account of the self, along with some common objections to, and developments of, his view, and note how the problems faced by Locke's account affect historical-constructionist views more generally.

---

[1] The *OED* (Oxford University Press, 2015) cites two earlier uses of 'the self' in the philosophical sense, but their authors are obscure, and the philosophical nature of the first attestation is doubtful.

A connected and somewhat overlooked issue with Locke's account—but of interest beyond it—is the relation between self and person. In much of II.xxvii, Locke seems to be using 'self' and 'person' interchangeably. And while the term 'Lockean person', designating a person in Locke's sense, is a staple of philosophical discussions of personal identity and ontology (e.g. Parfit, 2012), talk of a 'Lockean self' is much less common. If a Lockean self simply is the same as a Lockean person, this matters little. But, although Locke himself seems to encourage a conflation of person with self, there are hints in his treatment that the terms are not wholly coextensive even in their Lockean senses. And nor, in a general sense, should they be. For a person has characteristics, rooted in the social context in which it is expedient to talk of persons, that are not required of a self. This distinction between person and self will be the topic of my discussion in § 2.4: I will explain in general terms why it is an important distinction and why ignoring it is at the root of much that is wrong with historical-constructionist accounts of the self. But I will also point out that Locke himself vacillates somewhat between the synonymous and the distinctive use of 'person' and 'self'. This is not surprising, since the self–person distinction is ultimately informed by Locke's own conceptualization of persons: it is the 'forensic' aspect of the Lockean person that does not necessarily apply to the self. Here, again, there is an element in Locke's account that is worth preserving—his view of the nature of persons—though Locke himself does not follow it through to its conclusion—that self and person do need to be distinguished.

But let me begin with a summary of Locke's account.

2.2 Locke on personal identity and the self

Locke begins his disquisitions on identity uncontroversially by giving a standard definition of what philosophers now usually call diachronic identity, i.e. identity over time, as something's being one and the same thing at different times. But he notes that the principle of individuation, and thus the conditions of identity, of things depend on what sort of thing is being talked about: simple or compound, 'mass of matter' or 'living body' (1690/1706, II.xxvii.3). In the case of compounds, particularly living organisms, diachronic identity is not a question of a thing's retaining pre-

cisely the same particles of matter over time. Plants and animals, as well as machines, remain the same by preserving a 'fit organization of parts' that make them a particular plant, animal, or machine, rather than by preserving each of the parts. He thus allows, in keeping with common sense, that a complex thing can be the same thing over time even while its constituent parts change.

In discussing the identity of a man (or, we ought to add, a woman), Locke does not yet depart from the principles that apply to the identity of plants, animals, and machines. Here too identity consists 'in nothing but a participation of the same continued life . . . like that of other animals in one fitly organized body' (II.xxvii.6). Though men (and, we should add, women) are generally endowed with reason, that in itself is not the quality that decides their identity over time, which is rather a question of 'shape and make':

> 'tis not the idea of a thinking or rational being alone, that makes the idea of a man in most people's sense; but of a body, so and so shaped, joined to it. (II.xxvii.8)

Locke supports this point with the first of many thought experiments[2] in this chapter of the *Essay*: a rational parrot would still be thought a parrot, a dull man still a man. Thus, a man could lose all his reason and still, satisfying the conditions of having 'the same continued life' in the same body 'so and so shaped', remain the same man, that is, the same human organism.

One might leave it there, and say that the same conditions apply to the identity of a *person*. Locke, however, does not, and here provides us with an important distinction that has bedevilled the philosophy of personal identity ever since: that between a human being and a person. A person is something other than a man, and consequently the identity conditions for a person are different from those that apply to a human being generally. For here is where the qualities of rationality and thought come to matter. A person, according to Locke, is

> a thinking intelligent being, that has reason and reflection, and can consider itself as itself, the same thinking thing in different times and places; which it does only by that consciousness, which is inseparable from thinking . . . and as far as this consciousness can be extended backwards to any

[2] I discuss Locke's—and others'—use of thought experiments in ch. 3.

> past action or thought, so far reaches the identity of that *person*; it is the same *self* now it was then; and 'tis by the same *self* with this present one that now reflects on it, that that action was done.
>
> (II.xxvii.9; original emphasis)

This passage is crucial in two ways. First, it stipulates the identity conditions for Lockean persons: I will discuss the consequences of these presently. Secondly, Locke here employs the terms 'person' and 'self' in the same breath. But what is the self?

> *Self* is that conscious thinking thing, (whatever substance made up of whether spiritual, or material, simple, or compounded, it matters not) which is sensible, or conscious of pleasure and pain, capable of happiness or misery, and so is concerned for itself, as far as that consciousness extends. (II.xxvii.17)

We here have a *functional* definition of 'self' which—taken, for the moment, without the ensuing identification of the self with the person—is not inconsistent with my system view of the self. The Lockean self is something conscious of itself and concerned for itself: these are important higher-level self-representational processes among those I have named in the previous chapter as defining the self as a complex system. Locke's agnosticism about the precise ontology of the self probably owes less to 17th-century limitations in scientific knowledge than to Locke's desire to eschew contemporary metaphysical debates between materialists and immaterialists (Boeker, forthcoming). Having defined the self in functional terms, Locke has no need to commit himself to a particular ontological position.

He does, however, in the remainder of the paragraph just quoted, equate 'self' with 'person'. Let us return, then, to the identity conditions of Lockean persons (and thus of Lockean selves).

Why do persons have identity conditions different from those of a human being? It is because, in Locke's view, 'person' is a 'forensic term'; it 'belongs only to intelligent agents capable of a law' (II.xxvii.26). When we refer to someone as a person, we ascribe to him or her qualities that go beyond the continued life of a body that would be sufficient for someone to remain (just) the same human being; qualities which allow a person to be held *responsible* for his or her actions. Thus, the 'forensic' nature of the term 'person' refers to the practice of holding someone to account for what she does or has done. Thus, a Lockean person, too, is a *functional kind*.

The conditions for personhood, then, must include whatever is necessary for someone to be ascribed responsibility for her actions: in Locke's view, rational intelligence and consciousness. Rational intelligence is needed, presumably, because responsibility for one's actions implies that one acts according to reason and can give a rational account or explanation of how an action connects with a goal or motivation, so satisfying the 'forensic' nature of responsibility-taking. And consciousness is required because it seems inequitable to reward or punish someone for an act performed unconsciously (II.xxvii.19).[3]

It now becomes clear why the diachronic identity of persons must, on Locke's account, involve the 'same consciousness' being present at any point in time at which some person is the *same* person.[4] For if one had lost all consciousness of a previous action, even if that action had at the time been carried out consciously (and intelligently and rationally) it would not, Locke maintains, be appropriate to apportion praise or blame for that action to that man now:

> But if it be possible for the same man to have distinct incommunicable consciousness at different times, it is past doubt the same man would at different times make different persons; . . . human laws not punishing the mad man for the sober man's actions, nor the sober man for what the mad man did, thereby making them two persons . . . (II.xxvii.20)

Same consciousness is thus an absolutely *necessary* condition for diachronic personal identity—but is it also a *sufficient* condition? Locke thinks so: the *substance* (be it a material body or an immaterial soul) that may at any time be associated with a particular person is irrelevant to that *person*'s identity over time, which is wholly determined by consciousness (II.xxvii.10–16).

If, then, consciousness is both a necessary and a sufficient condition for personal identity, the identity conditions of persons and those of human beings no longer overlap, and their identities can therefore come apart. One Lockean man, having completely lost all consciousness of his former actions before a certain point in time, and from then on developed a new con-

---

[3] But Locke's account also seems to imply, more controversially, that one cannot be held responsible for actions of which one has lost all recollection.

[4] Quite what Locke means by 'consciousness'—and therefore quite what it is that is 'the same' when the personal-identity condition obtains—is somewhat debatable (see § 2.3.3 below).

sciousness of all that he had experienced and done since, would be two (consecutive) Lockean persons (or selves) (II.xxvii.21). Conversely, it is conceivable (and a much used trope in certain genres of science-fiction literature) for someone's consciousness to be transferred from one man to another, or even from one kind of substance to another.

The potential coming apart of the identities of a man and a person do not, of course, accord with our ordinary experience (though it might in situations when someone is '*besides himself*' (II.xxvii.20)), but they do follow from Locke's commitments on what constitutes the conditions for personhood. One can lose the characteristics of personhood—one's consciousness of past actions, one's reason—and stay the same human being. It is in the context of this distinction that the most commonly cited objections to Locke's account arise.

2.3 Objections and interpretations

2.3.1 Loss of transitivity

A first criticism of Locke's account that has repeatedly been made is that his notion of personal identity does not preserve transitivity. As a formal relation, identity is transitive: if A = B and B = C, then A = C. For Lockean personal identity, this is not the case, as Thomas Reid (1785, § 3.6) points out in his now canonical counterexample of the 'brave officer', who as a young officer remembers being flogged for stealing apples when a schoolboy, and who when later made a general recalls his exploits as a young officer, but no longer remembers the schoolboy flogging. According to Locke's account, this makes the old general the same person as the young officer, and the young officer the same person as the delinquent schoolboy, but not the old general the same person as the schoolboy—thereby violating the transitivity requirement.

Several kinds of reply are possible here. The first is to modify Locke's account so that transitivity is preserved. John Mackie (1976) attempts this 'by taking the unit of consciousness to be determined not by the relation *could remember*, but by its ancestral—that is, by the relation which is to *could remember* as *ancestor* is to *parent*.' (p. 180). That is to say, there is a continuity

relation—the 'ancestral' of the memory relation—between between the young officer and the old general, and between the schoolboy and the young officer, but also transitively between the schoolboy and the old general. This, however, as Mackie admits, is 'a revision, not an interpretation, of Locke's account' (p. 181), and directly contradicts Locke's view in II.xxvii.20 that irretrievable memory loss indeed severs the link of personal identity.

The other possible reply to the loss of transitivity is to bite the bullet. One way of doing this is to eschew the formal identity relation altogether, as Derek Parfit (1984) does when he insists that 'identity is not what matters' (*passim*). What matters is the *continuity* of persons, their *survival*. In Parfit's account this is assured by *psychological* continuity having '*any* cause' (p. 208). Parfit's notion of psychological continuity here bears some resemblance to the ancestral relation of Mackie's revision. Both go beyond what Locke means by 'consciousness' (more on this in § 2.3.3, below). But the Parfitian continuity/survival relation does not rule out the fission or fusion of persons. The same applies to Lewis's (1976) neo-Lockean account that takes persons to be four-dimensional entities.

Arguably, Locke's own view does not rule out that persons could fissure or be fused, so long as the 'same consciousness' relation obtained between their pre- and post-fissure or -fusion stages. But it would be semantically difficult to defend the consequence, given that a Lockean *person* is a Lockean *self*, that a (post-fusion) self at one time would be the *same* self as *two* selves at a previous time, or that a (pre-fission) self at one time would be the same self as two selves at a later time. It is worth reminding ourselves that such counterintuitive—because empirically highly unlikely—scenarios arise only if we consider the continuity of a person or self as ontologically separable from his or her biological organism—that is, if the identity conditions of a person are wholly independent of those of the entity that happens to have the characteristics of being a particular person. But all historical-constructionist accounts in which the person or self is not anchored to a particular organism must admit of such scenarios—whereas the system view, as explained in the opening chapter, avoids such bizarre consequences, since a self subsists only in virtue of the continuing existence of an active brain—and it is extremely unlikely for living brains to become

fissured or fused.[5] And anchoring the self in the continuing brain, rather than its productions, avoids problems of transitivity altogether.

Meanwhile, a more moderate way of accepting the loss of transitivity is to take it to be the very point Locke is trying to make. In Galen Strawson's (2011) view, contradicting the transitivity principle

> isn't an objection to Locke's view of personal identity. It is, rather, an illustration of its fundamental and forensic point, the commonsense point . . . that human beings won't on the Day of Judgment be responsible for all the things they have done in their lives, but only for those that they're still Conscious of and so still Concerned in. (p. 54)[6]

This echoes Mackie's (1976) verdict that Locke's theory is 'hardly a theory of personal identity at all, but . . . a theory of action appropriation' (p. 183). *Person*, as a 'forensic term', is not about formal identity, but about responsibility. But, if so, we must ask whether this applies to the Lockean *self*, too. I address this point in § 2.4.

2.3.2 Circularity?

Perhaps the most notorious objection to Locke's account, most often associated with Bishop Butler (1736),[7] is that it is circular:

> And one should really think it Self-evident, that Consciousness of personal Identity presupposes, and therefore cannot constitute, personal Identity, any more than Knowledge in any other Case, can constitute Truth, which it presupposes. (p. 302)

The way Butler frames this makes it seem a sound, even devastating, objection. If Locke's view were indeed that personal identity is constituted by consciousness *of* personal identity, it would be obviously circular. But that is not Locke's account. Personal identity is not constituted by consciousness

---

[5] The so-called 'split-brain' condition does not amount to a complete fissure of the brain (see § 8.3.1).

[6] Strawson uses initial capitals in 'Consciousness' and 'Concern' to indicate that they are to be understood in Locke's particular sense of these words.

[7] A similar objection is also found in Berkeley (1732) and Reid (1785). As for Butler, 'the objection to Locke for which he is well known is not his, having been put by John Sergeant in 1697 and Henry Lee in 1702, among others' (Strawson, 2011, p. 2).

of itself, as Butler's lazy reading of Locke would have it, but by consciousness *of past actions and thoughts*. There is thus no obvious circularity here.

At the same time, if we were to ask *whose* consciousness it is that constitutes personal identity over time, we cannot answer 'the person's' without avoiding circularity. As Mackie (1976) puts it:

> Yet behind [Butler's] unsound criticism, there is an element of truth. We are reluctant to believe that our identity through time is constituted by this sort of memory, and are more inclined to regard the memory as evidence for an identity which is already there, constituted by something else and somehow making that memory possible. (p. 187)

Mackie here echoes Hume's (1739) thought that 'memory does not so much *produce* as *discover* personal identity' (I.iv.6). In other words, memory is an *evidential* rather than a *constitutive* criterion of identity. So, while Locke does not make the formal mistake he stands accused of, his account does seem to presuppose that *some* continuant entity underpins (enables, subserves) the possibility of consciousness of past actions being preserved and repeatable.

Another way of making Mackie's point is to say that 'it seems to belong to the very idea of remembering that you can remember only your own experiences' (Olson, 2015). This being so, even taking memories as just evidence (rather than constitutive) of personal identity is 'trivial and uninformative' (ibid.): what I can genuinely remember are necessarily my own experiences. Therefore, to say that my remembering something of my past makes me myself has no explanatory purchase.

Neo-Lockeans have countered this objection by introducing the notion of 'quasi-memory'. Sydney Shoemaker (1970) defines quasi-memory as

> knowledge of past events such that someone's having this sort of knowledge of an event does involve there being a correspondence between his present cognitive state and a past cognitive and sensory state that was of the event, but such that this correspondence . . . does not necessarily involve that past state's having been a state of the very same person who subsequently has the knowledge. (p. 271)

The idea of quasi-remembering does not, therefore, necessarily require the identity of the rememberer with the person having the experience being remembered. How does this help? It helps insofar as it removes the circularity objection. But, in exchange for that, it leaves us with a purely theoretical notion of 'quasi-remembering' whose only actual instances that feature

in ordinary experience are the 'special case' (ibid.) of *actual* remembering. Shoemaker concludes:

> In the actual world it is true both that (1) [quasi-remembering] is always [actual] remembering . . . and that (2) the primary focus of a person's 'self-interested' attitudes and emotions is his own past and future history. It is surely no accident that (1) and (2) go together. (p. 285)

Thus, while memory, in Shoemaker's account, is no longer a *criterion* for personal identity, it clearly *matters* to one's self-interested concerns. And, of course, remembering is an important high-level self-representational activity. (I will discuss personal memory in detail in ch. 4.)

Before moving on, it is worth noting that variations of the circularity objection apply to historical-constructionist accounts of the self more generally. For any account that takes the self to be constructed from elements of one's autobiography (memories, self narratives) faces the question who or what is doing the construction (the remembering, the narrating). I will return to this point in Chapters 4 and 6. For now, let us return to Locke.

2.3.3 Consciousness, memory, psychological continuity

Commentators on Locke, at least since Reid (1785), have summed up his account of personal identity by ascribing to Locke the view that one's person or self is constituted by one's *memory* of past actions. I have in the foregoing discussion followed this practice. But this is something of an oversimplification. In his II.xxviii, Locke repeatedly names *consciousness* as what assures ('makes') personal identity—not memory. H. P. Grice (1941) suggests that we may interpret '"consciousness" as meaning "memory", or "memory or introspection"' (p. 341). More recently, Marya Schechtman (1996) and Galen Strawson (2011) have taken a different view, claiming that Locke's 'consciousness' precisely does not mean 'memory'. It may be useful, therefore, to try and shed some light on the relation between Lockean 'consciousness' and memory.

Of course, consciousness of one's past actions *requires* memory. As Locke himself says explicitly in II.xxvii.20, the complete loss of one's memory of a past action wipes that action out of one's conscious personal history. So we are right in taking memory to be a necessary condition for the consciousness that makes a Lockean person. But memory being neces-

sary does not mean that it is also a *sufficient* condition for the 'same consciousness' that for Locke assures the same self:

> For as far as any intelligent being can *repeat the idea of any past action with the same consciousness it had of it at first*, and with the same consciousness it has of any present action; so far it is the same *personal self*.
>
> (Locke, 1690, II.xxxvii.10; first emphasis added)

Locke here effectively requires us to *relive* our past actions 'with the same consciousness' which we had of them at the time. A mere *factual* memory of a past action is not enough; it would be wrong to assume that 'repeating' past actions is merely a demand on the accuracy of our memory. It is also, perhaps primarily, a requirement that one be *affectively concerned* in the action in question in the same way as when one first performed it (Schechtman, 1996; Strawson, 2011).[8]

Unfortunately, Locke himself does not elaborate on what he means by 'consciousness'. As Mackie (1976, p. 178) observes, Locke sometimes uses 'consciousness' as a mass noun, and at other times as a countable noun, as when he speaks of 'distinct incommunicable consciousness' (Locke, 1690, II.xxvii.23). Some tidying-up is required here.

Mackie suggests that Locke's use of 'consciousness' as a concrete noun stands for the fairly commonsensical notion of 'an entity consisting of someone's being conscious of a number of actions and experiences together' (1976, p. 178). Thus, someone's consciousness amounts to this someone 'having a series of co-conscious experiences' (ibid.). One might then be tempted to suppose that there are 'well-defined units of consciousness, distinct consciousnesses, that is, separate mental histories, such that there was perfect co-consciousness within any unit and no co-consciousness between units' (ibid.). But, as Reid's brave-officer case illustrates, there are no such well-defined and unchanging mental histories. Mackie's reply to Reid was that we should take the 'unit of consciousness to be determined . . . by [the]

---

[8] It might be said that the 'same consciousness' requirement could be met by what has become known as 'episodic memory', since such recollections are said to involve 'autonoetic' awareness, i.e. awareness of oneself in the event being recalled as though one were experiencing it in the present (Tulving, 1972; 1985). But then, Butler's circularity objection does apply, for we would have thus redefined consciousness as consciousness of the same self. (For a discussion of episodic memory and autonoetic awareness, see § 4.2.2 below.)

ancestral [of] the relation *could remember*' (p. 180). He then explicates this point by developing it into a full-blown revision of Locke's account.

Mackie proposes that we should think of one's personal history as a series of 'I-occurrences'. Each moment of consciousness is such an I-occurrence, which 'fills some short stretch of time', so that successive I-occurrences during any continuous period of consciousness 'overlap and fade into one another'. And it is these overlapping I-occurrences which 'constitute' one's identity 'throughout any one waking day' (ibid.).

But of course we are not continuously conscious throughout our lives, having nightly periods of unconsciousness that amount to as much as a third of our 'continued life', and perhaps not even throughout all of any waking day, during which the merest 'drowsy nod' (Locke, 1690, II.i.13) may disrupt the flow of overlapping I-occurrences. These interruptions, Mackie suggests, are bridged by memory. There then emerges an elegant, but not quite Lockean, account of what a person is:

> A network of overlapping specious presents and day-to-day memory bridges builds up what we can now take to be a single consciousness: we can thus generate a relation *is the same person as* and another *belongs to the same person as* which are both transitive and symmetrical . . .
>
> (Mackie, 1976, p. 181)

Mackie thus not only restores transitivity; he also accounts for both consciousness and memory—although he does so at the cost of revising, rather than interpreting Locke. Mackie gives an account of personal identity proper, rather than personal 'identity' in the 'forensic' sense in which Locke was interested (see § 2.4).

Mackie's interpretation is similar to Shoemaker's (1984) account of personal identity as '*psychological continuity*'. Such continuity requires neither 'same consciousness' nor an uninterrupted memory link from any past moment to the present; rather, it is sufficient for there to be an 'appropriate' causal chain from past psychological states to one's present psychological states. Relevant states include memories, but also traits, preferences, beliefs, and other cognitive states. Two such states can be directly psychologically *connected* (as in Reid's brave officer recalling his boyhood), and any two states related by a chain of such psychological connections, even if not directly connected, are then psychologically *continuous* (as in the officer's boyhood and his dotage). Shoemaker thus avoids problems of both transitivity

and circularity. Further, his account is designed to fit with a materialist metaphysics.

This too is a common factor with Mackie's revision of Locke, which in this respect is another departure from Locke's own account: Mackie suggests that when we come to try and give a *factual* account of how co-consciousness and memory bridges work, we need not stipulate any immaterial and/or unknowable substrate: 'what makes co-consciousness [and memory] possible is the structure of the central nervous system and the persistence of that structure through time' (1976, p. 200). This, of course, comes close to my system view: but here, the question of identity is already answered by the persistence of the neurocognitive self-monitoring system. And psychological continuity is a product of the operations of that system. It provides the appropriate causal mechanism for my psychological states at one time to be continuous with those at another time.

2.4 The person and the self

I come now to a crucial (and, I think, hitherto overlooked) point in this discussion of Locke's (1690/1706) II.xxvii, the question of the relation between 'person' and 'self'. But before I analyse Locke's own use of the terms, some more general remarks on the concept of a person and its relation to the self are in order.

2.4.1 Personhood (and why it is not the same as the self)

Questions about personhood are a wide-ranging philosophical topic in their own right. It is not the topic of my thesis, but, given the unfortunate conflation of 'person' and 'self' since Locke, I must at least explain *why* it is not my topic—that is, why being a person is a quality, or set of qualities, over and above those of having a self. To that end, I will not attempt to settle what the best definition of a person is. For present purposes, a brief survey of some common approaches to the question of personhood will be sufficient.

Daniel Dennett (1976) lists 'six familiar themes, each a claim to identify a necessary condition for personhood' (p. 177): being a *rational being*; ascrip-

tion of *psychological states*; having 'in some way an *attitude taken* toward it' (ibid.); a capacity to *reciprocate*; a capacity for *verbal communication*; and '*self*-consciousness of one sort or another' (p. 178). Note that all these criteria seem to require some kind of self-monitoring. Additionally, attitude-taking and reciprocity—the essential criteria for personhood in P. F. Strawson's (1962) account—and communication are *social* capacities.

An oft-cited account stipulating some psychological capacity as requisite for personhood is Harry Frankfurt's (1971). A person, in Frankfurt's view, must have second-order volitions, that is, desires *about* her desires. She is able to structure and control her volitions in ways that satisfy some overarching goals, rather than acting impulsively on every desire. She may want to enjoy the sunshine by taking a stroll in the park, but, deploying a second-order desire to do well at her job, she modifies her first-order volition, forgoes the walk, and gets on with some time-critical paperwork. Her second-order volition is that she *wants* herself to want to stay at her desk. Meanwhile, an agent who simply acts on his first-order volitions is not a person on this account, but a 'wanton'. Clearly, a Frankfurtian person requires the higher-level self-monitoring necessary for impulse control and deferred gratification. But even a Frankfurtian 'wanton' engages in self-monitoring to the extent that he knows what his desires are, and how to satisfy them. A self-monitoring system is thus absolutely necessary for Frankfurt's account of a person, but more than that: it must be exercised in a particular way. Just *having* a self-monitoring system is not sufficient for Frankfurtian personhood.

Frankfurt's personhood criterion is perhaps too stringent. At any rate, Tom Beauchamp (1999) notes that any definition of 'person' in terms of cognitive capacities is bound to fail in borderline cases. But if we look elsewhere for the criteria of personhood, we similarly find that characteristics over and above having a self are necessary. John Doris (2009) points out that '[t]alk of persons involves both descriptive and normative elements'; the latter 'mark a network of ethical expectations' (p. 58). Satisfying such expectations may well involve Strawsonian reciprocal attitude-taking; in any case they require a capacity for representing *others* as well as oneself. While there is some evidence from cognitive neuroscience that self- and other-representations engage the same neural networks (see § 7.3.1), their object is obviously not the same. It is entirely conceivable that a being could

have a functioning, even highly developed, self-representational system yet fail to respond to and understand others in ways that satisfy the social-normative expectations involved in person-talk. This may be the case for people with autism (see ch. 7).

To sum up: a functioning self-monitoring system—a self—is one of the necessary conditions for personhood,[9] but on no account is it a sufficient one. One may have a self and be a Frankfurtian wanton. One may be competent at self-monitoring but incapable of communication and therefore unable to fulfil the normative expectations surrounding persons. Or one may be capable of communication and still lack a sensibility for the normative framework of personhood, as may be the case with psychopaths. Furthermore, personhood, whatever its precise criteria, is a property or set of properties one can acquire, have, and lose—notwithstanding the continuity of one's organism (Olson, 1997) and its self-monitoring system. The grievous error of much of the personal-identity literature has been to focus entirely on what assures the continuity of this property of personhood, however defined, without first establishing what ensures the continuity of the self. In deciding what makes someone the same *person* over time—given the social-normative considerations inherent in the concept of personhood, such as responsibility for one's actions—the historical-constructionist approach makes some sense. The mistake is then to equate the social-normative history of the person with the continuity of the self.

Much of the blame for this person–self conflation must be laid at Locke's door. But, at the same time, the roots of the self–person distinction I have just outlined are, I shall now argue, to be found in Locke's account as well.

2.4.2 'Person' and 'self' in Locke

Locke seems to use the terms 'person' and 'self' synonymously and interchangeably. Thus, his definition of 'self' in II.xxvii.17 closely mirrors that of 'person' in II.xxvii.9, and in both passages, the terms appear as comple-

---

[9] *Legal* persons (companies etc.) are an exception to this, although arguably a collective entity that qualifies for legal personhood will have management and auditing procedures that are functionally equivalent to self-monitoring in an organic being.

mentary: 'so far reaches the identity of that *person*; it is the same *self* now it was then' (II.xxvii.9; original emphasis); 'consciousness . . . makes the same *person*, and constitutes this inseparable *self*' (II.xxvii.17; original emphasis). In these passages, Locke may seem to be saying that what he is talking about goes by two common nouns, 'person' and 'self'. They may to some extent explicate each other, hence his juxtaposing both terms in this way; but overall they refer to the same thing and can be used interchangeably (cf. Strawson 2011, ch. 8).

But there are other passages that do not quite tally with this reading. Thus, Locke also uses 'personal' as a *qualifier* for 'self' (II.xxvii.10). Are we to understand 'personal self' as a particular *kind* of self, or is the qualifier redundant? If the latter, why should Locke use the epithet 'personal' at all? If a self is the same as a person, a 'personal self' is a 'personal person'. It seems unlikely that Locke would deliberately construct such a meaningless expression.

Further, if Locke's use of 'self' and 'person' is indeed intended to be synonymous, it is remarkable, given the centrality of both terms to his argument, that he doesn't explicitly tell us this. The closest he comes to explicating the relation between the terms 'self' and 'person' directly, in what may well be the most important section in the chapter, is in saying precisely not that the person *is* the self, but: '*Person*, I take it, is the **name** for this *self*.' (II.xxvii.26; original italics; bold emphasis added)

Locke would not put it this way without good reason, and the reason for putting it this way is given soon after: *person* 'is a forensic term appropriating actions and their merit' (ibid.). It is thus in the 'forensic' context of action-appropriation that the self is labelled 'person'. There are, I think, two related but subtly different readings of this available. The first reading is that while 'self' and 'person' refer to the same thing, the use of each label is determined by different perspectives: 'self' is an *inward-looking*, reflexive term, which picks out the self 'from the inside',[10] while 'person' is a name for the self given the forensic context of a person being judged, as it were, from the *outside*. Thus, 'person' and 'self' are co-extensive, but not synonymous. This reading seems directly supported by the sentence that intervenes between Locke's definitions of 'person' as 'a name' for the self and

[10] This term of Shoemaker's (1970) is used by both Mackie (1976) and Strawson (2011) to explicate the Lockean notion of 'consciousness'.

as a 'forensic term', where he says: 'Wherever a man finds what he calls *himself*, there I think **another** may say is the *same person*' (ibid.; original italics; bold emphasis added). Thus the difference in use between 'self' and 'person' is explained by whether one is looking at oneself (one's *own* person) or at another person.

But there is a second possible reading of the difference, which while less explicit is nevertheless suggested by what Locke says about the characteristics of persons here and elsewhere in the chapter. It is that there may be Lockean selves that do not qualify as Lockean persons. The giving of the 'name' *person* to a self, on this reading, is itself a kind of forensic act, an attitude-taking, one which recognizes the self so named as one that is capable of appropriating its own actions so as to satisfy the moral and legal responsibility that is contained in Locke's notion of a person.

Locke's singly necessary and jointly sufficient conditions for something to be a *person*[11] are that it be 'thinking', 'intelligent', that it have 'reason and reflection', 'can consider itself as itself, the same thinking thing, in different times and places' (II.xxvii.9). Explicitly defining the *self*, Locke says it is 'that conscious thinking thing, . . . which is sensible, or conscious of pleasure and pain, capable of happiness or misery, and so is concerned for it*self*, as far as that consciousness extends.' (II.xxvii.17) Finally, in defining 'person' as a 'forensic term', Locke goes on to say that this term '**belongs only to** intelligent **agents capable of a law**, and happiness and misery' (II.xxvi. 26; bold emphasis added). We notice that being 'capable of happiness and misery', previously said of the self, recurs here in Locke's second definition of 'person', thus underpinning the close relation between the two terms. But 'capable of a law' is an addition that only occurs here. It explicates the forensic nature of Locke's notion of 'person'.

Self, on the other hand, is clearly not a forensic term. It is quite conceivable that there are selves in the Lockean sense that are 'thinking' and 'sensible, or conscious of pleasure and pain, capable of happiness or misery' while lacking the action-appropriation condition of a Lockean person. Children before the age of accepting responsibility for their doings seem an obvious example.

---

[11] As noted in § 2.4.1, conditions for personhood are notoriously hard to define. Locke, at least, has a pretty good stab at it.

And this is where Mackie's (1976) revision makes sense. By his own admission, his account is not one of a person in the Lockean forensic sense, but of something 'corresponding to all the personal pronouns' (p. 183). But in this it also corresponds to Locke's definition of a 'self', and most of his definition of a 'person'—with the crucial exception of the forensic condition. If we follow Locke in reserving 'person' for its forensic use, we can still adopt Mackie's revision as an account of the Lockean *self*—the transitivity of whose identity over time, as has been mentioned, is preserved.

Locke himself, despite the hints I have exposed here, does not explicitly draw a distinction between *self* and *person*, and most readings of Locke also fail to do so. But since his account of personal identity emphasizes the question of legal and moral *responsibility* of persons, the conflation of person and self is unfortunate. Had Locke been a little clearer on this point, we might not now be confronted with a vast philosophical literature that takes the continuity of the characteristics of personhood, however defined, as the main issue in determining the continuity of the sort of thinking beings that we are.

2.5 Conclusions

The point of this chapter has been to highlight two rights and two wrongs with Locke's account of personal identity and the self. He is right in giving a functional account of the self as something 'concerned for itself'. This requires the higher-level self-monitoring involved in awareness of one's own past and making projects for one's future. But such temporally extended self-consciousness is a *product* of the self: Locke is wrong in holding that this consciousness '*makes*' the self. This historical-constructionist strategy faces the difficulties with transitivity and circularity discussed earlier, but, more important, it leaves open the possibility of branching selves and successive discontinuous selves in the same organism. The system view, taking the self to be the neurocognitive system that *enables* our self-consciousness over time, rather than that consciousness itself, avoids these problems.

There is more than a hint towards the end of his chapter on personal identity that Locke (1690/1706), though studiedly agnostic on the question of materialism, realized that knowledge of the substrate of our cognitive capacities might well render some of his discussion superfluous:

> Did we know what it [the nature of that thinking thing that is in us] was, or how it was tied to a certain system of fleeting animal spirits; or whether it could, or could not perform its operations of thinking and memory out of a body organized as ours is; and whether it has pleased God, that no one such spirit shall ever be united to any but one such body, upon the right constitution of whose organs its memory should depend, we might see the absurdity of some of those suppositions I have made. (II.xxvii.27)

Locke's own account does not guarantee that there should be a single self per human being. But if we take the continuing brain to be 'the right constitution of organs' on which the self depends, we can put aside quibbles about successive and fissured selves.

Locke's ontological agnosticism has a purpose, however. For his main concern, as Mackie (1976), Schechtman (1996), and Strawson (2011) point out, is the accountability of *persons* for their actions. Here the historical-constructionist strategy makes sense: in the social-normative arena, our doings are evaluated by what reasons we proffer, that is, by the continuity we construct from our motives to the results of our actions. The second right element of Locke's account is his characterization of persons as 'intelligent agents capable of a law', the recognition that what it is to be a person is conditional on the social-normative framework in which we speak of persons and take the relevant attitudes of commendation and blame towards one another.

What is wrong with this part of Locke's account is the conflation he encourages of 'person' in his sense with 'self'. For, as I have shown, there can be selves without the attributes of personhood. It is the lack of a clear distinction between person and self in Locke that seems to have led many of his followers to emphasize the continuity of personal attributes above all else, without first considering the continuity of the self. And so the criteria for the continuity of persons, wrongly, become criteria for the continuity of selves.

Yet, even stripped of the attributes of personhood, it may seem that the continuity of the self is determined by the 'same consciousness' that Locke marks as its criterion. In the next chapter, I will examine whether and how the notion that the self depends on memory and psychological continuity may be intuitively appealing.

Chapter Three

Thought experiments and experimental philosophy

3.1 Introduction

One explanation for the pertinacity of the Lockean view of the self and its historical-constructionist descendants is that there may be something intuitively plausible about the idea that one is *oneself* only so long as one is in some way psychologically continuous with one's personal history. Colloquial expressions suggest that some such intuition may be widespread among ordinary people. One may say of a relative suffering from dementia that 'she is no longer herself'. An exculpatory 'I forgot myself!' may be said to appeal to a quasi-Lockean 'folk' conception of personal responsibility.

But if we are intuitive Lockeans about the self, how deep and robust are those intuitions? In this chapter, I will examine and question the use of thought experiments in eliciting intuitions about philosophical puzzles (§ 3.2). Then, after introducing the method of 'experimental philosophy' (§ 3.3), I will discuss recent experimental studies of intuitions about the persistence of self (§§ 3.4, 3.5) and note that while there is some evidence suggesting a popular quasi-Lockean conception of the self, this evidence is not conclusive.

3.2 Intuitions and thought experiments

Intuitions are thought of as spontaneous and pre-theoretical responses to situations, questions, and problems—including philosophical puzzles. They are thus distinguished from reflected philosophical theories, although

they may well be informed by implicit metaphysical commitments. It is often the way of philosophical enquiries either to vindicate or to modify our initial intuitions about a question, particularly when a question seems to throw up conflicting intuitions. *Thought experiments* take the form of an imagined scenario designed specifically in order to elicit intuitions about some philosophical problem or puzzle (or to illustrate the consequences of a philosophical view), and so to help develop (or dismiss) a reflected philosophical position.

The literature on self and personal identity in particular abounds with thought experiments. Here too Locke is a precursor: his chapter on personal identity (1690/1706, II.xxvii) contains a number of thought experiments to support his position. To illustrate that the identity conditions of human beings *per se* (as opposed to persons) do not rest in rationality alone, Locke reproduces William Temple's tale of Prince Maurice's parrot as an example of an allegedly rational (and talking) being that is very obviously not a human being (II.xxvii.8). Having established his notion of self, Locke emphasizes its distinctness from that of an immortal *soul* with another thought experiment: although someone alive now might somehow have inherited the soul of either of the Trojan War heroes Nestor and Thersites, that would not make him of the same *self* unless he also retained the *consciousness* of Nestor or Thersites (II.xxvii.14). Next, to illustrate the distinction between self and *body*, Locke presents the first of many body-swap thought experiments in the personal-identity literature, involving a prince and a cobbler (II.xxvii. 15): were the prince's soul to 'enter and inform' the cobbler's body and, crucially, carry the prince's consciousness with it, we would then take the cobbler to be 'the same person with the prince'—but not the same man. Further emphasizing the role of consciousness for the continuity of the self, Locke repeatedly invokes different versions of Socrates: awake and asleep (II.xxvii.19), and as an infant and after resurrection (II.xxvii.21). Improbable though most of these scenarios are, their point is to elicit what matters about the concept of a person, as opposed to that of a man or his soul.

Subsequent treatments of personal identity in the philosophical literature have largely followed Locke in appealing to thought experiments. We find fission cases in Lewis (1976) and Parfit (1984), here alongside merger cases and teletransportation, and more body-swapping in Williams (1970), of which more shortly. All of these are designed to illuminate some prob-

lem or puzzle arising from what we seem intuitively to believe about the nature and persistence conditions of persons. And almost all of them describe situations which are not only fictitious, but quite unlike any situation which any of us is likely to encounter in the course of a human life. Kathleen Wilkes (1988), a notable dissenter from the use of thought experiments in the personal-identity literature, argues that thought experiments lead to inconclusive results and are irrelevant to actual conditions in people's lives that pose questions about the boundary conditions of personhood, such as infancy, mental deficiency, dementia, insanity, and dissociation.[1]

Here then are two major problems with the use of thought experiments and associated intuitions. First, if thought experiments typically involve puzzle cases, can we be sure that there is but one 'intuitive' solution to the puzzle? Or that there is any solution at all? And if there are several, how are we to adjudicate between them? Secondly, if thought experiments are far-fetched scenarios removed from ordinary experience, how are we warranted in extrapolating from these cases to more general and familiar ones?

The first problem is most apparent in the observation that *intuitions about puzzle cases vary between individuals*. For instance, in a tutorial group of first-year philosophy students, Plutarch's *Ship of Theseus* puzzle will elicit different intuitions from different students, with some opining that the ship made of parts that have gradually replaced all the original parts is the proper Ship of Theseus, others that it is the ship reconstituted from the original discarded parts; some few will even claim that both are.[2] It may be tempting here to suggest that such differences will disappear once the students have been thoroughly tutored in philosophical analysis. But it is by no means to be taken for granted that even philosophers agree among each other in their intuitions about a given puzzle. After all, puzzle cases are interesting precisely because they tend to elicit conflicting intuitions. But then the use of thought experiments only confirms that there *is* a puzzle and by itself offers no conclusive answer to it.

The second problem with philosophical thought experiments arises from the fact that they deal with highly specific scenarios. Thus, the intu-

---

[1] I discuss some of these cases in chs. 7 and 8, but with respect to the self, not personhood.

[2] One might also hold that *neither* is Theseus's ship, though I have not come across this intuitive response among my undergraduates.

itions thought experiments elicit may be of limited general applicability. We may have one intuition about one thought experiment and a conflicting one about another (or another version of the first thought experiment), where both are designed to elicit intuitions about the same concept. Indeed, the two thought experiments may be explicitly designed with the purpose of eliciting conflicting intuitions. Thus, Bernard Williams (1970) presents a case of two individuals' swapping bodies from two different perspectives. One presentation elicits the intuition that psychological continuity is necessary and sufficient to ensure continuity of the self. The other presents the case from a first-person perspective involving fear of future pain despite the loss of all psychological continuity; here the intuition Williams seeks to elicit is that since we can fear future bodily harm, it is bodily continuity that matters for the continuity of the self. Such cases suggest that, rather than eliciting dependable intuitions about a philosophical problem, thought experiments can serve to produce a state of cognitive dissonance by triggering conflicting intuitions. That may be Williams's point in this particular case: but it illustrates the general danger of 'thought experimenter bias' (Nichols & Bruno, 2010, p. 304), where a thought experiment is (deliberately or inadvertently) framed in a way that leads to the desired intuition about a problem. I discuss Williams's two presentations in more detail below (§ 3.4.2).

Further, there is Wilkes's (1988) point that the kinds of scenarios imagined in thought experiments, especially those concerning personal identity, are often highly improbable or empirically impossible. We do not ordinarily come across a prince's consciousness entering a cobbler, or the similar body-swapping scenario imagined by Williams. Nor are we accustomed, outside of science fiction, to teletransportation—Derek Parfit's (1984, ch. 10; 2011; 2012) favourite thought experiment. Now it might be argued that such scenarios nonetheless help us elucidate, in the abstract, just what it is that is important about the continuity of a person or self. Thus Locke's and Parfit's thought experiments draw us towards attaching the greatest importance to psychological continuity. But of course they are *designed* to do so, deliberately removing the bodily continuity of the characters in their scenarios, blithely assuming that this is indeed separable from their psychological continuity. In so doing, they do not establish that psychological continuity is *all* that matters generally—only that, in constructed

cases where psychological continuity is all that is *available*, we're inclined to assume that the person or self continues through it. They also do not tell us anything at all about how anyone's psychological continuity is ensured in the actual world.

In short, the second problem with thought experiments is their specificity, which may at best elicit *local* intuitions about a particular case, rather than about the *general* underlying question; and further, that it may be impossible to generalize from the local case because it is one not encountered in ordinary life.

How can these problems be addressed? Over the last decade or so, a new method has been adopted by a number of philosophers that seeks to address the first problem identified here—that of diverging intuitions: *experimental philosophy*. Used in the right way, it can also help mitigate the second problem I've discussed—that our intuitive responses are specific to the scenarios presented.

3.3 Experimental and empirical philosophy

Experimental philosophy has been described as a 'new movement' (Knobe & Nichols, 2008) and, with some hyperbole, a 'methodological revolution' (Prinz, 2008). Its motivation is to make use of experimental methods in tackling philosophical problems. In probing intuitions about puzzle cases, this means sampling the intuitions of ordinary, philosophically untrained people about these cases, rather than relying solely on individual philosophers' intuitions about them, which may or may not be representative of those held by a wider population.

The two principal aims of experimental philosophy in this kind of enquiry are to find out (i) what people's intuitions about philosophical problems actually are (as opposed to what a particular philosopher *thinks* they are), but also (ii) whether these intuitions are warranted (Sosa, 2008). The commonest method for achieving the first aim is borrowed from social psychology and involves questionnaire-based surveys. Participants in such surveys are typically presented with a philosophical thought experiment and probed for their intuitive response to the case presented. For the second aim, one may test the intuitive responses obtained against data

either from existing studies in e.g. the psychological sciences or from specially designed experiments. Though unlike the traditional image of the philosopher as a solitary being given to reasoning from an armchair, the method of gathering data on people's intuitions and, where appropriate, using empirical and experimental studies to test whether such intuitions are warranted by the facts, is an extension of—rather than a discontinuity with—more traditional philosophical methods (Nichols, 2004; Knobe & Nichols, 2008). This continuity is evident from the fact that experimental philosophy studies begin just like traditional philosophical analysis: a problem or puzzle is identified and intuitions as to its solution are sought. (As Jesse Prinz (2008) observes, the traditional philosophical method of introspection is itself an observational study, albeit with a sample of just one participant.) The novelty of experimental philosophy lies in the methods used thereafter, i.e. the gathering and analysis of ordinary people's intuitive responses to philosophical puzzles.

The usefulness of such an enterprise should appeal to any philosopher who has ever witnessed or been party to a disagreement among colleagues about what intuitions we 'commonly' have about some conceptual puzzle or problem, and whether these prior intuitions hold when the puzzle is modified or refined, and if not, what new intuitions are elicited. It should thus be useful to know whether what some philosophers suppose our intuitions to be about a certain problem is in fact matched by the population at large. Experimental evidence can illuminate who holds what intuitions, and how common or uncommon certain 'commonly held' intuitions actually are.

The experimental method thus tackles the first of the two problems identified in the previous section: the variability of intuitions between individuals. Obtaining responses to particular philosophical puzzles from a larger sample of individuals minimizes the risk of individual bias in intuitions. Of course, there is still likely to be disagreement within a given sample. However, the experimental method allows such disagreement to be quantified. With results from an experimental study, we can show what proportion of a sample respond in what way to a given philosophical problem. Supposing there are two typical intuitive responses A and B to a philosophical puzzle, the survey method combined with statistical analysis will produce one of three outcomes: that the majority of a sample hold intuition

A rather than B, that the majority hold intuition B rather than A, or that there is no significant difference in numbers between adherents of A and B.

But it will be observed that even though we may gain a more or less representative insight into 'folk' intuitions about a particular problem in this way, we still only obtain responses to a particular scenario. Indeed, the second problem identified in the previous section, concerning the specificity of thought experiments, appears at first sight not to be avoided by experimental philosophy. For a typical experimental philosophy survey presents participants with a vignette that is just as specific as a philosopher's armchair thought experiment. Nevertheless, the experimental method can mitigate the problem of overly specific thought experiments by submitting different vignettes to the same experimental group and comparing participants' responses to different scenarios, or comparing responses to a concrete scenario with responses to an abstract question. (I'll discuss an example of such a study in the next section.) Again, statistical analysis is used to determine whether differences in responses to different scenarios are significant.

As for the second aim of experimental philosophy—to scrutinize our intuitions about some puzzle case with respect to whatever facts we can gather about that matter: this is already an integral part of many philosophical enquiries. Empirical data, whether specially gathered or generated by relevant scientific research programmes, provide an obvious testing ground for both philosophers' and common-sense intuitions about philosophical problems. Here one may usefully follow Prinz's (2008) distinction between *experimental* philosophy and *empirical* philosophy, where the former engages in 'data collecting' and the latter in 'data mining' (p. 196). Having collected data on ordinary people's intuitions about philosophical problems, the second aim of testing whether these intuitions are warranted may be served by the 'mining' of data already generated in the empirical sciences, specifically psychology and neuroscience.

This may at first seem to contrast sharply with the time-honoured method of armchair theorizing about whether common-sense intuitions hold up in light of available facts. But again, there is continuity with traditional philosophical method: it is simply that by referring to empirical studies, our philosophical theorizing (which we may still carry out while seated in an agreeably cushioned item of furniture) benefits from additional data

that bear on the appropriateness or otherwise of our intuitions. Nor is the use of available empirical data in philosophical enquiries really all that novel and revolutionary, having been practised, to varying degrees, by Descartes, Locke, Hume, James, and many others. What is perhaps novel is the sheer wealth of data the behavioural and brain sciences now provide that was not available to philosophers in earlier centuries. Scouring these data for those that are relevant to a given philosophical problem may present its own difficulties. But where they do bear on philosophical problems, it cannot be in philosophy's interest to ignore them.

The remainder of this chapter is concerned with experimental philosophy studies on personal identity and the self. From these, it will be apparent that the 'Lockean' intuition that the persistence of the self requires persistence of one's memories indeed seems a fairly common one, though —as Williams suspected—not applicable to all cases, and with a sizeable minority of test subjects taking a different view. The following chapters will then engage in what Prinz calls *empirical* philosophy, garnering evidence from the cognitive sciences to expose the problems faced by Lockean, neo-Lockean, or quasi-Lockean popular conceptions of the self which anchor the persistence of the self in psychological continuity.

3.4 Experimental philosophy and the self

3.4.1 Experimental studies on a 'Lockean frame' scenario

Do ordinary people share Locke's and other philosophers' intuition that memory is necessary for the persistence of the self?[3] An obvious way to test this using the method of experimental philosophy is to present philosophically untrained participants with two versions of a thought-experiment-style vignette that differ only in the condition of a protagonist's memories, which are either preserved or lost. Participants should then judge in each case whether the protagonist is or is not the same person at the end of the vignette as at its beginning. The independent variable of such an experi-

[3] One reason to suppose the folk might be 'Lockean' in this way is that they might have an implicit commitment to mind–body dualism, perhaps for religious reasons.

mental design is the preservation or loss of the protagonist's memories; the dependent variable the participants' judgement of the sameness or otherwise of the protagonist. The experimental hypothesis is that memories are commonly and intuitively deemed necessary for the preservation of the self or person.[4] A result where participants judge the protagonist not to be the same person in the lost-memories case, but the same person in the preserved-memories case, would support the hypothesis. A result with no significant difference between the conditions would support the null hypothesis, i.e. that it is not the case that memories are commonly and intuitively deemed necessary for the preservation of the self or person.

Intending to design such an experiment from scratch, I was fortunate to find this to be unnecessary, since an experimental study with an appropriate design had already been carried out; not once, but twice—by Sergey Blok and colleagues (2005) and by Shaun Nichols and Mike Bruno (2010). I was able to replicate the study once more. The vignette presented to participants was this:

> Jim is an accountant living in Chicago. One day, he is severely injured in a tragic car accident. His only chance for survival is participation in an advanced medical experiment called a "Type 2 transplant" procedure. Jim agrees.
>
> It is the year 2020 and scientists are able to grow all parts of the human body, except for the brain. A stock of bodies is kept cryogenically frozen to be used as spare parts in the event of an emergency. In a "Type 2 transplant procedure," a team of doctors removes Jim's brain and carefully places it in a stock body. Jim's original body is destroyed in the operation.
>
> After the operation, all the right neural connections between the brain and the body have been made. The doctors test all physiological responses and determine that the transplant recipient is alive and functioning. The doctors scan the brain of the transplant recipient and note that the memories in it are the same as those that were in the brain before the operation.
>
> (Blok et al., 2005)

In the second condition, the conclusion of the vignette was altered thus:

---

[4] The design of the experiments presented here does not allow for the self–person distinction I have drawn in the previous chapter. I address this point in my discussion below (§ 3.5.3).

> The doctors scan the brain of the transplant recipient and note that *no* memories in it are the same as those that were in the brain before the operation. *Something must have happened during the transplant*.

For each condition, participants were then asked to rate their agreement with the statement 'After the operation, the Type 2 transfer recipient is Jim,' using a ten-point Likert scale ranging from 0 ('strongly disagree') to 9 ('strongly agree').

Nichols and Bruno (2010) introduced a further variable into the experimental design: whether the vignette was presented as a third-person or a first-person scenario. The motivation for this addition is Williams's (1970) suggestion that presenting a given thought experiment in the first-person perspective may change our intuition about the necessity of memories for the persistence of the self—perhaps because, as Nichols (2008) suggests elsewhere, the indexical *I* is 'descriptively exceedingly thin' (p. 523) and so may mislead one into being able to imagine *oneself* under that thin description of 'I' in scenarios lacking one's normal 'thick' descriptors (i.e., one's psychological states). Nichols' and Bruno's first-person versions omitted the first sentence of the vignette and substituted 'you, your, etc.' for 'Jim, Jim's, he, his, etc.'.[5]

My own study replicated the four conditions used by Nichols and Bruno (2010), but unlike their and Blok and colleagues' (2005) studies used a *between-participants* design, so that each participant was given just one of the four scenarios (same memories or lost memories in either the third-person or the first-person version).[6] Blok and colleagues and Nichols and Bruno used a *within-participants* design where each participant was presented with both the same-memories case and the lost-memories case. Though

---

[5] Although these experimental conditions use the *second*-person pronoun, they are (here and in Nichols & Bruno, 2010) referred to as *first*-person conditions, since they are designed to elicit a first-person perspective on the scenario in the participant. One may wonder, however, whether this second-person prompt of a first-person perspective quite mirrors what Nichols (2008) refers to as 'the poverty of the I-concept' (p. 523) when one engages in *I*-imaginations by oneself.

[6] I also substituted 'Birmingham' for 'Chicago' and '2030' for '2020'. The first substitution was motivated by the desire to avoid giving my (UK-based) participants a scenario obviously set in a foreign country, the second was made in order to replicate roughly how far in the future the original study had set the scenario, relative to the date of the study being conducted.

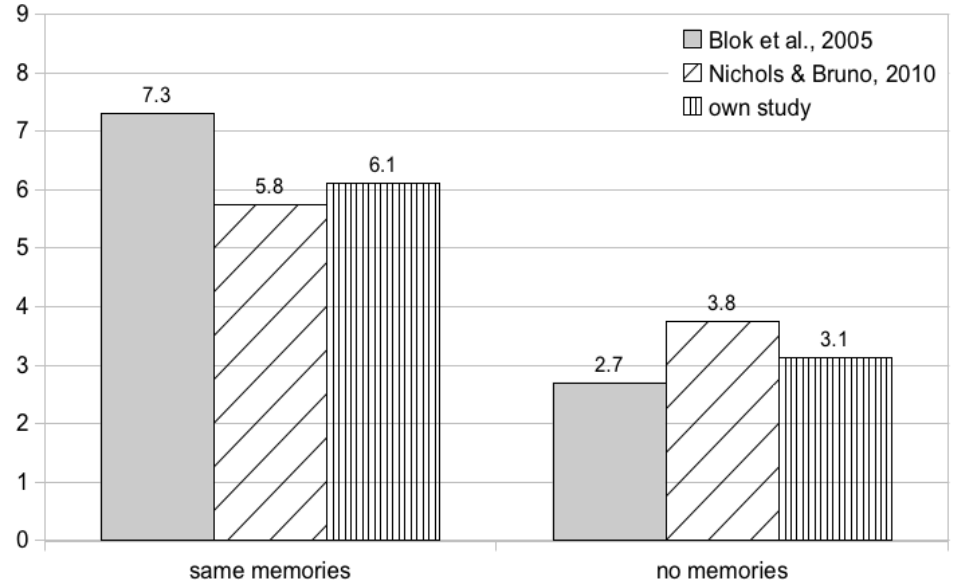

Figure 3.1 Mean values of responses to 'Lockean frame' surveys.

both previous studies controlled for order effects by balancing the order of presentation of the vignettes across participants, a within-participants design will obviously mean that the participant is already familiar with the general scenario when presented with the second vignette. To avoid any potential confounds from this familiarity effect, and following up a suggestion made by Nichols and Bruno (2010, note 16), I tested responses to the different scenarios entirely *between* participants, each participant having been randomly assigned one of the four scenarios.

The participants in the three studies were: for the relevant part of Blok and colleagues' study, 33 introductory psychology students; for Nichols and Bruno, 70 students in an introductory philosophy class; in my study, 73 first-year philosophy undergraduates who at the time of the survey had not been exposed to philosophical discussions of personal identity.

The results of the three studies are broadly congruent (fig. 3.1). Blok and colleagues obtained a mean agreement of 7.3 for the same-memories condition and 2.7 for the no-memories condition. The means in Nichols and Bruno's study were closer to the midline, but not significantly, at 5.8 (same memories) and 3.8 (no memories); they report no significant difference

between the third- and first-person conditions (the figures given here are therefore aggregate means for both conditions). In my study, the aggregate mean for the same-memories conditions was 6.1, for no memories, it was 3.2.[7]

All three studies thus support the experimental hypothesis that persistence of memories is, in the context of the given thought experiment, deemed necessary for the persistence of self; Nichols and Bruno's and my study additionally support the hypothesis that it makes no significant difference whether the relevant scenario is presented in the third or the first person. In addition, the *between*-participants design of my study also made no difference. I did, however, observe an interesting feature in the *distribution* of responses to the lost-memories cases in my study, where about a quarter of respondents form a clearly separate group in taking the opposite view to the majority. I discuss this in § 3.5.1 below.

3.4.2 Williams's 'pain frame': conflicting intuitions

The difference between third- and first-person presentations is not the only difference between the two 'frames' Williams (1970) gives of his thought experiment. The other difference is that in one scenario, the emphasis is on the continuity of a person's memories, whereas in the second presentation, it is a subject's affective responses that supposedly elicit a different (non-Lockean) intuition about the continuity of the self. It is Nichols' and Bruno's merit to have deconflated these two variables in their study.

Williams's second presentation involves the fear of future pain and the explicit loss of a subject's memories:

> Someone in whose power I am tells me that I am going to be tortured tomorrow. I am frightened, and look forward to tomorrow in great apprehension. He adds that when the time comes, I shall not remember being told that this was going to happen to me, since shortly before the torture something else will be done to me which will make me forget the announcement. This certainly will not cheer me up, since I know perfectly

---

[7] There was a significant main effect of whether or not memories were preserved ($F(1, 69) = 32.26$, $p < 0.001$, $\omega^2 = 0.31$). There was no significant main effect of third/first-person difference ($F(1, 69) = 0.00$, $p = 0.984$, $\omega^2 = 0$) and no significant interaction effect between whether or not memories were preserved and third/first-person difference ($F(1, 69) = 0.07$, $p = 0.793$, $\omega^2 = 0$).

> well that I can forget things, and that there is such a thing as indeed being tortured unexpectedly because I had forgotten or been made to forget a prediction of the torture: that will still be a torture which, so long as I do know about the prediction, I look forward to in fear. He then adds that my forgetting the announcement will be only part of a larger process: when the moment of torture comes, I shall not remember any of the things I am now in a position to remember. This does not cheer me up, either, since I can readily conceive of being involved in an accident, for instance, as a result of which I wake up in a completely amnesiac state and also in great pain; that could certainly happen to me, I should not like it to happen to me, nor to know that it was going to happen to me.
>
> (Williams, 1970, pp. 167–8)

The intuition Williams seeks to elicit by this thought experiment (labelled the '*pain* frame' by Nichols and Bruno, as opposed to the '*Lockean* frame' of the first thought experiment) is that the self persists regardless of the loss of memories: 'it seems quite sensible to fear the pain that will be experienced by the person with your original body, despite the amnesia' (Nichols & Bruno, 2010, p. 295).

Nichols and Bruno (2010) proceeded to test this intuition by means of a second survey, which was again presented in first-person and third-person versions:

> Imagine that some time in the future your brain has developed a lethal infection and will stop functioning within a few hours. In the emergency room, you are alert and listening as the doctors explain to you that the only thing they can do is the following:
>
> > Render you completely unconscious, and then shave your head so that they can place electrodes on your scalp and shock your infected brain. Unfortunately, this procedure will permanently eliminate your distinctive mental states (including your thoughts, memories, and personality traits).
>
> You slip into unconsciousness before the doctors can discuss the matter further with you, and they elect to perform the procedure. It works exactly as expected. Several days after the procedure, the doctors perform some follow up brain scans and administer a series of painful shots. (p. 300)

The third-person version substituted 'Jerry, he, him, his' for 'you, your'.

Along with a number of questions designed to probe participants' comprehension of the scenario, participants were then asked to agree or

disagree with the statement, 'When the doctors administer the series of shots, *you* [*Jerry*] will feel the pain' (p. 301 [302]). The responses obtained from participants having demonstrated comprehension of the scenario were as follows: in the first-person condition, 75 % agreed with the statement '*you* will feel the pain'; in the third-person condition, 72 % agreed with the statement '*Jerry* will feel the pain'. Again, Nichols and Bruno report no significant difference between first- and third-person conditions (fig. 3.2 (a)).

Their interpretation of these results is that Williams is right in asserting that the particular framing of a thought experiment makes a difference to what intuitions about the persistence of the self are elicited. The difference here does not, *pace* Williams, seem to depend on whether the scenario is presented in the first or third person: none of the studies considered here that tested for this variable found that difference to have a statistically significant effect on responses. However, the 'pain frame' may encourage respondents to impute a persistent subject, because it is intuitively obvious that *someone* is going to *feel the pain*:

> After all, if *I* am not going to feel it then *who is*? Similarly for the third-person version, if *Jerry* isn't going to feel the pain, then *who*? There is plausibly pressure here to give a persistence response. (p. 304)

This of course is the second problem with thought experiments identified in § 3.2—their tendency to elicit local intuitions about a specific scenario, rather than about the general applicability conditions of a concept. But if there is thought-experimenter bias in Williams's two frames, which of them is the biased one? And how can experimental philosophy help overcome the specificity problem?

Nichols and Bruno conducted a further survey that addresses both these questions. If the problem is the concreteness of thought-experimental scenarios, the obvious solution is to remove concrete particulars from the questionnaire and pose the question in an abstract way. Nichols and Bruno did this by asking participants for a free response to the question, 'What is required for some person in the future to be the same person as you?' To avoid experimenter bias towards Lockean responses, their questionnaire made no mention of psychology. The result was this: 'In their free responses, over 70 % of participants explicitly mentioned psychological factors like memory or personality traits as necessary for

persistence' (p. 304). Participants were then asked 'whether they agreed or disagreed with the following statement:

> In order for some person in the future to be *you*, that person doesn't need to have any of your memories. (ibid.)

Here, more than 80 % of participants disagreed—an almost exact reversal of the responses to the 'pain frame' scenario (fig. 3.2(b)).

It seems, then, that once specific thought experiments are removed from consideration, folk intuitions about personal identity take on a strongly Lockean flavour. But, as Nichols and Bruno caution, there may be cases where responses to abstract questions are less reliable than responses to concrete questions in that they offer an unconsidered view that could easily be shifted when presented with a counterexample. And though the specificity of thought experiments is a problem, the fact that philosophers have found it useful to have recourse to such specific scenarios is due precisely to the fact that abstract considerations do not always suffice in teasing out our intuitions about philosophical problems. We do learn from examples.

Nichols and Bruno (2010) therefore ran a final survey in which participants were given both the 'pain frame' scenario and the 'abstract frame', followed by 'an exercise of reflective equilibrium':

> Now that you have answered these questions, we want to call your attention to the fact that it wouldn't really be consistent to say both that you would feel the pain of the shots and also that in order for a person to be you, that person must have some of your memories. In light of this, which one are you more inclined to agree with? (check one please)
> ___ More inclined to say that you would feel the pain in case #1.
> ___ More inclined to say that in order for some person in the future to be you, that person must have some of your memories.
>
> The point was to see whether people would show a preference for one judgment over the other. And we did find a preference. 64% of participants sided with the psychological response under these reflective conditions, greater than what one would predict by chance alone . . . See figure [3.2(c)].
>
> Once again, people's judgments favor the view that persistence of psychological features is required for persistence of self. (p. 306)

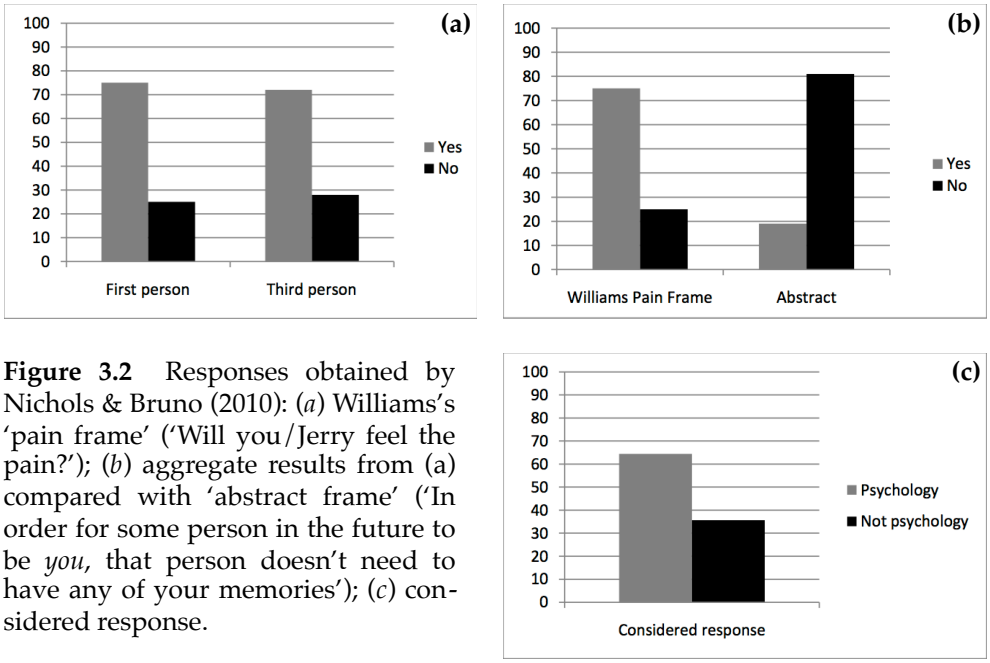

Figure 3.2 Responses obtained by Nichols & Bruno (2010): (*a*) Williams's 'pain frame' ('Will you/Jerry feel the pain?'); (*b*) aggregate results from (a) compared with 'abstract frame' ('In order for some person in the future to be *you*, that person doesn't need to have any of your memories'); (*c*) considered response.

To sum up, then, the results and implications of Nichols and Bruno's study: In a 'Lockean frame' survey (borrowed from Blok and colleagues, and also replicated by me), there is a clear result that most participants intuitively judge the persistence of memories to be necessary for the persistence of the person or self. Conversely, given a 'pain frame' scenario modelled on Williams's (1970) second presentation of a thought experiment designed to elicit the intuition that the self persists in the same body regardless of the loss of one's psychological states, the majority of participants respond accordingly, their intuitive response now at odds with the 'Lockean frame'. So far, then, the experimental method seems to confirm, with a wider sample than a single 'armchair' philosopher, that different framings of thought experiments elicit divergent intuitions.

Nichols and Bruno's third and fourth surveys seek to redress this problem of 'thought-experimenter demand'. The point of both the 'abstract frame' and the 'reflective equilibrium' surveys in their study is to find out which intuitions about the persistence of the self hold when the specifics of particular imagined scenarios are removed and generalization is required (thus addressing the specificity problem of thought experiments I highlighted earlier). And here their results suggest that in both an abstract frame and a considered condition of reflective equilibrium, popular intuitions revert to a quasi-Lockean position.

3.5 Discussion

Do these survey results provide evidence that the 'folk' concept of self is broadly Lockean, such that memory is considered necessary for the persistence of the self? That would be to overstate both the breadth and the depth of what these studies show. Concerning breadth, some remarks about the experimental populations, and variations within them, are in order. As regards depth, it is worth asking just what conclusions about people's intuitions we are warranted to draw from their survey responses. I'll now discuss these points in turn.

3.5.1 Experimental populations and dissenting minorities

The participants in all three studies were psychology or philosophy undergraduates at North American or British universities and not necessarily representative of the general population in their countries, let alone of non-Western populations. As Nichols and Bruno (2010) admit,

> This sample is homogenous on several important factors, including age, culture, and socioeconomic status. It's quite possible that people in different cultures or age or socioeconomic groups will respond differently from the population we studied. And even *within* the population we studied, there was far from uniform agreement about the questions. (pp. 307–8)

Disagreement within a culturally and socioeconomically homogeneous experimental population is also apparent in my replication of the 'Lockean frame' study, in the crucial 'lost memories' conditions. Recall that there was a statistically significant difference between the *means* of the responses to the two scenarios (6.1 for same memories, 3.1 for lost memories, on a 0–9 scale where 0 signified strong disagreement and 9 strong agreement with the statement that the post-operative patient was Jim/you). But it is in the nature of averages to mask underlying differences. These become apparent when one looks at the *distribution* of replies to the different scenarios (fig. 3.3).

The left-hand column in the figure shows the frequency distributions for the two 'same memories' conditions and their aggregate, the right-hand

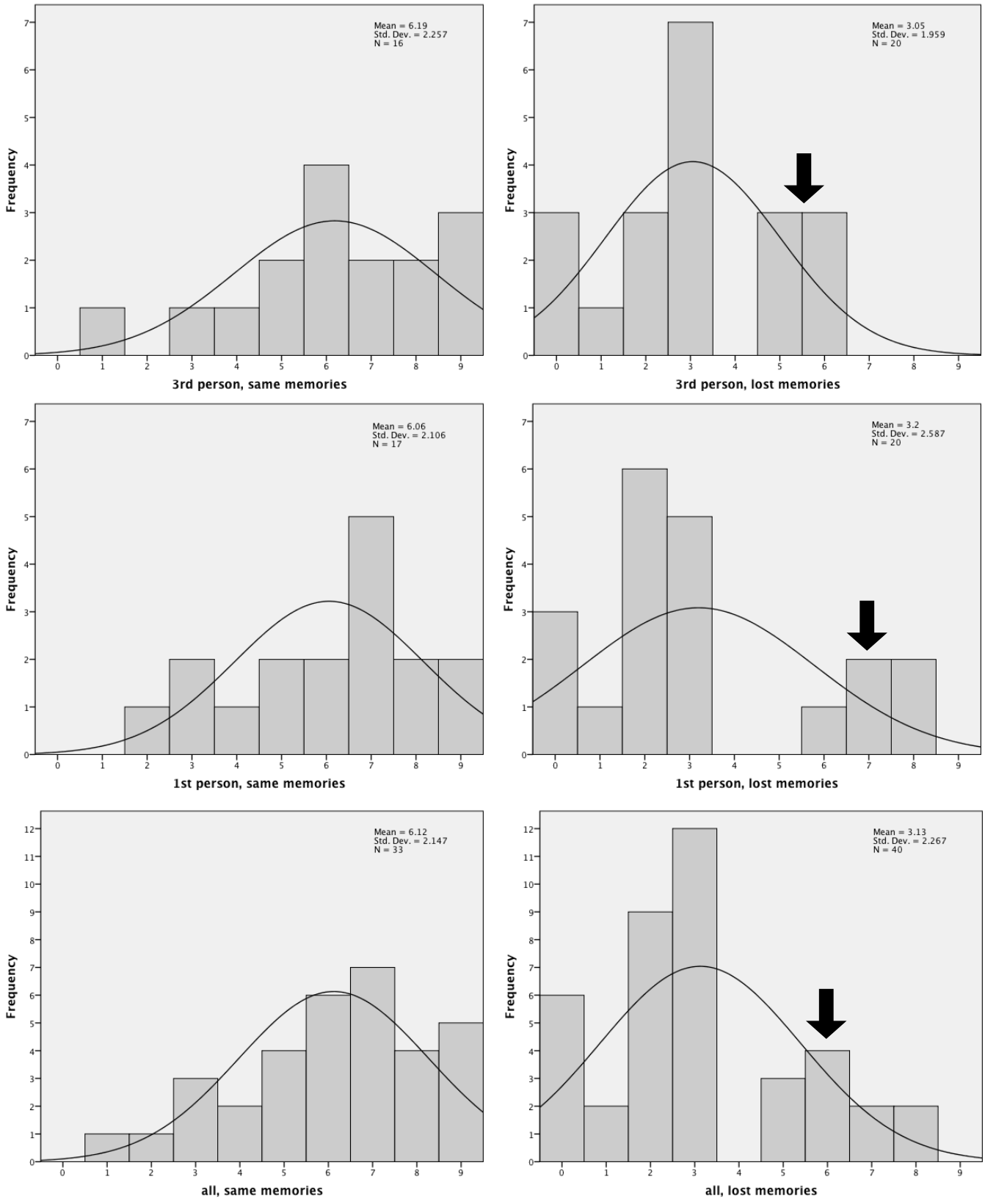

Figure 3.3 Frequency distributions of replies to own 'Lockean frame' survey with 'twin peaks' in lost-memories conditions indicated by arrows.

column for the two 'lost memories' conditions and their aggregate. The black curve in each diagram shows the normal distribution of the responses with its peak at the mean; the underlying histogram the number of responses for each point on the Likert scale, where in all conditions the largest number of responses for any one scale point lies close to the overall mean. Additionally, for the 'same memories' conditions, the histograms approximately follow the normal distribution—where the distribution curves taper off towards the low end of the scale, so do the histograms. But a different picture emerges for 'lost memories': here we find a gap in the histogram at about the midpoint of the scale, with a second, smaller peak in the distribution on the other side of the midpoint from the overall mean. About a quarter of participants in these conditions dissented from the majority view and agreed more than disagreed with the statement that the post-operative patient was Jim/you despite the loss of his memories, particularly so in the first-person condition.

This 'twin peaks' effect in the frequency distribution illustrates that the 'lost memories' scenarios trigger clearly divergent responses among different participants. Given the philosophically puzzling nature of the case, one might expect nothing less—after all, these studies were conducted precisely *because* philosophers differ about our intuitions regarding the persistence conditions of persons and selves. One might therefore say that these results merely confirm that what we took to be a puzzle is indeed a puzzle. But it is striking that, unlike in the 'same memories' conditions, participants in the 'lost memories' cases divide so neatly into two (unequal) camps. Here, then, the experimental method apparently has not done away with the problem of diverging intuitions—indeed, stating just the means of the responses obtained completely masks the underlying disagreement.

But what is the nature of this disagreement? It might be tempting to suppose that the dissenting minority of participants in my study are intuitively anti-Lockean about personal identity. But this risks overstating the depth of what can be gathered from survey responses. We now need to consider just what is being probed in these surveys and how this is done.

3.5.2 Survey pragmatics

In a critical discussion of the survey method used in experimental psychology, Samuel Cullen (2010) warns that 'the assumption that intuitions can be simply read off from survey responses' (p. 275) is misguided. This, he argues, is because the pragmatics of surveys—the type and phrasing of the questions asked, the measurements used, the context of the survey—allow for responses to be influenced by factors other than participants' intuitions.

Let me first address the *pragmatics* of measuring survey responses. On this point, one might wonder whether the results of Nichols and Bruno's (2010) 'pain frame', 'abstract frame', and 'reflective equilibrium' surveys would be as clear cut if they had again used a Likert scale for answers, rather than a forced-choice yes/no option. One might also query whether there were cues in the phrasing of the vignettes that may have influenced participants' responses in one direction rather than another, perhaps resulting from experimenter bias (cf. Strickland & Suben, 2012).

But of course, the very point of Nichols and Bruno's study was to examine the *thought*-experimenter bias inherent in the Lockean and 'pain' frames. It cannot therefore be a criticism of the vignettes used that they were skewed to elicit one intuition rather than another about the persistence conditions of the self: that was the purpose of comparing them. Meanwhile, for their free-response 'abstract frame' survey, Nichols and Bruno carefully avoided all references to psychological traits in the question, yet obtained results leaning heavily towards psychological-continuity views of the self. It is also worth noting that they expected to find a significant difference in responses between third- and first-person conditions, but found none: on this variable, there is no evidence of experimenter bias towards the expected result.

What of Cullen's (2010) general point that survey responses do not equate to participants' intuitions, but rather are influenced by 'pragmatic cues and conversational norms' (p. 285)? Here we may admit that the survey methodology used in experimental philosophy is not *perfect* as a tool for gathering data on people's intuitions. But it is difficult to imagine circumstances under which people could ever *not* be influenced by 'pragmatic cues and conversational norms' (debates among philosophers, in print and in seminars, not excepted). Further, cues and conversational norms are

themselves open to empirical study. Experimental philosophy can offer a way of studying these by repeating trials with variations in cues and questions and so obtain a better understanding of how these factors influence results. Some of this has already been done with respect to the 'Lockean frame' surveys, as I'll now discuss.

3.5.3 Survey questions: persons and individuals and their identities

It seems an obvious observation that the responses one gets depend on the question that is asked. But it is less obvious whether the questions asked in experimental surveys reflect the philosophical categories one is trying to illuminate. In the 'Lockean frame' studies, participants were asked to indicate their level of agreement or disagreement with the statement 'After the operation, the Type 2 transfer recipient is Jim' (or 'you' in the first-person conditions). The prompt made no mention of 'identity', 'self', or 'the same'. Now, to assume that the responses indicate participants' intuitions about personal identity or the self means also to assume that identity judgements can be read off the use of proper names or personal pronouns. This may seem a reasonable assumption—but it is not clear whether the continued use of a proper name or personal pronoun carries with it a judgement of a continuing *person*, or merely a judgement of a continuing *individual*.

In a response to Nichols and Bruno, Renatas Berniūnas and Vilius Dranseika (2016)[8] report on a number of studies they conducted to elucidate this question. Among these, they too replicated the 'Lockean frame' survey discussed here, but with two additional statements, asking participants whether the transplant recipient 'is still *the same person*' and whether he 'is still *a person*' (p. 106).[9] The responses they obtained for the original prompt were in line with the previous 'Lockean frame' experiments by

---

[8] I thank my MA student Miranda Fisher-Levine for bringing this paper to my attention.

[9] Blok and colleagues' (2005) original study also tested this last statement, here in order to differentiate between conditions with a human transplant recipient and a robot transplant recipient. In the 'human recipient' conditions, agreement with 'still a person' was high for both 'same memories' and 'no memories' conditions; in the 'robot recipient' cases, it was low.

Blok et al., Nichols and Bruno, and myself. Responses to the 'same person' prompt differed insignificantly from these, but showed slightly lower agreement in both 'memories retained' and 'memories lost' conditions. This suggests that the assumption that the use of proper names carries an identity judgement of some kind is sound—but not necessarily a judgement of *personal* identity. For responses to the 'still a person' question showed an only marginally significant difference between the 'memories retained' and 'memories lost' conditions. To Berniūnas and Dranseika, 'the results suggest that *continuity of memory* is important in tracking an individual, but not as important in categorizing the post-transformation individual as a person' (p. 106).

Following up this suggestion, they conducted further surveys, which used 'scenarios that are more down-to-earth (not "science fiction") and explicitly present a reverse case of radical psychological discontinuity and bodily continuity' (p. 107), such as a patient in a persistent vegetative state (PVS).[10] From their results, along with those of other studies reported in Blok et al. (2005), Berniūnas and Dranseika (2016) conclude that 'it would be helpful to distinguish between two types of identity judgments: those of *personal identity* and *individual identity*' (p. 113). Such a distinction can account for someone's remaining the same *individual* while ceasing to be a *person*, as might reasonably be said of a PVS patient. This tallies perfectly with the observation made in the previous chapter (§ 2.4.1) that the quality of personhood is dissociable from the continuity of an organism (cf. Olson, 1997).

Berniūnas and Dranseika put this distinction to the test by re-running the 'Lockean frame' and their PVS case surveys, this time probing participants' agreement with statements that distinguished between the patients' being 'the same person *numerically*' and 'the same person *qualitatively*'. They were able to do this without having to explicate the philosophical notions of numerical and qualitative identity in complex terms, since their and their participants' native Lithuanian allows such a distinction in ordinary terms, with 'the same' translating as *tas pats* when used numeric-

[10] Although they make no reference to Wilkes (1988), this choice of scenarios would seem to address her (and my own) complaint about the over-reliance of the personal-identity literature on otherworldly, science-fiction cases.

ally, and *toks pats* when used qualitatively.[11] The results showed that judgements of someone's being 'the same person *numerically*' (*tas pats*) in the 'memories lost' conditions were greatly increased and no longer statistically significantly different from the 'memories retained' conditions.

The implications of the numerical–qualitative or individual–person distinction on Nichols and Bruno's results are that their results in the 'abstract frame' and 'reflective equilibrium' conditions must be read with caution, since participants may have responded to the survey questions with a qualitative rather than numerical reading of identity in mind. Further, in Berniūnas and Dranseika's conclusion,

> it seems not to be the case that, in thinking about hypothetical cases of transformation, the folk rely on anything like the philosopher's notion of personal identity. In other words, identity judgments in hypothetical cases do not track identity conditions supplied by the concept *person*, since there is a double disassociation between identity and personhood judgments.
>
> (p. 119)

We should note that Berniūnas and Dranseika's research is far from a criticism of the method of using experimental surveys to probe popular intuitions. It is, on the contrary, an illustration of how refinements of the questions used in experimental methodology can yield more precise results.

Their conclusion that there is a distinction in popular judgement between the continuity of *individuals* and that of *persons* supports my point that personhood is a secondary characteristic that we apply to individuals, whose persistence conditions *qua* individuals are *prior* to the attributes of being a person. Unfortunately, none of the studies discussed here explores the distinction between *person* and *self*, or for that matter the question whether a continuing individual, bereft of the qualities of personhood, retains the same self. The person–individual dissociation elaborated by

---

[11] The closest equivalent to this in English might be the somewhat awkward 'self-same' for a numerically identical item. My native German, like Lithuanian, has two words for 'same'—*selbe* for numerically identical items, and *gleiche* for qualitatively identical ones, but their correct use is the subject of much every-day pedantry, since colloquially and regionally *selbe* may be used in both senses. Whether this applies to *tas pats* and *toks pats* in Lithuanian I do not know, but Berniūnas and Dranseika sought to forestall any confusion about the use of these expressions with an explanatory paragraph reminding participants of the difference by means of an easy-to-understand example involving billiard balls.

Berniūnas and Dranseika could still allow for the persistence of the self to be judged by the same attributes as that of the person. Or the continuity of the self could instead ride on the persistence conditions of the individual—though we can speculate that in the case of a patient in a persistent vegetative state, popular judgement might be hesitant to ascribe fully the same self to that patient as before. Finally—and consonant with my system view—the persistence of the self could be judged to lie somewhere in between: not requiring the full attributes of personhood, but retaining more brain functions than are available to a patient in a PVS.

3.6 Conclusions

I have in this chapter considered the relevance of thought experiments to philosophical problems in general, and in particular to the problems concerning personal identity and the self. Many, but not all, thought experiments in the literature support a psychological-continuity view of the self, but there is a worry that the framing of these thought experiments, particularly those involving empirically highly improbable scenarios, skew our intuitions in a particular direction that may have little bearing on puzzling cases we may encounter in actual situations. The recently developed method of experimental philosophy can offer additional insights into what intuitions ordinary people have on philosophical puzzle cases. As I have demonstrated, this methodology is not without its pitfalls. Thus, while Nichols and Bruno's (2010) studies suggest that, on balance, there seems to be a popular intuition (at least among Western undergraduates) associating the persistence of the self with that of psychological states (particularly memory), subsequent research by Berniūnas and Dranseika (2016) shows that variations in the survey questions posed paint a more complex picture of popular intuitions, particularly in that they appear to dissociate between persons and individuals.

In the experimental studies discussed here, much importance has been attached to *memory*—whether its preservation is or is not a signifier of identity (personal or individual). Yet little or no attention is paid in thought experiments and experimental philosophy vignettes to what memory amounts to. This lack of precision is in the nature of such experiments: we

are assumed to have an intuitive handle on what it means for someone to retain or lose his or her personal memories. But if we assume that the persistence of memory does matter, in a quasi-Lockean fashion, to a popular conception of the persistence of the self, it is then reasonable to enquire into the nature of memory and what implications this has on the nature of the self. This enquiry is the next stage of my experimental-empirical project and the topic of the next chapter.

Chapter Four

Autobiographical memory

4.1 Introduction

The Lockean self, whether on Locke's own view or in neo-Lockean accounts, cannot persist without memory. Arguably, the necessity of memory is given more prominence in more recent accounts (Grice, 1941; Shoemaker, 1984) than in Locke's own, but Locke's remarks on total amnesia make it clear that the 'same consciousness' that marks a Lockean self, whatever else it involves, depends *inter alia* on intact memory. Damasio's (1999; 2010) neurologically grounded account of the self likewise includes remembering as the key activity of an 'autobiographical' self. And in the previous chapter, we have seen that it seems quite a common intuition that the loss of one's memories forecloses or at least seriously threatens the continuity of the self.

Many philosophers have, until fairly recently, taken memory for granted—that is, for the purposes of, say, devising a model of psychological continuity (see § 2.3.3), it seemed to suffice to appeal to conditions in which one did or did not have memories of past events, without enquiring too closely into what it means, conceptually and/or empirically, to remember. Thus, it was enough for Grice (1941) to redefine Locke's 'consciousness'—as characteristic of the self—as 'memory'. Mackie's (1976) revision of Locke depends on periods of unconsciousness being bridged by memory—of whatever kind necessary. For these accounts, and many others,[1] what matters is *whether*, rather than *how*, one remembers.

But, as so often happens when one studies the nature of seemingly simple phenomena more closely, when we enquire empirically into the

[1] Notable exceptions are Vico (1744) (see § 4.4.1 below) and Wittgenstein (1974).

nature of memory, things turn out to be far more complex than the simple binary 'does/does not remember'. From such enquiries into memory, three main themes emerge that are relevant to this thesis. First, the kinds of remembering that concern the self are surprisingly difficult to accommodate in standard classifications of memory. The kind of memory relevant to the self is what is often called *autobiographical memory*, that is, memory of or about oneself (Robinson, 1992; Conway & Rubin, 1993; Nelson, 1993; Klein et al. 2004). This typically involves 'episodic' memories of one's actions and experiences—which when remembered should produce the consciousness of past actions that is central to Locke's account of the self. But we also have 'semantic' (factual) memories that are autobiographical—one can remember facts about one's upbringing, schooling, relationships, and professional career without their being embedded in specific episodes. Indeed, it seems that in organizing our autobiographical memories, these are, over time, transformed from recollections of particular episodes to recollections of more 'general events' and 'lifetime periods'—which are no longer episode-specific (Conway & Rubin, 1993). Recent clinical studies also suggest that some of our semantic self-knowledge can remain intact when specific episodic memories are no longer retrievable (Klein, 2013c). Further, it appears that even episodic remembering involves the operation of several dissociable psychological mechanisms (Klein & Nichols, 2012). Thus, autobiographical memory sits uneasily in standard classifications of memory. I begin this chapter, therefore, with an overview of how memory can be classified (introducing some standard terminology along the way) and what kinds of memory may usefully be termed 'autobiographical memory' (§ 4.2).

Secondly, the *availability* of autobiographical memory imposes severe constraints on a quasi-Lockean view of the self as something somehow fixed by the continuity of remembering. I will review the processes and conditions of memory encoding and retrieval, the different ways in which we can be aware of our memories, and how our autobiographical memories are typically distributed over a lifetime—all of which suggest that autobiographical memory is uneven, patchy, and highly mutable (§ 4.3).

Thirdly—and most important—the operations of so-called 'episodic' remembering are largely *reconstructive*, and require the joint activity of several dissociable self-representational capacities. These processes are guided

as much by current inputs as they are by recalled events in the past. I discuss the reconstructive nature of episodic recall and its implications on the reliability of autobiographical memory (§ 4.4). These observations seriously call into question the historical-constructionist notion of our memories' being *prior* to the self. If we are—as this evidence suggests we must be—constructionist about autobiographical memory, we cannot at the same time be constructionist about the self. Thus, in the final section of this chapter (§ 4.5), I conclude that the self, rather than being a construct from autobiographical memory, is better conceived of as the system that actively constructs our autobiographical memories.

4.2 What is autobiographical memory?

4.2.1 Taxonomies of memory

Memory may be classified in a number of different ways to suit different purposes. Sven Bernecker (2009, ch. 1) lists four standard ways of memory classification: by the length of time a memory is retained; by the degree of awareness one has of a memory; by what kind of prompt triggers memory retrieval; and by content.[2]

The first classification distinguishes between *working* and *short-term* memory, and *long-term* memory. The terminology concerning working and short-term memory is somewhat inconsistent (Sutherland, 1995). Among psychologists, working and short-term memory are generally taken to operate over periods of only seconds. On this view, autobiographical memory in any useful sense is long-term. Neurobiologists, on the other hand, use the terms more distinctly and somewhat more in line with popular usage. One model has working memory, along with a 'sense of the present', operating over fractions of seconds to minutes, short-term memory over

[2] I here ignore Bernecker's own novel taxonomy—by grammatical object of the verb 'to remember'—partly because my concern here is psychology, not language use; partly because autobiographical memory does not fall neatly into one of Bernecker's 'grammatical' classes any more than it maps on to the standard classifications by content which his classification is supposed to improve upon. (For a discussion of Bernecker's theory of memory in relation to personal identity, see Schechtman, 2011a.)

minutes to hours, and long-term memory over days to years—short-term memory here is 'effectively long-term memory that is too new to have become well-established' (Murphy & Naish, 2006, p. 3). According to this model, autobiographical memory can be short-term—e.g. Mackie's (1976) memory bridges between (working-memory) 'I-occurrences'—as well as long-term. But most autobiographical memory—e.g. the memories whose loss so dramatically features in the thought-experimental scenarios discussed in the previous chapter—is long-term memory. It is then worthwhile to ask which short-term memories are committed to long-term memory and why, and I return to this question below (§ 4.3.1). Where in what follows I refer to short-term memory, 'short-term' is to be understood in Murphy and Naish's neurobiological sense of 'minutes to hours'.

Bernecker's second and third ways of classifying memory, by degree of awareness and by what prompts its retrieval, are relevant to autobiographical memory in that these factors impose certain constraints on its availability. I will therefore postpone their discussion to the section on availability constraints (§ 4.3).

Finally, perhaps the most oft-used way of classifying memory is by its content. The standard psychological taxonomy here makes systematic distinctions on two levels. At the first level, there is a distinction between *procedural* and *declarative* memory. The terms are (nearly) self-explanatory. Procedural memory is memory for executing (usually motor) skills; its retrieval is automatic and non-conscious (Sutherland, 1995). Textbook examples of procedural memory include riding a bike or playing the piano. Declarative memory is memory for facts, which can be consciously accessed and articulated (ibid.). Within declarative memory, a further standard distinction, first introduced by Endel Tulving (1972), is between *episodic* and *semantic* memory.

Semantic memory may be defined as 'memory for facts' or 'organized world knowledge', which can be retrieved without the need to 'remember any particular past event'. Episodic memory, in contrast, is memory of 'the events of one's life' which includes 'spatial and temporal landmarks that identify the particular time and place when an event occurred' (Squire & Kandel, 2009, p. 118). Thus it might seem that episodic memory is straightforwardly autobiographical, while semantic memory, comprising factual

information about the world, seems 'detached from autobiographical reference' (Tulving, 1972, p. 389).

Drawing the episodic–semantic distinction by content or reference is not tenable, however. It should be obvious that there can be elements of *semantic* memory that have an autobiographical reference: knowledge of one's name, of where and when one went to school, etc. Such memories are semantic in that their encoding and retrieval can indeed be (and often is) detached from any temporal or spatial setting or any other episodic features. Yet, in so far as they are memories of the facts of one's life, they are clearly autobiographical.

Tulving (1983) recognized that the original distinction by content or reference was 'inchoate, . . . rudimentary, imperfect, incomplete, and rather disorganized' (p. 27). To refine the distinction in light of new clinical observations, he therefore proposed that episodic and semantic (and non-declarative) memory are different memory *systems* marked by the *type of consciousness* with which their retrieval operates: episodic memory correlates with *autonoetic* ('self-knowing') consciousness, semantic memory with *noetic* ('knowing') consciousness, and procedural memory with *anoetic* consciousness (consciousness limited to present inputs) (Tulving, 1985). Both the episodic–semantic and the declarative–procedural distinctions are thus redefined in terms of the different phenomenologies of remembering, with each type of memory hypothesized to be using a distinct neurocognitive system.[3]

4.2.2 Where does autobiographical memory fit in?

Autonoetic consciousness, which is exercised in episodic recall, is 'the capacity that allows adult humans to mentally represent and to become aware of their protracted existence across subjective time' (Wheeler et al., 1997, p. 335). Exercising this capacity is often referred to as *mental time travel* (Tulving, 1983; Suddendorf et al., 2009; Klein, 2013a). The idea behind this somewhat fanciful label is that the phenomenology of episodic remem-

[3] The redefinition of the episodic–semantic distinction in terms of phenomenology rather than content has since been 'widely adopted by memory researchers and has shown [itself] to be a particularly fruitful way of generating testable hypotheses and theoretical models of . . . long-term declarative memory' (Klein, 2013b, pp.1–2).

bering involves, as it were, transposing oneself into one's own past, of consciously 'reliving' past events—and in this resembles the phenomenology of consciously simulating *future* events, a capacity which indeed seems to share a neural substrate with episodic remembering (Schacter et al., 2012—see § 4.4.1).

There is a decidedly Lockean flavour to the notion of autonoetic consciousness. Episodic memory understood as mental time travel approximates 'repeat[ing] the idea of any past action with the same consciousness [one] had of it at first' (Locke, 1690, II.xxvii.10). Autonoetic consciousness involves awareness not just of past events, but of *oneself experiencing* those events (see § 4.2.3). So, although episodic memory is no longer defined by its autobiographical reference, its redefinition in terms of autonoetic consciousness still makes it the most obvious class of memory to be associated with the self.

And indeed there is a considerable overlap between autobiographical memory and episodic memory even under its new definition. It is probable that most of our autobiographical memories are episodic in nature, and that most episodic memories are autobiographical. But the overlap is not complete. There may be episodic memories that are not strictly autobiographical, such as vivid recollections of a scene in a film. Of course, there is an autobiographical context to such recollections (one's having seen the film in such and such a cinema at such and such a time of one's life), but it seems plausible that one may in recalling a film scene have the vivid phenomenology of 'being there' without any connection to the surrounding events of one's own life. It may be said that such consciousness is not strictly autonoetic either (it does not involve 'self-knowing')—but its vividness resembles autonoetic consciousness more than the comparatively austere phenomenology of semantic recall. But I need not insist upon this point.

The converse case—autobiographical memories that are *not* episodic—is more important. As has already been mentioned, there are facts about one's life whose retrieval from memory does not require autonoetic consciousness and episodic memory. My name, my date of birth, my place of birth, certain facts about my career and places of residence, etc., are all the stuff of semantic memory, but are nonetheless autobiographical. Such memories are often *reinforced* by episodic recollections (e.g. recalling one's former address may trigger a visual recollection of what the street looked

like), but they can also be recollected as mere facts, without necessarily being accompanied by autonoetic consciousness. Such semantic memories are relevant to the present discussion not just because they are *about* oneself: the presence of such memories is crucial to a continuous *sense* of self, as is obvious when we consider what their loss would mean. The news media periodically report on cases of 'transient global amnesia'—some wandering individual having been found somewhere who has (or appears to have) forgotten his name, address, occupation, indeed every fact about himself. Whether these cases are all really cases of amnesia may be disputable, but the point is that it is easy to imagine that anyone afflicted with such a loss of semantic memory would have a seriously defective sense of self, in having forgotten *who* he is.

I am not alone in holding that some important components of autobiographical memory are semantic rather than episodic in nature. Along with factual self-knowledge of the kind just discussed, Klein and Nichols (2012) point out 'a second kind of semantic self-knowledge, knowledge of one's own traits':

> Research over the past twenty years has provided evidence that the semantic memory system contains a specific subsystem that stores information about one's own personality in the form of trait generalizations (for example, *Self: usually stubborn*). (p. 681)

Similarly, John A. Robinson (1992) includes a semantic element of 'self-description' in his account of autobiographical memory. Such self-description or trait generalizations may, of course, be augmented and reinforced by episodic recollections that illustrate one's particular traits. But the descriptive, semantic elements and the episodic recollections seem to be dissociable and accessible independently of each other. (This is often apparent in amnesia and dementia cases—see §§ 4.2.4 and 8.2.3).

For example, I may have a very vivid episodic recollection of a particular event, say, a triumphant and unexpected sprint victory on a school sports day. I may recall the phenomenology of the victorious sprint—the heat of a summer's day, the smell of mown grass, the shot of the starting gun, the feel of the Tartan track under my running shoes, the navy blue colour of my opponent's shirt as I overtook him just a few metres before the finish line. I could have this episodic recollection without recalling any facts about when or where the event remembered took place. More often, such

episodic recollection will be accompanied by knowledge of facts about the situation and other persons present which allow me to place the event more or less accurately in my biographical history: the name of the school, the name of the boy in the blue shirt, the fact that it was the same year in which some other event happened, and so on. But these are semantic memories. And they could, in turn, be recalled without any accompanying episodic memory, as could the mere fact of winning the hundred metres at the school sports day at such-and-such a school in the summer of 198x. Autobiographical memory, then, includes semantic as well as episodic elements, even if, in normal circumstances, these are likely to be mutually reinforcing.

One may perhaps go further. Arguably, there will be cases where procedural memory, too, can be 'autobiographical' or self-defining. Again this becomes obvious when one considers the loss of such memory. A concert pianist having, through some brain lesion, lost the ability to play the piano, or a *Tour de France* competitor having lost his capacity for cycling, or an artisan who apprenticed for years acquiring the fine motor skills necessary for her trade becoming similarly incapacitated—all these would experience not merely a loss of procedural memory but of part of their 'self'. That said, the import of such a loss of procedural memory would become apparent to the patient only in the presence of intact semantic and/or episodic memory (of having previously had the now defunct procedural capacity). Thus, I will now mostly limit my discussion to declarative (semantic and episodic) autobiographical memory.

4.2.3 The phenomenology of autobiographical memory

While the episodic–semantic distinction is no longer usefully cashed out in terms of *content* or *reference* of the memories in question, those terms still seem to be useful in defining *autobiographical* memory. In the simplest analysis, we could say that any memory, whether episodic or semantic (or procedural), is autobiographical whenever its content refers to oneself *as oneself* (*de se*)—one's biographical data, capacities, personality traits, and the events of one's life. But this does not seem quite sufficient to capture the phenomenology of autobiographical recall. If I were to memorize another's biographical details, catalogue his capacities and personality traits in detail,

and contrive—after his description and any other data I might assemble—vividly to imagine a vast number of events of his life from his perspective, would I then be accessing (replicating, simulating) that person's autobiographical memory? Or would there be something lacking in the *experience* of that memory *as* autobiographical? Would I, in short, have the same sense of 'mineness' in accessing another's autobiographical memory as with my own?

Perhaps, then, autobiographical memory is not defined merely by its content or referent, but also, again, by its phenomenology. An empirically useful way of characterizing that phenomenology is in considering what other psychological capacities are recruited in accessing declarative memory so as to produce an experience of that memory being mine, of my life, and not of another's. In their theory of autobiographical memory, Stanley Klein and colleagues (2004) suggest that there are three such capacities that 'transform declarative knowledge into an autobiographical experience' (p. 463): the ability to reflect on oneself ('self-reflection'), a sense of agency and ownership, and a 'sense of personal temporality' (p. 465). Drawing on clinical data, Klein and colleagues argue that these capacities are 'individually necessary and (perhaps) jointly sufficient for autobiographical memorial experience' (p. 468). They note that impairments of self-reflection in frontal-lobe pathologies and autism, disturbances of the sense of agency and ownership in schizophrenia and certain delusional states, and impairments of 'personal temporality' (the sense of one's continued existence through time) in certain amnesias all correlate with impairments of autobiographical episodic recall.[4]

Correlations do not, of course, in themselves favour one direction of causality or necessity over the other. It could be that the capacities discussed themselves depend on functioning autobiographical memory. But there are good reasons for the hypothesis that these capacities are prerequisites for autobiographical memory experience, rather than the other way around. The ability to self-reflect, which again taps into Tulving's notion of autonoetic consciousness[5]—and by which is meant merely a higher-

---

[4] Cf. chs. 7 and 8.

[5] And shares its neural basis in the frontal lobes—self-reflection, autonoetic consciousness, and episodic memory are all associated with frontal-lobe activity (Wheeler et al., 1997; Klein et al., 2004).

order cognitive capacity that allows one to be the object as well as the subject of one's consciousness—is basic to all forms of self-knowledge, not just autobiographical memory. It is conceivable that one could have the capacity for self-reflection in the present without having any autobiographical memories. But the converse case is not conceivable: a memory would not qualify as autobiographical if one did not know or sense that it was about oneself.

The sense of agency and ownership is, perhaps surprisingly, dissociable from self-reflection. A disturbed sense of agency is evident in delusions of control, where 'patients . . . experience their own thoughts and actions as having been caused by an external agent' (Klein et al, 2004, p. 465). Loss of a sense of ownership of one's thoughts occurs in the classic schizophrenia symptoms of thought insertion and auditory hallucinations (see § 8.3.4). As Klein and colleagues point out, impairments of episodic memory are also 'disproportionately pronounced' in schizophrenia (ibid.), suggesting a link between loss of a sense of ownership and impairments of episodic memory. However, the relationship between a sense of ownership and episodic memory is a complex one. It is possible to have episodic memories *without* a sense of ownership. Klein and Nichols (2012) report on such a case, the patient R.B., who after a brain injury lost the sense of ownership of his episodic memories.

> I can picture the scene perfectly clearly … studying with my friends in our study lounge. I can 'relive' it in the sense of re-running the experience of being there. But it . . . did not feel like it was something that really had been a part of my life. Intellectually I suppose I never doubted that it was a part of my life.
>
> Things that were in the present, like my name, I continue to own. Having been to MIT had two different issues. My memories of having been at MIT I did not own. Those scenes of being at MIT were vivid, but they were not mine. But I owned 'the fact that I had a degree from MIT'.
>
> (p. 686)

Though R.B. has *semantic* autobiographical knowledge (such as having a degree from MIT), and he therefore *knows* the events of his episodic memory to be *his*, he does not *experience* them *as his own*. Thus, it seems that episodic memory does not in itself require—or confer—a sense of ownership, but that a sense of ownership is required for episodic recollections to be experienced as autobiographical.

R.B.'s case is relevant to the neo-Lockean notion of *quasi-memory*, memory-like states that have some proper causal connection to the events of which they seem to be memories, without however presupposing the rememberer to be the same person as the one who experienced the event (see § 2.3.2). Marya Schechtman (1990) argues against the possibility of quasi-memories by noting that individual memories aren't isolated, but involve many associations with, and references to, 'other parts of [one's] life and [one's] personality' (p. 81). From this, she concludes that a quasi-memory must either be stripped of these associations—in which case the quasi-memory will be qualitatively rather different from an actual memory, and so cannot serve as a substitute for actual memories in establishing psychological connectedness—or it must somehow reproduce all the 'personal elements' (p. 83) of actual memories. In that case, Schechtman goes on, 'the mineness of the experience seems to be part of the content of the memory' (ibid.). If that is so, the quasi-memory is either delusional (because the quasi-rememberer experiences it as 'his' when, in fact, it is not) or it presupposes personal identity (because it is indeed the quasi-rememberer's own memory), and so again fails to do the job of establishing psychological continuity without presupposing it.

R.B.'s case seems a clear counterexample to Schechtman's contention that 'mineness' is part of the content of the memory: a sense of ownership is precisely what is missing from the phenomenology of his episodic recollections. Thus, Klein and Nichols (2012) respond to what they call 'Schechtman's dilemma'—that quasi-memories rich enough to reproduce a sense of remembering are either delusional or presuppose personal identity—by pointing out that R.B.'s 'ownership'-less episodic memories are not delusions (they were factually corroborated), but nor do they involve the presupposition that R.B. is necessarily the same person whose episodic memories he seems to be experiencing—indeed, he appears to experience those recollections as though they were someone else's.

But we must be careful here to distinguish between what Klein and Nichols call 'the *sense* of personal identity' and the *presupposition* of personal identity. Yes, R.B. lacks a *sense* of identity with the protagonist of his episodic recollections. But his case arouses interest precisely because he and his clinicians *do* presuppose that he *is*—as a matter of *fact*, not as a matter of phenomenology—continuous with the person who had the experiences he

now remembers. What is so strange about R.B.'s phenomenology is that we know—and *he* knows, or can deduce—that these *are*, in fact, his own memories (even though they do not feel that way to him). R.B.'s case does not presuppose a sense of 'mineness' (that is a point against Schechtman's argument)—but for it to count as a non-delusional case, it does involve the presupposition that R.B. the rememberer is the same person as the R.B. who had the experiences he now recollects (which is a point in Schechtman's favour).

What is instructive about R.B. in discussing the nature of autobiographical memory is that there seems to be a clear dissociation between self-knowledge (such as R.B.'s knowledge of having been at MIT, which is intact) and a sense of ownership of episodic recollections (the scenes at MIT do not *seem* 'his'). Thus, R.B. provides evidence that a 'sense of personal agency/personal ownership'—second in Klein and colleagues' (2004) list of capacities subserving autobiographical memory—is indeed among the normal self-representational capacities at work in personal remembering, though it is noteworthy that this sense becomes apparent only by its absence or impairment:[6] R.B. experiences his recollections as odd because there is something missing from them, a self-representational capacity (and the phenomenology it gives rise to) whose operations, in non-pathological circumstances, go unnoticed.

Klein and colleagues' (2004) third prerequisite for autobiographical memory is having a sense of 'personal temporality'. They characterize this as the 'capacity to become aware of the temporal dimensions of one's own experience' (p. 466). On a quasi-Lockean view of the self, this sense of self-over-time would assume intact autobiographical memories. But the Lockean then faces a vicious circularity: that autobiographical memory is necessary for a sense of self-over-time, and a sense of self-over-time necessary for memory to be properly autobiographical. One way of avoiding this circularity would be to suggest that a sense of personal temporality is a result, not a prerequisite, of autobiographical memory. Klein and colleagues consider this hypothesis, but note that

> a review of the literature reveals that episodic memory loss is not necessarily associated with impairments of temporal consciousness. For example, patients with retrograde amnesia cannot remember their personal

[6] As it also is in schizophrenia (see § 8.3.4).

> past, but they can remember events occurring after the brain trauma that left them amnesic; other amnesic patients report other types of temporal gaps in their personal narrative. (p. 466)

Thus, not only can a sense of personal temporality be intact in cases of autobiographical memory loss, but it also seems to be precisely that sense of temporality that allows amnesic patients to *notice* their loss of personal memories. A sense of personal temporality does indeed seem to operate independently of the contents of autobiographical memory.

4.2.4 Re-enter the episodic–semantic distinction

An illustration of the foregoing is provided by Klein's (2013c) case of an amnesic patient, D.B.:

> D.B. was a 79-year-old man who became profoundly amnesic as a result of anoxia following cardiac arrest. Both informal questioning and psychological testing revealed that D.B. was unable to consciously recollect a *single* thing he had ever done or experienced from any period of his life. In addition to his dense retrograde episodic amnesia, he also suffered severe anterograde episodic memory impairment, rendering him incapable of recollecting events that transpired only minutes earlier . . . (p. 797)

D.B. thus had no accessible episodic autobiographical memories nor the means of forming new ones. Yet, Klein and his colleagues found that he retained *semantic* self-knowledge related to his personality traits:

> To test D.B.'s semantic self-knowledge, we asked him on two separate occasions to judge a list of personality traits for self-descriptiveness. We also asked D.B.'s daughter (with whom he lives) to rate D.B. on the same traits. Our findings revealed that D.B.'s ratings were both reliable and consistent with the way he is perceived by others . . . D.B. thus appeared to have accurate and detailed knowledge about his personality despite the fact that he had no conscious access to any specific actions or experiences on which that knowledge was based. (ibid.)

In addition to this dissociation between episodic and semantic autobiographical memory, Klein and colleagues also found a dissociation *within* D.B.'s semantic memory, for his ability to recollect general knowledge or factual information relating to his own life was likewise impaired by his amnesia. Only 'his knowledge of his own personality was intact' (p. 798).

It may be observed that 'knowledge of personality traits' is a tricky notion, since it is questionable whether anyone really has the stable personality traits we assume ourselves and others to have (Doris, 2002). But that question is not at issue here. For we do *think* ourselves to have stable traits, and it is the knowledge and recollection of that *self-image* that is in play here. Thus, hereinafter, Klein's term 'trait self-knowledge' may be understood as 'knowledge of one's trait self-image'.

Klein (2013c) observes that his patient R.B. (§ 4.2.3) likewise 'possessed both accurate and reliable trait self-knowledge' despite the impairments to his episodic memory, and concludes that '[p]erhaps, semantic trait self-knowledge provides the bedrock from which a *sense* of diachronicity springs' and hence, 'long-term memory (with the possible exception of semantic trait self-knowledge), due to its potential for loss *without* accompanying loss of sense of identity, appears unnecessary for a *sense* of personal identity across time' (p. 799, original emphasis).

But D.B.'s sense of self, taken globally, is nevertheless seriously impaired by his amnesia. Thus, Klein reports that D.B.

> was greatly troubled by the absence of information that, as D.B. describes it, "I don't know, but I should, shouldn't I?" (D.B. often broke down in tears over his inability to recollect knowledge of his personal past); information, in short, that failed to inform his subjective self-awareness.
>
> (p. 798)

Again, there is a dissociation of self-representational capacities in evidence here. D.B. has a sense of temporality and intact trait self-knowledge, but lacks other semantic self-knowledge and episodic memories—and is deeply troubled by this. On the system view, there *is* a self operating here—D.B. is able to situate himself in time, retains knowledge of his trait self-image, and, crucially, is *aware* that something important is missing from his self-representations. What is missing are the components of a Lockean or popular quasi-Lockean self: his personal memories (apart from trait self-knowledge). Without these, his *sense* of self is incomplete. It seems, then, that the quasi-Lockean intuition that memories are important for a sense of self over time is justified—but we must also note that the only way for D.B. to recognize their absence is for *other* conscious self-representational capacities to continue—troublingly for him—to function. Thus, the *Lockean* self does not exhaust the self.

4.2.5 Summary: the diverse nature of autobiographical memory

As this discussion shows, autobiographical memory is a much more complicated matter than it may at first seem. Episodic memories of the events of one's life form an important, but not exclusive, part of autobiographical memory. Memory that pertains to oneself and one's biography can also be semantic self-knowledge (in some cases, perhaps even procedural memory). Autobiographical memory thus straddles the boundaries of standard psychological classifications of memory.

But deficiencies in autobiographical memory also reveal unexpected dissociations *within* episodic and semantic memory. For semantic autobiographical memory, the case of D.B. shows that it is possible to have no memories of facts about oneself *except* knowledge of one's character trait self-image. Episodic recollection, meanwhile, involves different dissociable self-representational capacities: self-reflection, a sense of agency and ownership, and a sense of temporality. As shown by R.B.'s case, the sense of ownership can be absent despite the availability of episodic recall (as well as semantic self-knowledge).

The diverse nature of autobiographical memory processes and their possible faults pose a first problem for the view that memory supports the self over time. On the contrary, it seems that more basic self-representational processes are required to support autobiographical memory. But there is worse to come: there are constraints upon the availability and accessibility of autobiographical memory, and important questions about its reliability. I discuss these issues in the next two sections of this chapter.

4.3 Availability and accessibility constraints

It is a common enough observation that not all individuals have equally good autobiographical memory. Some of us have excellent semantic memory, others may have particularly rich and vivid episodic memories without, perhaps, being able to place them very accurately in their semantic context. Some people's memory seems generally poor; some care greatly about this, others do not. Even barring the serious progressive memory loss

associated with dementias such as Alzheimer's disease, memory tends to degenerate with age, though it is often the ability to form *new* memories that is most afflicted, so that age-related memory loss manifests itself most acutely with reference to recent events, while its sufferers seem to retain autobiographical memories of long past events with great clarity.

But no human being, not even one gifted with extraordinary recall for both facts and events, retains a complete record of his or her life in autobiographical memory. Whereas the aptly named android 'Data' in *Star Trek: The Next Generation* is incapable of forgetting any fact or event he is ever exposed to, no such total recall is known outside fiction, and nor would it be desirable or to its bearer's advantage. As Marya Schechtman (1996) puts it, were one to strive to recall 'each and every event befalling the human being in full detail—such a goal would result in a Proustian paralysis in which the recognizable general features required for a coherent story [of one's life] would be lost in the richness of information' (p. 124).[7] The intellectual paralysis of one who remembers the details of every minute of his life but is unable to abstract or generalize from them is nicely illustrated in Jorge Luis Borges' (1942) short story 'Funes the Memorious'.

There is, and indeed needs to be, selectivity concerning both what facts and events are encoded in autobiographical memory and how and when (and why) they are retrieved. Nor are our autobiographical memories distributed evenly across a lifetime. I shall look at these issues in turn.

4.3.1 Encoding and retrieval[8]

If, as just suggested, autobiographical memory is selective, the question arises: what factors contribute to an event or fact in short-term memory being committed to long-term memory for later autobiographical recall, instead of forgotten? For instance, I currently have a short-term memory of an amiable but trivial conversation I've just had with the staff at my favour-

---

[7] Schechtman's use of 'story' here is not metaphorical; the remark comes from her 'narrative self-constitution' account, which I discuss in the next chapter (§ 5.3.4).

[8] The terms 'encoding' and 'retrieval' for the formation and recall of memories, borrowed from computing language in the 1960s, reflect the then newly fashionable paradigm of the brain as an information-processing system (Brown & Craik, 2000).

ite coffee shop, which is already fading into just a recollection of the *fact* of having had a friendly exchange of words, rather than of *what* was said; and soon I'll have forgotten both the fact and the content of that conversation completely. It will form no part of my long-term autobiographical memory. But why not? What factors would have to have been present for this episode to be stored in autobiographical memory?

Scott Brown and Fergus Craik (2000) identify a number of factors for 'good encoding' (p. 96) of memories in general: how a fact or event relates to one's goals and purposes; how the memory is rehearsed ('elaborative' rehearsal, involving some cognitive processing of the information, is more effective than mere 'maintenance' rehearsal; so is rehearsal distributed over a longer time-span); the organization of memories (where constructing 'meaningful relationships between items' (p. 97) increases effectiveness of encoding); the distinctiveness of the episode or fact in question; and, overall, how the encoding of memories is 'guided by an individual's prior knowledge, values, and expectations' (p. 98). Relating these factors to my coffee-shop episode, we notice that some components of good encoding are missing. Though a friendly exchange of words with the staff may be conducive to my immediate goal of obtaining coffee, that goal ceases to be once the coffee has been obtained. For the same reason, the memory of the episode wouldn't usually be rehearsed (though it has now been by my writing about it). As for organizing my memories, the episode may serve to consolidate a generic semantic memory of the staff at this particular coffee shop being generally personable, without my recalling the specifics of today's conversation—because it was not distinctive. Indeed, an episode seriously conflicting with my prior knowledge and expectations of the coffee shop would have been much more memorable.

But though this humdrum coffee-shop episode will not reside in my long-term episodic memory, I just suggested that it may help consolidate some more general semantic memory about the coffee shop. Likewise, many episodes that we recall for only a short while and that fit some but not others of Brown and Craik's 'good encoding' factors may contribute to autobiographical memory without ever being recalled as individual events in Proustian detail: our memory of certain periods in life or repeated activities is built up in this way.

Thus, Martin Conway and David Rubin (1993) theorize that our 'autobiographical knowledge base' is structured into three layers differing in the level of detail of their contents. The most general layer, one's autobiographical knowledge of 'lifetime periods', will include general knowledge associated with the different periods of one's life, such as the significant others and overarching goals that define each lifetime period and which 'may represent major thematic divisions of a person's life' (p. 105)—years at school, at college, time spent working at X or living with Y, and so on.

The second, somewhat more specific layer of autobiographical memory is that of 'general events'—repeated events such as a particular work routine or extended events like holidays, which are organized thematically and chronologically and will generally be attributable to a lifetime period at the top layer. They also provide context for event-specific memories, which mark the third, most detailed layer of autobiographical memory. These memories are *episodic*, relating to a particular event, rather than thematic, and are typically recalled for a much shorter time span than general event memories and lifetime period memories.

An important consequence of this plausible structuring of autobiographical memory is its *mutability*—at the most detailed, event-specific level, episodes are typically recalled for a few hours or perhaps days, then may contribute to general event memories lasting some months or years, some of which in turn feed into the most general and abstract autobiographical knowledge of lifetime periods. Thus, my coffee-shop episode ceases as episodic memory after few hours but contributes to the general event memory of repeatedly going to that coffee shop during the lifetime period of being a PhD student at Sheffield. Autobiographical memory, then, is subject to constant change: it is not merely a question of accumulating more and more memories over a lifetime, but of organizing and reorganizing one's autobiographical knowledge base. In general, the less recent an episode, the less likely is it to be retained as episodic memory, though it may shape and alter general event memories.

Of course, there are some events that may be recalled episodically for a much longer time than our humdrum, day-to-day experiences. This is especially the case with some early-life memories (see § 4.3.3) and memories of significant events—one's wedding, say, or the funeral of a close friend or relative, or any other event that in some way has a particular salience.

Here, two factors from Brown and Craik's (2000) catalogue are decisive: the distinctiveness of the event in question, and the rehearsal of the memory. And both these factors will be reinforced by another element crucial for retaining episodic memories: affect. The presence of strong (positive *or* negative) emotions contributes to the salience of our experiences, and consequently to the likelihood of our remembering them.[9] According to this 'intensity hypothesis for memory' (Hardcastle, 2008, p. 63), not only does the presence of a strong affective response reinforce the memory of an event,[10] it also makes us more likely to retell (in Brown and Craik's terms, 'elaboratively rehearse') the episode in question, thereby once more rehearsing the affective response that accompanied the original experience, and so again reinforcing our memory of it (fig. 4.1).

The rehearsal of memories in this way is a process involving both the retrieval and subsequent re-encoding of a memory. But how does retrieval work in general? Clearly, we do not *constantly* recall everything we *can* remember (this would lead to a cognitive overload much worse still than being able to remember everything we experience). What, then, are the conditions under which memories are retrieved? Once again, it was Tulving

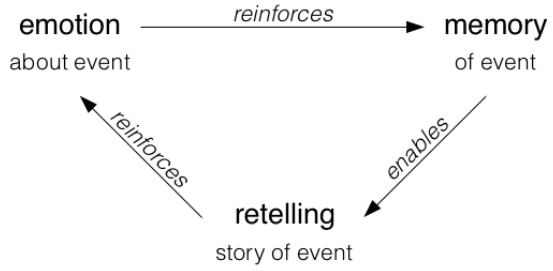

Figure 4.1 Hardcastle's (2008) model of affect–memory interactions.

[9] There are exceptions. In cases of post-traumatic stress disorder, the affective salience of an experience can block rather than enhance memory encoding ('mnestic block syndrome', Markowitsch, 2000, p. 470), though the mechanisms of this are not yet fully understood, and some traumatic experiences do produce especially vivid episodic memories (Schooler & Eich, 2000).

[10] The neural substrate of such reinforcement is found in the projections from limbic structures (associated with affect responses) to the orbitofrontal cortex, which mediates input from the affect and reward systems for learning, memory, decision-making, and planning (Hardcastle, 2008; Kringelbach, 2005).

(1983) who introduced an important distinction here, between the *availability* of memories (a memory trace has been encoded) and their *accessibility* (the memory can actually be accessed/'retrieved'). The accessibility of memories is dependent on what cues are present at the time of retrieval. And what cues are effective for the retrieval of a particular memory depends on 'the extent that information in the cue was incorporated in the [memory] trace . . . at the time of its original encoding', in other words, 'successful retrieval depends on the similarity of encoding and retrieval operations' (Brown & Craik, 2000, p. 99; cf. Morcom, 2014).

One's affective state is one important factor here—a memory encoded while one is joyful or despondent will more likely be recalled when one is once again in a similar affective state. Other information encoded along with memories that can be used as retrieval cues includes co-occurring activities and/or cognitive processes—doing household chores, or listening to a piece of music—or environmental, especially spatial, features—I may be more likely to recall my coffee-shop episode when I find myself in the vicinity of the same coffee shop again. Besides this, 'the depth (or type) of initial encoding also plays a major role' in the retrieval of memories (Brown & Craik, 2000, p. 99), so for strongly encoded memories, especially those involving a strong affective component, a relatively weak cue will suffice for their retrieval.

But is 'retrieval' really the right word for the process of accessing our memories? For semantic memory, it may be: the recall of facts can plausibly be described as a retrieval of a memory trace. But for episodic memory, 'at least some memory imagery is due to constructive processes rather than simple retrieval and reactivation of memory traces' (Robinson, 1992, p. 238). Indeed, it seems that most or all episodic recollection is *constructive* or at least *reconstructive* (Schacter & Addis, 2007; Klein, 2013b): unlike a film, in which every detail of a scene is replayed exactly as encoded, episodic memory works by reconstructing mental imagery from accessible memory traces and present cues. This is no mere retrieval, but a self-representational process of a more complex kind that involves a number of dissociable capacities (§ 4.2.3). And if episodic recollection is reconstructive in its contents, that reconstruction may be more or less faithful to the original event that is being remembered; in some cases it may be partly or entirely

fictitious or false. I discuss reconstructive remembering, and its implications, in more detail below (§ 4.4).

To sum up: autobiographical memory is selective at both the encoding and the retrieval stages. What facts and events are committed to memory depends on a variety of factors, both internal and external, the most important for encoding event memories being an occurrence's distinctiveness and/or one's affective response to it. Once memories are encoded, and therefore 'available' in Tulving's sense, they are not necessarily always *accessible*. Our constantly updated and reorganized 'autobiographical knowledge base' is not designed for random-access searches for specific memories (as we notice in those maddening moments when we're trying to remember minor autobiographical details, like the name of a former neighbour whose kids were keen to partake in the harvest from our pear tree). Their retrieval depends on effective cues, which again range from one's own present affective and cognitive states to external, environmental factors, and may be quite unanticipated and unplanned (as when the pear-loving kids' father's name—Tony—suddenly pops back into conscious awareness). Finally, the recollection of episodic memories is a constructive or reconstructive process rather than one of mere 'retrieval'.

4.3.2 Types and degrees of awareness[11]

As discussed above (§ 4.2), semantic and episodic memory (in Tulving's revised definition) differ in the character of our awareness of each. According to Tulving (1985), awareness of semantic memories, including those about oneself, is merely *noetic*; noetic consciousness need not recruit the sense of ownership and agency or of personal temporality that appear to be subsidiary capacities for *autonoetic* consciousness, which is exercised in episodic recollection.[12] This poses a problem for Locke's (1690/1706) criterion that past actions constitute the self 'as far as any intelligent being can repeat the idea of any past action *with the same consciousness* it had of it at

---

[11] I here follow Wheeler et al.'s (1997) usage of *consciousness* as referring a 'general capacity' and *awareness* to 'a particular manifestation or expression of this general capacity' (p. 335).

[12] But note that patient R.B. (§ 4.2.3) does attribute 'ownership' to his *semantic* self-knowledge.

first, and with the same consciousness it has of any present action' (II.xxvii. 10; emphasis added). This is only possible (if at all) in the case of autonoetic awareness, where a sense of ownership and agency comparable to that of present awareness accompanies the episodic memory. (As noted in § 4.2.2, the notion of autonoetic consciousness has a decidedly Lockean flavour.) But this would mean once more excluding semantic memories from the Lockean self. Given the large number of semantic memories one has about oneself and the events of one's life that are not or no longer available for episodic recollection, yet still part of one's 'autobiographical knowledge base' (Conway & Rubin, 1993), this seems an undesirable consequence. For example, I have no episodic recollection of playing on the beach at Oostduinkerke at the age of just under four, but the *fact* of having been on that summer holiday (corroborated by family records and photographs) does form part of my autobiographical knowledge base. But, because of my lack of episodic recall, it would not form part of my Lockean self.

Within autonoetic consciousness, awareness varies also by degree. Most of us are familiar with having some vague memories of past events without quite being able to recollect the details of these episodes. Such vague memories may sometimes, if more effective cues are produced, resolve into sharper recollections. Or they may in time become inaccessible as episodic memories, their trace merely adding to the semantic memory of what Conway and Rubin call 'general event' knowledge.

These varying degrees of our awareness of past events are explicable by a number of factors already discussed—the strength of the original encoding, the presence or absence of effective cues for recall, the activity of Klein's (2013b) 'enabling' systems for episodic recollection—and also the temporal distance of the event in question (cf. § 4.3.3 below). This poses another problem for Locke's 'same consciousness' criterion. Even supposing that autonoetic awareness of a past action satisfies the condition of being the same *type* of consciousness, for many instances of episodic memory, the same *degree* of awareness cannot be assumed.

Does this matter? It could be argued that so long as an episodic recollection is veridical (and this is a big if—see § 4.4), the degree of awareness one has of the past event is unimportant for its forming part of one's Lockean self: what matters is that one recalls oneself, the present agent, as the agent of the past episode. This might not satisfy Locke himself (though it is

somewhat unclear quite what he means by 'same consciousness'), but it would satisfy many latter-day Lockeans in that 'psychological continuity' between the past event and the present is maintained.

However, there is another problem. For it can be argued that whatever the degree of autonoetic awareness one has of an episodic memory, it *never* is the same type of awareness as that which one has of a present action. Episodic recollection makes use of one's sense of personal temporality (Klein et al., 2004), which means that as well as being aware of the memory, one has a 'feeling of pastness' (Fernández, 2008) about it. It is precisely that sense of recalling an event as a *past* event that makes it phenomenally different from any present experience. As it happens, Locke (1690/1706) himself makes that point in his account of memory:

> the mind has a power, in many cases, to revive perceptions, which it has once had, *with this additional perception annexed to them, that it has had them before*. (II.x.2, emphasis added)

Indeed, if we did not have a 'feeling of pastness' accompanying episodic recollection, we would be hopelessly confused between our personal past and present. But that means that Locke's 'same consciousness' is never ensured.

There are other ways in which the awareness of an episodic recollection can differ from present experience. As Mohan Matthen (2010) notes, some of us may, in some cases, have episodic memories from an 'observer' or allocentric perspective, in which 'you yourself are one of the things in the image, and you view yourself doing things in the way somebody else would' (p. 12). We do not, except in pathological cases of depersonalization or schizophrenia (see ch. 8), have awareness of the present in this way. Non-pathological present awareness of oneself always has an egocentric perspective.

Finally, there are episodic recollections in which one's sense of ownership—without being completely lost, as in the case of R.B. (§ 4.2.3)—is weakened. A vague and diffuse childhood memory may often, even with intact autonoeticity, feel detached from one's present self, almost as if it belonged to a different person. This benign feeling of detachment probably owes much to the passage of time—and it is in this context that the nature and structure of autobiographical memory present some more potentially troubling characteristics.

4.3.3 Distribution of autobiographical memories over a lifetime

Another constraint on autobiographical memory is the uneven distribution of our memories over a lifetime. It has already been established that we do not have equal availability and accessibility of memories from all periods in our life. In view of Conway and Rubin's three-layer model of autobiographical memory discussed above (§ 4.3.1), one might envisage a temporal distribution of autobiographical knowledge in which memories of recent events (those that may still be episodically available) are the most numerous, with the number of accessible memories then decreasing as a function of how much time has passed between the event being recalled and the moment of its recall. As we shall see shortly, there is evidence to support this view, but only up to a point.

For it will be recalled that there are significant events in our lives the memories of which are more richly (often affectively) encoded, and which are therefore not forgotten at the normal rate. Thus, one might visualize the distribution of autobiographical memories over time as a descending curve interrupted at irregular, idiosyncratic intervals by spikes of significant memories. But memory research paints a somewhat different picture.

David Rubin and colleagues (1986) conducted a meta-analysis of several earlier, independent studies of the number of memories recalled by 50- and 70-year-old participants. By aggregating the data sets of these studies, they obtained an S-shaped curve for autobiographical memory distribution (fig. 4.2), which

> represents the contribution of three factors: a retention function, a reminiscence factor, and a childhood amnesia factor. The general picture they present is that from adolescence into old age there is a moving 20-year memory window which exhibits a generic forgetting rate independent of the person's current age. However, that rate cannot be extrapolated back to infancy. First, people aged 40 and older exhibit a reversal of the trend for those years corresponding to late childhood through adolescence and into early adulthood. Rubin et al. attribute this to a reminiscence factor. Second, adults of any age report disproportionately fewer memories from the preschool years than would be projected by a uniform forgetting rate. That is a manifestation of childhood amnesia. (Robinson, 1992, p. 226)

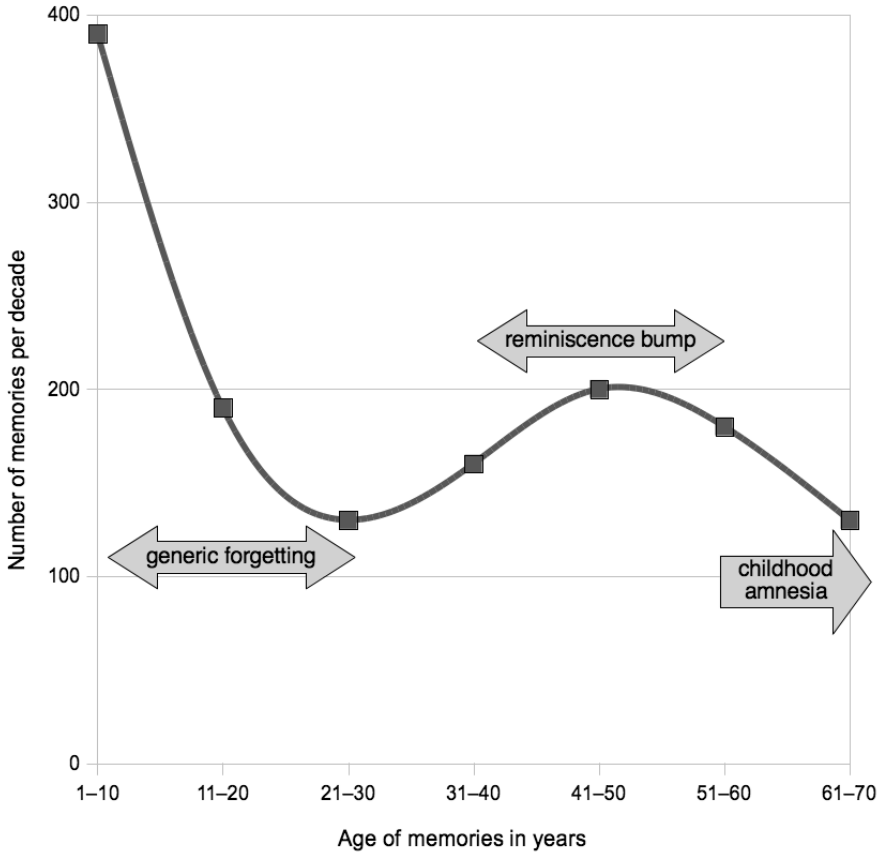

Figure 4.2 Distribution of autobiographical memories over a lifetime in 70-year-old respondents: aggregated data from four studies analysed by Rubin et al. (1986).

Thus, besides more-or-less linear 'generic forgetting', there are two other factors in play in shaping the autobiographical memory distribution curve: a 'reminiscence bump' (Conway & Rubin, 1993; Roberts & Feeney, 2009) between late childhood and early adulthood, and so-called 'childhood amnesia'.

The reminiscence bump coincides with a lifetime period that is generally replete with 'formative' experiences—school, college, first romantic relationships, leaving the parental home, etc. This might lead to such memories being selectively retrieved more frequently, i.e. by reminiscing (Rubin et al., 1986). More likely, however, is that memories from this lifetime period are recalled and rehearsed not for nostalgic reasons but for

their salience, 'the central place those events have in the person's life story or self-narrative' (Robinson, 1992, p. 231, citing Fitzgerald, 1988). Other proposed explanations of the bump appeal to the relative cognitive novelty of events in that period of one's life, biological maturational factors, and cultural normative expectations or 'life scripts' (Berntsen & Rubin, 2002). I will return to this debate in the next chapter (§ 5.2.3).

Before the bump (i.e. to its right in the temporally reversed curve in fig. 4.2), the memory curve tails off and ends rather abruptly some time in the first decade of life (usually around the age of three years), reflecting the absence of any autobiographical memories of early childhood, a phenomenon known as childhood amnesia (Robinson, 1992; Conway & Rubin 1993).[13] Here too a number of competing explanations have been attempted. Katherine Nelson (1993) reviews four standard theories of childhood amnesia (repression, forgetting, no memories, inaccessible memories) in light of evidence from developmental psychology and finds all of them either lacking in evidence or contradicted by it. She then considers the social-interaction hypothesis (Hudson, 1990; Pillemer & White, 1989), according to which autobiographical memories are rehearsed and retrieved in narrative conversations with parents and other care-givers. While admitting that such narrative memory talk is not universal, and not *culturally* universal, Nelson (1993) endorses a version of this hypothesis which one might call social-functional: 'the original functional significance of autobiographical memory is that of *sharing memory with others*, a function that language makes possible' (p. 376, emphasis added).

But whatever the exact mechanism of childhood amnesia (and a combination of cognitive, developmental, and social factors seems plausible), its existence limits the availability of autobiographical memory to about age three upwards.

---

[13] Given the usual meaning of 'amnesia' as *loss* of memory, this label is somewhat misleading, because it implies that previously formed memories are no longer available, whereas the explanation of childhood amnesia may well be that no persistent autobiographical memories are encoded in the first place. We should therefore, in this context, understand 'amnesia' simply to mean *lack* of memory.

4.3.4 Availability and accessibility constraints: summary

The availability and accessibility of autobiographical memory is, first of all, governed by processes operating at the time of both the encoding and the retrieval of memories. The significance and salience of an event or fact, especially where underwritten by affective responses, enhance its committal to memory. But for a memory to be accessible as well as available, similarly salient and significant processes must be in operation for its retrieval.

Encoding and retrieval processes also determine what kind of awareness we have of an autobiographical memory, whether it is recalled semantically, with noetic awareness, or episodically, with autonoetic awareness. And in episodic recollection, the degree of vividness and detail of the recall will vary according to factors such as time elapsed since the memory was encoded, the effectiveness of present cues for the retrieval of the memory, and again the event's significance and affective resonance.

More globally, the distribution of autobiographical memories over a lifetime appears to be governed by three factors. Generic forgetting means that the most recent period of one's life is generally the most present in autobiographical memory. Reversing this general trend, there is a 'reminiscence bump' of memories from early adulthood, adolescence, and late childhood. And childhood amnesia means we generally hold no autobiographical memories at all from the first three years of our lives.

Taken together, these constraints suggests that a Lockean self, made up of those autobiographical events we can remember (whether 'with the same consciousness' as at the time of the event or with any degree of awareness at all), is uneven in both its temporal distribution and its accessibility, highly variable over time, and (depending on how stringently we apply the 'same consciousness' criterion) rather patchy.

4.4 Reconstruction and reliability

At the beginning of the preceding section I noted what I took to be a commonly known fact, for which I then presented psychological evidence: that no one retains a *complete* record of the events of his or her life in autobio-

graphical memory. I shall now argue that it is also highly unlikely that anyone holds a completely *accurate* record of the events of his or her life.

How reliable is autobiographical memory? We can have accurate recall of some *semantic* autobiographical memories (important dates in one's life; schools and colleges attended; the names of family, friends, lovers, colleagues, pets; etc.). Other semantic memories may be less accurate—for instance, I may not accurately recall the year or month I went on a particular holiday without, where available, checking my records.[14] And in *episodic* recollection, probably only very recent or the most salient memories are recalled with anything approaching accuracy. For, in addition to the constraints on availability and accessibility just discussed—the complex factors that attend the encoding and retrieval of autobiographical memories—the *reconstructive* nature of most of our episodic recollections seriously constrains their reliability. I will now discuss this important feature of episodic remembering, and then sketch a continuum of reliability among reconstructed episodic recollections.

4.4.1 Remembering as reconstruction

The idea that memory involves constructive or reconstructive processes is far from recent. Giambattista Vico (whose lifetime overlaps with those of both Locke and Hume) may be noted as a precursor for his observations, in his *New Science* (1744), on what he sees as the triadic nature of memory:[15]

> Memory thus has three different aspects: memory [*memoria*] when it remembers things, imagination [*fantasia*] when it alters or imitates them and invention [*ingegno*] when it gives them a new turn or puts them into proper arrangement and relationship. (1948 edn, p. 280)

Writing two centuries later, and on the basis of empirical research in social and cognitive psychology, Frederic Bartlett (1932) is widely cited as the first

---

[14] An interesting question arising in this context is whether such external self-referring records amount to something like an 'extended self'. I will address this point in my final chapter (§ 9.2.2).

[15] I owe this discovery to Neimeyer and Metzler (1994), who in turn credit M. J. Mahoney.

modern psychologist to have espoused a reconstructive or constructive theory of memory:

> The first notion to get rid of is that memory is primarily or literally reduplicative, or reproductive. (1995 edn, p. 204)

> Remembering is not the re-excitation of innumerable fixed, lifeless and fragmentary traces. It is an imaginative reconstruction, or construction . . . (p. 213)

Bartlett studied mostly people's repeated recollection of stories and images and found these to vary considerably over time. Such recollections share structural features with *episodic* remembering, which involves both imagery and a temporal sequence.

The view that episodic recollection is a reconstructive process is now widely accepted (e.g. Loftus, 1974; Neimeyer & Metzler, 1994; Schacter & Addis, 2007; Klein, 2013b). The precise mechanisms of reconstructive episodic remembering are, as yet, not so clear. There seem to be several possibilities concerning what and how much of our episodic recollections is reconstructed. There may be islands of episodic memory traces which, on retrieval, are imaginatively enhanced, with missing imagery filled in around a core episode. Or perhaps all retrieval is semantic and then used as a cue for imaginative episodic construction.

The second possibility would be consistent with Stanley Klein's (2013b) recent hypothesis that there are not, after all, discrete systems for semantic and episodic memory, but that it is *retrieval* operations—supported by the 'enabling systems' of self-reflection, sense of ownership and agency, and sense of temporality—that lend episodic recollections their episodicity:

> What makes a memory experience episodic or semantic is not the nature of the content, or the hypothesized system in which content resides while in "storage," but rather an act of temporal (or atemporal) awareness that becomes associated with the content *once* it has been retrieved. (p. 3)

Episodic recollection is then something like, as Rob Hopkins (forthcoming) puts it, 'imagining the past'. Much recent cognitive neuroscience offers support for this view, having found a significant overlap in brain activity between episodic remembering and imagining (for a review, see Schacter et al., 2012). Yet the overlap is not complete. For example, some patients with amnesia resulting from hippocampal damage are still able to imagine fu-

ture and fictitious experiences (Maguire et al., 2010; Cooper et al., 2011; Hurley et al., 2011). This is perhaps not surprising—in the absence of (semantic?) memory traces, a capacity for imagining is insufficient, on its own, for imaginative recollection. But, conversely, unilateral prefrontal lesions seem to result in impaired imagination but no impairment of episodic recall (Berryhill et al., 2010), which suggests that imagination is not always necessary for episodic recollection.

A plausible position may be that we sometimes have episodic memory traces, which are reconstructively enriched, but at other times reconstruct episodic memories wholly around semantic retrieval. But in both cases, reconstructive processes contribute to the episodic recollection we experience. Given this, it is easy to see how inaccuracies will creep into our recollections, as at the time of remembering we (unconsciously) fill in gaps in the recalled episode with more or less likely details.

Are such recollections still *genuine* memories? It seems plausible that a popular notion of the veridicality of memories appeals to a suitable causal link from the event being remembered to the recollection.[16] Adapting Grice's (1961) causal theory of perception to memory, George Botterill suggests that genuine memories 'are states which are causally related [in the right way] to the incidents of which they are memories' (Botterill & Carruthers, 1999, p. 37). Thus, the *gist* of an episodic memory may count as genuine, while some details are filled in that do not satisfy the causal requirement of being traceable back to the event in question. I will call this largely benign filling-in 'embellished recall'.

More serious instances of unreliable recollections are misremembered, confabulated or outright false 'memories'. Thus emerges a continuum ranging from largely reliable to wholly unreliable recollections, which I now discuss in order of decreasing reliability. And it should be noted that, without recourse to external evidence, we have little ability to discriminate between reliable and unreliable recollections (Ross & Buehler, 1994).

---

[16] An experimental survey to test this presumed popular notion is in preparation. A pilot study conducted with 15 philosophy postgraduates and final-year undergraduates, while involving too small and homogeneous a sample to allow generalization, suggests that the right kind of causal connection indeed seems crucial for a 'vivid image' to be deemed an instance of 'remembering'.

4.4.2 Embellished recall

In ordinary episodic recollections, many details will be missing: either because they were not encoded in memory in the first place, or because they have since been forgotten, or because retrieval conditions are less than perfect. At the same time, the rememberer will have some generic knowledge relevant to the episode in question, as well as memories of similar episodes. It is plausible, therefore, that when one attends closely to a partially remembered episode, the details missing from the memory trace are filled in from these other sources.

Suppose I try to recollect a dinner at a restaurant with a group of friends that took place some months ago. I remember the faces of my friends and where everyone sat in relation to each other. I remember some of what I ate. I seem to remember fragments of what was talked about over dinner, perhaps even snatches of the conversation that were particularly memorable, because funny or outrageous. I do not remember exactly what clothes people were wearing, what the waiter looked like, what everyone else ate and drank, or anything about the decor of the restaurant. But as I try to recollect the scene, I fill in some of these gaps quite easily from other parts of my autobiographical knowledge base. Mike doesn't drink red wine, so I 'remember' him drinking white. Josh usually wears knitted jumpers, so I picture him in knitwear. And from relevant generic knowledge, my recollection gains a white table cloth, a small, dark-haired Italian waiter, some gauche wall decorations, and a scene over splitting the bill in which those wanting to pay by cash hand over banknotes to those wanting to pay by card. And so forth.

Such instances of embellished recall are normally benign. Embellished recollections aren't exactly *false*. Though they contain elements from one's autobiographical knowledge base that don't derive from the precise episode that is currently being recalled, these elements are nevertheless autobiographical and usually veridical (such as Mike's preference for white wine). Other reconstructed details may be so trivial that it is of little importance to the rememberer whether or not they are veridical (such as the colour of the table-cloth)—the original detail of the episode having been forgotten (or not having been encoded in the first place) precisely because it was of no particular importance. Embellished episodic recollection, while

not accurate in its details, can thus, in many cases, still be said to be genuine remembering, at least in its essentials.

4.4.3 Misremembering

A more pronounced manifestation of the 'filling-in' that occurs in embellished recall is what we ordinarily term 'misremembering'. Suppose that, in recollecting the restaurant scene, I seem to remember that my friend Charlotte was present at the dinner, when she wasn't. I probably have veridical recollections of her at dinner on other occasions; however, placing her at table in my recollection of this particular dinner seems a more serious departure from the facts than the filling-in of the colour of the table-cloth, for example. On a social occasion like a dinner with friends, it is no mere trifle who was and who was not present. In such a case, it would be too generous to speak of proper remembering, since one essential feature of the recollection is clearly a misremembered one.

Yet, in ordinary circumstances (but not, for instance, the more unusual circumstance of giving evidence in a court of law), even such misremembering is usually benign in its consequences. In particular, we tend to admit to having misremembered, and to correct our memories accordingly, when presented with evidence or testimony contradicting our recollections. Thus, if everyone else present at that dinner assures me that Charlotte wasn't present, or Charlotte herself points out that she couldn't possibly have been there as she was out of town that evening, I would be happy to accept that my memory was unreliable on that point and conclude that my episodic recollection mixed up different events. Such admission of misremembering and revision of memory is not so likely to be forthcoming in pathological cases of false remembering, which I'll consider next.

4.4.4 Confabulation

Besides the embellished recall and misremembering just described, which I contend happens to most of us some of the time, more serious defects of episodic recollection occur in certain pathologies. Patients with frontal-lobe damage are prone to *confabulating* episodic 'memories', with unhesitating confidence and often in considerable detail, most or all of which are untrue

(Conway & Rubin, 1993; Baddeley et al., 2009). Such confabulation is usually marked both by the patients' clearly believing their own story (there is no deliberate deception) and by the fact that to those around them it is often completely obvious that the patients' 'memories' cannot be true.

> Consider, for example, patient RR, who had extensive bilateral damage to his frontal lobes following a driving accident . . . When asked about the accident, he happily provided a detailed account that involved his getting out of his car and carrying out a polite but extremely repetitive conversation with the driver of the lorry that had hit him, with each apologizing to the other multiple times. He had in fact been unconscious for a lengthy period following the accident and could almost certainly not remember it. He was no longer capable of driving and gave a totally implausible account of how he had subsequently driven himself to the rehabilitation center . . . (Baddeley et al., 2009, p. 158)

The cause of this kind of spontaneous confabulation is usually neurological damage to the frontal lobes, which have been associated with *executive function*, our ability to control and manage cognitive processes. Baddeley and colleagues theorize that executive dysfunction resulting from frontal-lobe damage leads to the patient's having difficulties both with generating effective retrieval cues and with evaluating his own 'memories'. This is how confabulation, along with the patient's belief in the veracity of his confabulated memories, comes about. (Note that, although there is still a causal chain from the event to its confabulated 'recollection', we obviously no longer consider confabulations to be genuine memories, because the causal link—from accident to frontal lobe damage to confabulation—does not obtain 'in the right way'. Causal connections are necessary but not sufficient for genuine memories.)

Here then we have a defect of episodic autobiographical memory that is the result of a fairly well-understood pathological condition. Thus, one might think that any way one's sense of self is adversely affected by these confabulations is likewise a pathological symptom. Indeed, executive dysfunction, however caused, may be associated with deficits of the self in other ways.[17] There are, however, cases of false memories that are not the result of such clearly defined neurological injuries. What I'll discuss next are such false memories, particularly of one's childhood, which often refer to

---

[17] See §§ 7.2.3, 8.2.2, 9.2.1.

more prolonged events than those in confabulation, in some cases with troubling consequences.

4.4.5 False memories

I must begin this discussion with a confession. Though a tolerable sprinter in my school days (especially when running for a bus), I never did win the hundred metres against my peers. Nor do I have a detailed memory of any particular school sports day. The vivid episodic recollection I describe in § 4.2.2 is a fabrication. Yet, in making up that story of the fictitious sprint victory, I readily imagined the sensory details described there. And I can readily do so again: I seem now to have an episodic pseudo-memory of this made-up event.

The apparent ease with which episodic memories can be fabricated is supported by evidence from psychological studies. Elizabeth Loftus (1996; 1997) describes a number of memory studies in which false childhood memories were successfully implanted in between a fifth and a quarter of participants. The subjects were given vignettes of four episodes, three of which were about events from their own childhood (obtained from participants' parents), one was fabricated by the experimenters (participants' parents having confirmed that no such event had occurred in the participants' childhood). Participants were then interviewed several times about what they recalled of the events in question. In one such study,

> subjects remembered something about 89% of the true events during the first interview. Somewhat higher percentages were remembered during the second (93%) and third (95%) interviews. As for the false events, . . . no subject recalled these during the first interview, but 25% did so by the third interview. (Loftus, 1996)

Loftus reports that some of the participants who 'remembered' the false event at interview gave considerable detail of themselves in the fabricated episode, suggesting that the rehearsal of a false memory creates an autonoetic awareness like that of recalling an actual event.

The fact of actively *imagining* the fabricated event has a reinforcing effect on one's confidence that the event occurred. In another study, Loftus and her colleagues asked participants to

> indicate the likelihood that certain events happened to them during their childhood. The list contains 40 events, each rated on a scale ranging from "definitely did not happen" to "definitely did happen." Two weeks later we asked the participants to imagine that they had experienced some of these events. Different subjects were asked to imagine different events. Sometime later the participants again were asked to respond to the original list of 40 childhood events, indicating how likely it was that these events actually happened to them. (Loftus, 1997)

Participants who had been asked to imagine a particular event were subsequently significantly more confident that such an event had occurred than participants who had not been asked to imagine the event. So, were I to forget having deliberately made up the sprint episode to illustrate a chapter of my thesis, I might well, by virtue of having imagined it so vividly (and now repeatedly), come to believe that it was a true memory. In this imaginative rehearsal of a fabricated memory, Vico's *fantasia* and *ingegno* aspects of memory find empirical corroboration.

Should all this worry Lockeans? One might say that the examples of false memories given here—a fabricated memory to illustrate a doctoral thesis, some laboratory studies in which only up to a quarter of subjects were susceptible to the creation of false memories, all designed specifically with the aim of creating false memories—show only that false memories *can* be created in this way, in some people, under specific conditions. They do not suggest that this is what routinely happens in people's lives.

But there is a back story to Loftus's work. In the 1980s and early 1990s, hundreds, perhaps thousands, of cases were recorded in North America of psychotherapy patients (often diagnosed with 'multiple-personality disorder'[18]) who claimed to have, in therapy, 'recovered' previously 'repressed' memories of being subject to satanistic and/or sexual abuse when children. Many of these cases led to court proceedings against the alleged perpetrators (often the patients' parents); in some 29 US states new legislation was enacted delaying statutes of limitation, making it possible to sue for civil damages within a given number of years of the 'discovery' of the 'repressed memories', no matter how long ago the alleged abuse was supposed to have taken place. Many of these 'recovered' memories were subsequently found to be false, leading to hundreds more lawsuits, this

[18] See § 8.3.3 for a discussion of this disorder.

time against the therapists responsible for encouraging the false memories, and the formation of a pressure group of families affected by such cases, the False Memory Syndrome Foundation (Hacking, 1995a, ch. 8; Loftus, 1996).

Now, it might again be argued that, tragic though such false-memory cases were for those affected, they were the result of some underlying psychopathology and misguided therapeutic methods, and that no adverse conclusions should be drawn from pathological cases about the reliability of episodic recollection in 'healthy' people. This is a dangerous argument. Anyone could, through no fault of his or her own, succumb to some psychological disorder that encourages the formation of false memories. And Loftus's studies show that it is possible to create false memories in healthy subjects, too. Although these were created in controlled laboratory conditions, it is at least conceivable that similar circumstances occur in our ordinary lives. A conversation in which someone recounts experiencing a particular instance of abuse in childhood could, given appropriate motivating factors (envy, sympathy…), convince one of having experienced that kind of abuse in one's childhood, too, even if one did not. Even if most of us do not have completely false memories, it would be implausible to assert that anyone is immune from them.

4.4.6 Reconstruction and reliability: summary

The reconstructive nature of episodic recollection and its close kinship with imagination call into question the reliability of such recollections. As just seen, there is a considerable spectrum of diminishing veridicality of autobiographical memories, ranging from the wholly accurate to the wholly false. Some semantic memories, as well as recent and particularly salient episodic memories, may be entirely accurate. Most episodic memories are likely to become embellished in recollection with the filling-in of details from other sources in one's autobiographical knowledge base. Some episodic recollections, by the same process, will be misremembered ones, but open to being revised when the rememberer is confronted with contradictory evidence.

More serious, but admittedly pathological, cases of false remembering include the confabulation arising from frontal-lobe damage, and false memories allegedly 'recovered' in psychotherapy. However, as Loftus's

studies show, a significant minority of subjects are susceptible to the creation of false childhood memories even in the absence of an underlying pathological condition.

We cannot, therefore, take autobiographical memories for granted as fixed and unchanging or reliable reproductions of past events. Together with the constraints on the availability of autobiographical memory discussed earlier, this should worry Lockeans. And reconstructive remembering suggests that the order of priority of self and autobiographical memory in our cognitive architecture is really quite the reverse of what historical-constructionist accounts of the self assume. These issues are the subject of the final section of this chapter.

4.5 Autobiographical memory and the self

Having discussed the nature and constraints of autobiographical memory, I must now turn to its relation with the self. There is a large psychological literature exploring various empirical approaches to this relation (for a wide-ranging sample, see Neisser and Fivush, 1994). But my aim here is to adjudicate between different conceptualizations of the self. In this regard, there seem to be two opposite ways in which one might view the self–memory relation, depending on which of the relata is given logical and empirical priority over the other. Is *memory* prior, and is the self a *construct* from our autobiographical memories, as the Lockean-inspired historical-constructionist tradition would have it? Or is the *self* prior, in that the construction of autobiographical memory is an activity of the self? I will now examine which of these approaches better fits the evidence from psychology that I have presented here.

4.5.1 The self as a construct from autobiographical memories?

A common reading of Locke's account of the self is that 'the self is a logical construction, and is to be defined in terms of memory' (Grice, 1941, p. 340). I have already discussed the well-rehearsed philosophical objections to Locke's account (§ 2.3). Now, it seems, the Lockean view and its descendants face serious *empirical* challenges as well. I noted earlier (§ 4.3.2) how

variations in awareness of our autobiographical memories and the fact that episodic memory has a 'feeling of pastness' about it make it impossible for there ever to be the 'same consciousness' of the past as of the present that Locke's own account requires. So let us discard the 'same consciousness' criterion and follow Grice in taking *memory*—whatever the type and degree of awareness we may have of it—as the crucial ingredient of the self. But, given the nature of autobiographical memory, can it serve as 'raw material' (Hirst, 1994) for the construction of a self?

Raw materials for construction—if they are suitable, and if the construction is to be stable—are not expected to change much. This is not true of autobiographical memory. As discussed above (§ 4.3.1), our 'autobiographical knowledge base' (Conway & Rubin, 1993) is subject to constant updating and modification. Recent episodic memories are discarded or transformed into components of 'general event' memories. More distant, half-forgotten memories may suddenly be retrieved with new urgency when an appropriate relevant cue is presented. Anything constructed out of such mutable material is likely to shape-shift accordingly. But perhaps the self is like that, itself being constantly reshaped and updated? Granted, even my system view (§ 1.4) supposes that the self is dynamic rather than static. But, if the self is *constructed* from autobiographical memories *alone*, the selectivity with which we encode and retrieve these makes for a very 'gappy' self—a construction not only constantly shape-shifting, but shot through with rather large holes, which over time become larger still as 'generic forgetting' takes its toll. Over a lifetime, the gaps in one's memory probably far outnumber available memories, and not all available memories are always accessible: we end up with a very flimsy and unstable construction.

Defenders of an historical-constructionist view of the self may be happy to bite that bullet and accept that the self, constructed from memories, is both gappy and unstable. But there is an element in the uneven distribution of autobiographical memories over a lifetime (§ 4.3.3) that is particularly troubling here: childhood amnesia. The simple consequence of childhood amnesia is that a self constructed from autobiographical memories begins no sooner than about the age of three. Whatever happens in the first few years of our lives—and developmental psychology suggests an awful lot happens during that time, as various cognitive functions come

online, are trained, honed, and exercised—it cannot, on a view of the self as based on available memories, form part of our selves. The foundational cognitive developments of early childhood are simply excluded here. This is a deeply troubling implication of the view that the self is constructed from autobiographical memory: for it now seems that this construction—already shape-shifting and riddled with holes—also lacks a foundation.

But the most critical argument against the historical-constructionist position comes from the realization that autobiographical memory is itself a construction. As discussed in the previous section, our episodic memories are reconstructed rather than fixed. And the organization of our 'autobiographical knowledge base' (§ 4.3.1), containing our semantic self-knowledge as well as the elements for episodic recollection, is a constructive process, too. Constructionism about the self here seems to face another circularity objection. The construction of autobiographical memory involves, in Stanley Klein's notation (Klein et al., 2004; Klein, 2013b—see §§ 4.2.3; 4.4.1), the capacity to self-reflect, a sense of ownership and agency, and a sense of temporality. These capacities sound very much like elements of a self. Thus, it seems that a self is necessary for the construction of autobiographical memory—but autobiographical memory is supposedly what the self is constructed from.

A constructionist about the self could try to weasel out of this circularity by holding that neither self nor memory is logically or empirically prior to the other, but that they are somehow co-constructed. Klein's three capacities may *seem* to be elements of a self, the reply goes, but they only appear as such when they are exercised in constructing autobiographical memory, which in turn constructs a self. I do not find this reply very convincing. Does the self thus construed now encompass the very tools exercised in its construction? This would imply placing *capacities* (such as are involved in constructing autobiographical memory) on a par with *content* (the self supposedly constructed *from* autobiographical memory). But the capacities necessary for constructing memory *are* prior to this construction. And the position that self and memory are somehow interwoven constructions has nothing to say about the capacities required for constructing either. It is thus far more sensible to situate the self, conceptually, at the level of these capacities, as I will demonstrate shortly (§ 4.5.2).

There is, however, one more problem to mention that affects the view that the self is, in Grice's phrase, 'defined in terms of memory'. A self constructed from memory *alone* is vulnerable to the concerns about the reliability of our memories highlighted in the previous section. We have encountered a spectrum ranging from embellished and misremembered recall to confabulation and false memories. In cases where someone actually has false memories, we would have to say that such a person's self is defined partly in terms of false memories—which would be psychologically and philosophically problematic. Psychologically, because of the cognitive dissonance it creates between what one 'remembers' and what really happened. Philosophically, because it makes a mockery of any notion of self-*knowledge* and is thus, as Ian Hacking (1995a) puts it, 'contrary to our best vision of what it is to be a human being' (p. 267).[19]

But even in cases where no false memories are present, a memory-based self is at risk from their mere possibility. For we can rarely be sure which of our episodic memories might not be false, at least in part. (Thus Hacking's 'best vision' may turn out to be an unrealistic one.) If we construct a self wholly from the memories we seem to have of our past actions, which may be embellished, misremembered, confabulated, or false, the resulting self is uninsured against being contradicted by the facts.

4.5.2 The self as constructor of autobiographical memory

If we reject the view that the self is a construct from autobiographical memory, does this entail dismissing the apparent popular intuition we encountered in Chapter 3—that the loss of one's memories is indicative of a discontinuity of self? Not necessarily. For such an intuition is consistent with a view that preserved memory is *evidence* of a self, rather than constitutive of it. Hume (1739/1978) makes this point in response to Locke:

---

[19] Personhood is not my topic, but I should add that the possibility of false memories seems even more troubling for Locke's concept of a *person* (in its 'forensic' sense that emphasizes responsibility for one's past actions—see § 2.4). For if one falsely remembered a past action one did not in fact commit, then—unless the Lockean account is augmented by a requirement for external evidence that such 'remembered' actions are truly one's own—that action would be appropriated into one's Lockean person. But then the notion of a person as the totality of the actions for which one can be commended or blamed becomes absurd.

'memory does not so much *produce* as *discover* personal identity' (I.iv.6, p. 262). Then the loss of one's memories is a symptom, not the cause, of a disrupted or malfunctioning self. Let us then turn the memory–self relation the right way up: the construction of autobiographical memory is an activity of the self, a function of our self-representational system.

On this view, the operations of autobiographical memory discussed in this chapter make a lot of sense. *Autonoetic consciousness*, Tulving's marker of episodic recollection (§ 4.2.2), now appears as an instance of self-monitoring: the ability to process past experiences with reference to oneself and from an egocentric perspective, and correspondingly to simulate oneself in future activities. Suppose I have a rich autonoetic, first-personal recollection of delivering a mediocre (or successful) lecture. If I then catch myself, in my next lecture, proceeding in a manner that my recollection has marked as mediocre (or successful), I can adjust my style accordingly (or merrily carry on in that vein). A mere factual memory, detached from personal involvement in a specific experience, of what is and what isn't good lecturing, would be much less effective than the autonoetic awareness of *myself* having done well or badly in the past.

The fact that dissociable processes contribute to episodic remembering (§ 4.2.3) is easily accommodated by the system view, since it takes the self to be a complex system of self-representational capacities that can fail or malfunction individually. In the case of R.B. just one such capacity—the sense of ownership of episodic recollections—is faulty, while others—semantic self-knowledge, a sense of temporality—are unaffected.

The *selectivity* of encoding and retrieval operations (§ 4.3.1) also makes sense if we consider them in the context of what is relevant to one's self-interest both at the time of encoding and of retrieval. The friendly manner of the coffee-shop staff is worth encoding in memory as it provides one with a reason (along with the quality of the coffee) to return there—the precise words exchanged, on the other hand, are irrelevant and do not get recalled. In remembering a dinner party *qua* social occasion, being able to recall who was present is a useful self-representational exercise of situating oneself in one's social context. Recalling the colour of the table-cloth is not: that information is discarded from memory. Another way in which selective encoding and retrieval is adaptive is in 'storing the exception' to some pre-

existing self-image (Wagenaar, 1994): an event that represents a prediction error is *ipso facto* more salient and more likely to be encoded in memory.

The view that the self *qua* self-monitoring system constructs autobiographical memory may also help explain the two striking features of the *distribution* of autobiographical memories across a lifetime (§ 4.3.3): childhood amnesia and the 'reminiscence bump' of memories from adolescence and early adulthood. Concerning childhood amnesia, it is plausible to assume that our self-representational capacities are not all online at birth, but that the higher-level capacities required for autobiographical memory gradually develop over the first three years of a human life—hence the lack of autobiographical memories from those early years. As for the reminiscence bump: if the formation of autobiographical memories is a result of higher-level self-monitoring processes, it should not come as a surprise if that system is particularly active during adolescence, accordingly encoding more (or more durable) memories at that time than in later life, when self-monitoring is a more routine business as we encounter fewer novel situations (but see § 5.2.3 for other explanations of the bump).

As to the *reconstructive* nature of episodic recollections (§ 4.4): the system view has no difficulties here. Understood as our system of self-representational capacities, the self is quite obviously in the business of producing self-representations—and episodic memories are self-representations *par excellence*. Moreover, the manner in which we reconstruct autobiographical memories is, quite literally, self-serving. Memories are rearranged, shaped, and twisted by the self—by, as Anthony Greenwald (1980) dramatically calls it, 'the totalitarian ego'. Thus, Michael Ross (1989) found that our 'implicit theories' about ourselves determine the construction of autobiographical memory. *Pace* Wagenaar's above-cited 'storing the exception' model, Ross's experimental participants with a stable self-image tended, in their recollections, to overestimate the consistency of their attitudes. Conversely, the recollections of participants with a self-image emphasizing personal development were prone to overstate the changes in their attitudes over time. Such cognitive biases in remembering underline the priority of self-representational processes over autobiographical memory.

And it is here that differences between individuals are most apparent. For we differ not only in the experiences we have: what and how and how much we remember autobiographically reflects individual proclivities—as

Bartlett (1932) put it, the rememberer's 'appetite, instinct, interests, and ideals' (p. 210). Greg Neimeyer and April Metzler (1994) studied the differences in autobiographical recall between three types of personalities: 'information-oriented individuals', who actively explore different attitudes before committing to one, 'normative-oriented individuals', who are firmly committed to their attitudes without exploration, and 'diffuse-oriented individuals', who show neither commitment to nor active exploration of attitudes.

Participants were asked to recall personal memories relevant to cues of personality traits that were either positive or negative and either validated or invalidated participants' (pre-trial) ratings of self-descriptiveness of those traits. The number of memories recalled and the latency of recall were measured, as was an index of 'perceived self-change' based on another (post-trial) round of ratings of self-descriptiveness of the personality traits. The results showed differences in all these variables between the three different personality types. Information-oriented subjects had the most positive validating and the most negative invalidating memories (here more than double the number of memories recalled by normative-oriented individuals). Diffuse-oriented subjects had the fewest recollections overall, but showed the most marked change in their self-perception after recalling invalidating memories, while normative-oriented subjects showed the least change in their self-perceptions in those cases. Normative-oriented subjects were also 'the most inclined . . . to generate invalidating memories when it benefited them and the least inclined when it threatened them' (p. 130).

Neimeyer and Metzler's study illustrates a number of features of autobiographical memory: the reconstructive nature of its recall and how a rememberer's self-image can govern this reconstruction, but also how individuals differ in their cognitive biases and flexibility when engaging in autobiographical recollection. A variety of self-representational processes seems at work here—retrieval and reconstruction of autobiographical memories, and their appropriate enabling sub-processes, but also the maintenance or reappraisal of one's self-image—and not all in equal measure in different individuals. The self is a very busy system. Autobiographical remembering is an important activity of that system—but far from the only one.

4.5.3 Conclusion

Autobiographical memory is complicated. It involves several different aspects of memory, supported by even more dissociable self-representational processes. Its organization and the recall of episodic memories are constructive or reconstructive processes. Thus, it is difficult to maintain the position that the self is a construct from autobiographical memories.

The system view provides an account of the self much better suited to the nature and operations of autobiographical memory. Remembering, on this view, is an activity of the self, recruiting diverse self-representational functions both when memory is encoded and, crucially, at the time of recollection. The system view also avoids limiting the self to being 'defined in terms of memory', which would leave it vulnerable to the constraints of availability and accessibility and reliability that I have elucidated in this chapter.

Much more could be said about autobiographical memory. But I must now address a related matter that also pervades the philosophical and psychological literature on the self: *narrativity*. Looking again at how some autobiographical memories are reinforced and consolidated, it will be observed that the notion of narrative creeps up in a number of places. Hardcastle's intensity hypothesis of memory (fig. 4.1) makes use of the retelling of events to reinforce episodic memories. Some explanations of the reminiscence bump likewise appeal to narrative practices, as does Nelson's account of childhood amnesia, which suggests that the beginning of autobiographical memory in childhood may co-occur with the onset of narrative memory talk between child and parent.

The role of narrative practices may not be limited to being an aide to the consolidation of autobiographical memories: a number of scholars suggest that it is by means of narratives that we construct our selves. It may seem that narrativity offers a way out for historical-constructionist accounts of the self, if it can remedy the gappy and shifting nature of autobiographical memory by providing additional support to its structure, and perhaps a narrativist view can mitigate concerns about reliability. I will discuss narrative practices and critically examine the case for a narrative self in the next two chapters.

Chapter Five

Narrative practices, narrative selves?

5.1 Introduction

As hinted in the previous chapter, there are interactions between memory and narrative practices concerning the onset, organization, consolidation, and retrieval of autobiographical memories. But narrativity is frequently taken to be more directly connected with the self. Narrative practices, it is claimed, enable the emergence of a self-concept in child development (Fivush, 1994; Miller, 1994; Nelson, 2003) and allow us to structure and interpret our autobiographical memories in a meaningful way (Barclay, 1994; Freeman, 2003). Narratives provide the vehicle by which we give reasons for and make sense of our actions (Velleman, 2005), or indeed produce or 'constitute' the self (Dennett, 1992; Bruner, 1987; Bruner & Kalmar, 1998; Schechtman, 1996; 2007).

It is no easy undertaking to define the scope of, and distinguish between, these various claims, since there is considerable overlap between them. A philosophically important distinction to be made, which I take from Schechtman (2011b), is that between, on the one hand, accounts that link the self with various narrative capacities and practices in various ways, and, on the other hand, accounts that take the self to be constructed and structured narratively. In this context we must also ask quite what is meant by 'narrative'. I will discuss this question, distinguish between different views of the relation between narrative and the self, and illustrate these in the second half of this chapter (§ 5.3). But before doing so, I will consider

evidence from empirical psychology for the role of narrative practices in constructing autobiographical memory and how this might affect our conception of the self (§ 5.2).

5.2 Narrative, memory, and self: the view from psychology

5.2.1 Narrative practices and incipient memory construction in early childhood

In discussing the distribution of autobiographical memories across the lifespan and its absence in early childhood (§ 4.3.3), I mentioned that the onset of autobiographical memory in child development has been linked with narrative practices. I will now review relevant research in developmental psychology and the conclusions that researchers have drawn from it.

Robyn Fivush (1994) reports on studies of parent–child conversations about the past. One such study observed parents and their 2½–3½-year-old children (all from a white middle-class American population) engaging in conversations about events they had experienced. These were initiated and steered by the parent. The children's responses Fivush reproduces are most often repetitions of key words from parents' questions, and brief affirmative ('Yeah', 'Uh huh'), negative, or querying ('What?') rejoinders to these, but children are also reported as providing their own brief descriptions of what happened during the event under discussion. This occurred when parents were using an 'elaborative' conversational style, in which they themselves provided descriptive information about the event and kept adding to this in the course of the conversation, 'thus providing distinctive memory cues' (p. 143) which would lead to the child's volunteering his or her own descriptive memories. A 'repetitive' conversational style, in which parents asked brief questions and then repeated these without elaborating, moving to a different topic if the child provided no information, was not successful in this regard. Fivush concludes that 'elaborative parents are

teaching their children that the past is interesting and important to talk about and share with others' (p. 144).[1]

It seems plausible that such narrative practices encourage incipient autobiographical remembering in childhood, and the children's ages in this study correlate nicely with the usual onset of autobiographical memory at around 3 years of age. We should not, however, jump to early conclusions about the direction of causality between narrative practice and incipient autobiographical memory, and about the role of narrative practices in 'the continuously developing sense of self' (p. 136). I will return to these points below (§ 5.2.4). An initially more pressing question, given Fivush's entirely 'white middle-class' experimental population, is whether the narrative practices studied are culturally universal.

Peggy Miller's (1994) work suggests that practices of 'personal storytelling' do occur in diverse socio-cultural communities. With respect to child development, she classifies these narrative practices as taking place (1) 'around the child' in adult-to-adult conversations about the personal experiences of the narrating adult in the presence of children, neither 'told deliberately *for* children, nor . . . censored on behalf of children' (p. 167); (2) 'about the child' in adult-to-adult conversations with the child as the protagonist of the event being narrated (though children begin to contribute verbally to these narrative acts by 2½ years of age); and (3) 'with the child', where children act as 'co-narrators of their own experiences' (p. 171) from about 2½ years of age. Both 'around' and 'about the child' narration is

---

[1] The study also showed an unexpected gender differentiation: 'Both mothers and fathers were significantly more likely to use an elaborative style when talking about the past with daughters than with sons' (Fivush, 1994, p. 143). Fivush conducted a follow-up study of the use of emotion words in mother–child conversations about the past (15 boys and 15 girls aged 2½–3 years), which gave further support to this. While there was no significant difference between genders in the *quantity* of emotion words used overall, there were significant differences in emphasis, regarding both the kinds of emotions discussed (more discussion of sadness with girls, more of anger with boys), the types of resolutions to emotional episodes discussed (with regard to anger, these tended to be conciliatory with girls, retaliatory with boys), and the way in which the context of the emotional experience was discussed (a 'social-relational framework' with girls, an 'autonomous [individualistic] framework' with boys). Thus, Fivush argues, girls' 'remembered selves will be rich, detailed, and interpersonally oriented', whereas 'boys' remembered selves will be spare, limited in emotional tone, and autonomously oriented' (p. 154).

reported as occurring across socio-cultural groups (including both American and Taiwanese Chinese groups; among the former, both low- and middle-income groups, each in both white and African-American communities).

Unlike those in Fivush's studies, the narrative practices observed by Miller are not explicitly seen as an aid to building and recalling autobiographical memories:

> Events are not remembered for the sake of remembering but for the sake of creating tellable stories about the self. In South Baltimore, Daly Park, and West Side [low-income neighbourhoods; the latter two in Chicago, respectively white and African-American], where traditions of highly performed oral narrative flourish, speakers do not seem to define personal storytelling as a memory task. They infrequently use *remember* or related terms when telling or introducing stories and they rarely revise their accounts in the interest of accuracy. To do so would destroy the integrity of the story as an artistic performance. (p. 175)

Rather, the role of narrative here seems to be to illuminate personality traits, to organize one's experiences (one's 'autobiographical knowledge base'—see § 4.3.1), and to explore and define self–other relations. Memory reconstruction here is 'socially distributed' (ibid.) in vicarious story-telling, in placing oneself in relation to others, and in the narrative act itself, which is in part determined by its audience (notice Miller's reference to an 'artistic performance').

Miller contends that children included and participating in these narrative practices 'experience and reexperience self in relation to other' (p. 173). This suggests that narrative memory construction in childhood shapes one's experience and, more generally, one's self-consciousness. That theme is developed by Katherine Nelson (2003). According to Nelson, consciousness is an 'ongoing self-organizing developing system' with different 'emergent levels' (p. 18) of representation. In childhood, this system develops from early social awareness and intersubjectivity (as attested in joint attention and protodeclarative pointing) in the first year of life through rudimentary self-awareness (attested by self-recognition in mirrors) by the end of the second year to a more nuanced self-and-other-awareness that comes with narrative discourse and, ultimately, the emergence of a 'new level of consciousness . . . dependent upon language used

to exchange views of self and other, primarily through narratives but also through commentary on the self by others' (p. 33).

How, then, do narrative practices affect the structure of consciousness in child development? Borrowing terms from Bruner (1990), Nelson (2003) refers to two ways of conceptualizing experience that appear to come from narrative practices. The first of these is a '*landscape of action*', which we are to understand as 'the sequencing of actions to make a coherent and cohesive event' (p. 25). The other is a '*landscape of consciousness*', to be understood as 'the revealing of the mental states of the actors that are associated with the action, including their goals, their perspectives, their beliefs, their emotions, and so on' (ibid.). Both the understanding and the evaluation of agents' motivations—or, as Dan Hutto (2007b) puts it, 'reason-giving' (see § 6.4); or again our 'theory of mind' (see § 7.2.2)—belong to this second 'landscape'.

Crucially for Nelson's argument, a grasp of the 'landscape of consciousness' occurs rather later in child development than that of the 'landscape of action'. Nelson (2003) cites an example from her own research into infants' 'crib talk': a 2½-year-old child soliloquizes about her father's not being able to run in the New York marathon and having to watch it on television instead, without however grasping the reason for this. And in the parent–child conversations studied by Fivush and in the social narratives researched by Miller, while young children provide some of the narrative elements of *action*, the *evaluative* elements are provided by the adults. Overall, Nelson observes that in the narratives of 3- to 5-year-olds, 'temporal perspective' (an obvious crucial capacity for self-monitoring over time) as well as 'the mental . . . perspective of self and of different others' is still 'weak or nonexistent' (p. 28).

Unfortunately, indeed somewhat maddeningly, this is where Nelson's developmental story ends. Some time after the age of five, we are led to assume, the 'landscape of consciousness' begins to unfold in children's narratives, and with it emerges that 'new level of consciousness' which includes self-consciousness—but Nelson does not tell us when and how.

Commenting on research into children's narratives, Rebecca Eder (1994) observes that there are 'substantial developmental changes in children's self-concept, especially between 3 and 8 years of age and during adolescence' (p. 185). The process of developing a sense of self in childhood is

thus rather more drawn out than the developmental window (ages 2–5) studied by Fivush, Miller, and Nelson. Furthermore, the evidence from developmental psychology, while suggesting that narrative practices play a *role* in developing a sense of self, is inconclusive on whether their role is a necessary one. As Eder notes, narrative practices fulfil functions other than 'self-*construction*'. These include self-*presentation*, and in such cases the narrative is 'the *result* of . . . rather than a mechanism for self-construction' (ibid.).

Let me turn then to narrative practices in adults, and in this context return to the question of how they interrelate with autobiographical recollection.

5.2.2 Narrative capacities and autobiographical memory in adulthood

As discussed at length in the preceding chapter, there are numerous constraints on the encoding and retrieval of autobiographical memories, which include the *significance* of events. Here, then, is an obvious role for narrative practices: if an action or experience is significant in the context of some narrative structure—it may fit nicely with a particular story we like to tell ourselves and others about ourselves—such an event may be more likely to be consigned to memory, and, assuming the continuing relevance of its narrative context, to be recalled. Given the reconstructive nature of episodic recall, perhaps narrative practices shape that reconstruction. Further, the narrative *rehearsal* of a particular episode—actively telling oneself and/or others about it—is likely to consolidate one's memory of it. Finally, Conway and Rubin's (1993) layered structure of our autobiographical knowledge base (lifetime periods, general events, specific events) suggests that the way we organize our autobiographical memories may likewise benefit from narrative practices: lifetime periods and 'general events' are structured around themes, and these may well be the themes of an autobiographical narrative.

But, again, we should exercise caution about the direction of causality between narrative practices and memory organization. It is just as plausible that the thematic structure of one's autobiographical knowledge base shapes the stories one tells about oneself as it is that such self narratives shape the organization of autobiographical memories. On the other hand, where the retelling of experiences aids the encoding, consolidation, or re-

trieval of relevant autobiographical memories, there does seem to be a causal influence of such narrative practices on autobiographical memory. But here, too, we should pause before assigning a *necessary* role to narrative.

If narrative capacities—the ability to narrate one's memories, and/or the ability to reason narratively about one's experiences—were to be necessary for autobiographical memory, we should expect impairments of these capacities to correlate with memory loss. In a thoroughgoing review of the evidence from cognitive psychology and neuropsychology on autobiographical recollection and its disorders, David Rubin and Daniel Greenberg (2003) consider a number of systems that may be needed for autobiographical recollection, two of which would seem to underpin narrative capacities: the language system (here understood as the system for phonetics, syntax and semantics at and below the level of sentences), and what they call a 'narrative reasoning' system, where 'narrative reasoning' is understood as 'the ability to use structure above the level of the sentence', which is 'used to describe particular incidents of goal-directed behavior' (p. 62).

The evidence from cases of language loss (aphasia) and loss of narrative reasoning cited by Rubin and Greenberg suggests that (*a*) language and narrative reasoning are indeed separate systems: the loss of one does not necessarily entail the loss of the other; and (*b*) neither language loss nor loss of narrative reasoning necessarily leads to loss of autobiographical memory.

Only one particularly severe kind of aphasia, known as semantic dementia or progressive fluent aphasia, 'often results in severe impairment of autobiographical memory' (p. 67), especially for early memories. Other forms of aphasia, however, do not. In most cases of fluent aphasia (struggle to retrieve words) and conduction aphasia (deficit in 'inner speech'), 'aphasics manage to produce remarkably well formed autobiographical memories' (ibid.). Nor do these types of aphasia lead to a loss of narrative reasoning: aphasics' narratives 'reflect their language impairments' but 'frequently preserve discourse structure' (p. 70). So a difficulty with producing language is dissociable from deficits of narrative reasoning. Narrative reasoning is also spared in a variety of other disorders including Alzheimer's disease and Korsakoff's syndrome, which involve memory loss; thus, narrative reasoning ability is also dissociable from memory.

What of a loss of narrative reasoning itself? Rubin and Greenberg observe that narrative reasoning is surprisingly difficult to disrupt and its loss difficult to test for. However, assuming a neural basis of narrative reasoning in the right hemisphere and/or the frontal lobes, some conclusions can be drawn from patients with neural damage in these areas. Right-hemisphere patients present obvious symptoms of a loss of narrative reasoning: they have 'difficulty comprehending speech at the discourse level', 'fail to judge the intent of nonliteral statements', 'have substantial difficulty interpreting metaphors . . . and drawing inferences from stories', show 'substantial impairment of their ability to organize and reorganize data' and 'significant difficulty reproducing narratives' (p. 72). However, no deficits in autobiographical memory are reported in right-hemisphere patients. If narrative reasoning were necessary for autobiographical memory, there should be such deficits.

For frontal-lobe patients, meanwhile, there *is* a correlation between the loss of some components of narrative reasoning and impairments of autobiographical memory. As noted in the previous chapter, frontal-lobe damage leads to confabulation (§ 4.4.4), which is marked by noticeable deficits in maintaining plausible temporal and causal sequences in personal memories and narratives. But it would be questionable and unwarranted to suggest that the frontal-lobe damage first produces a loss of the requisite narrative reasoning ability, which then, *in turn*, produces the faulty memories. Rather, it seems that the executive dysfunction which is the result of frontal-lobe injury is responsible, by disrupting temporal and causal sequencing, for the impairment of *both* narrative reasoning and autobiographical recollection. That is to say that narrative reasoning and autobiographical memory here share a common substrate in the shape of executive function, not that one of them is a substrate of the other.

Overall, then, there are interesting interrelations between narrative capacities (as instantiated in what Rubin and Greenberg call the language system and the narrative reasoning system) and autobiographical memory, but they do not support the hypothesis that narrative capacities are *necessary*

for autobiographical recollection.[2] At best, we can say that *some* impairments of narrative capacities correlate with certain impairments of autobiographical recollection.

5.2.3 The 'reminiscence bump' revisited

Let us look, then, at a particular aspect of the structure of autobiographical memory that might be subject to shaping by narrative practices and that is particularly relevant for one's sense of self over time. As I said earlier, narrative practices seem a possible influence not only on the recollection of individual episodes, but also on the way we organize and structure our autobiographical memories across the life-span (e.g. into Conway & Rubin's (1993) 'general events' and 'lifetime periods'). Of particular interest here is the 'reminiscence bump' discussed in the previous chapter (§ 4.3.3)—the high concentration of autobiographical memories from the period between late childhood and early adulthood.

In a study with over 1,200 Danish participants between the ages of 20 and 93, Dorthe Berntsen and David Rubin (2002) set out to examine whether the distribution of memories across the life-span was affected by the *emotional* content of the memories. Participants were asked to state how old they were in their happiest, saddest, most important, and most traumatic memory. Further, they were asked about their most recent involuntary (unbidden) recollection: how old they were in the event they had recalled involuntarily, and whether this was a happy or a sad event.

The distribution of responses obtained by Berntsen and Rubin shows a clear bump in the twenties for the reported age in participants' happiest and most important memories across all age groups, but not for the saddest and most traumatic memories. For involuntary memories, most of these took place in the recent past for all age groups; there is, however, a smaller peak in the distribution for memories from participants' teenage years, but again only for happy memories. Berntsen and Rubin note:

---

[2] This in contrast to the evidence for a necessary role of what Rubin and Greenberg call the 'imagery system' in autobiographical memory: long-term visual memory loss does correlate with 'a general loss in autobiographical memory that extend[s] beyond visual memory to all areas of memory' (p. 66).

> The data reflect some bias exerted by memory and not simply how events are distributed over life. For example, it is unlikely that around 40 % of the happiest events in people's life actually take place between age 20 and 30, as would be the case if the memory data simply reflected a real-life distribution. (p. 647)

One might quibble here that, since the study asked participants to recall only *the single most prominent* event for each emotional category (*the* happiest, *the* saddest, etc.), the distribution of responses does not in fact say anything about the distribution of memories overall: barring the single happiest event participants were asked to recall, there could be an otherwise even distribution of happy memories across a life-span. But even so, it is still noteworthy that for 40 % of participants the single happiest memory of their lifetime should be one from their twenties. Let us accept, then, that there is something in the organization of autobiographical memory that accounts for this 'happiness bias' in early adulthood. But what?

Berntsen and Rubin note that various explanations of the reminiscence bump in general fail to predict the specific results of their study. A '*biological/maturational*' account of the bump—suggesting (as I do in § 4.5.2) that 'early adulthood might be especially favoured by nature with respect to cognitive skills' including memory—'does not specify any differences related to the emotional charging of the memory material or to involuntary versus voluntary retrieval' (p. 640). Similarly, *cognitive* accounts attributing the reminiscence bump to the relative novelty and distinctiveness of events experienced between the ages of about 15 and 25 do not predict a dissociation between happy/important and sad/traumatic memories. Cognitive explanations can be modified to account for Berntsen and Rubin's results: happy memories are more likely to be rehearsed than sad ones, possibly as a result of social censure; and there is some evidence that 'ratings of emotional intensity decrease more rapidly for negative than for positive events', suggesting that sad and traumatic memories are more easily forgotten, but then 'it is not clear what causes memories of negative events to fade more quickly than positive memories', which makes this suggestion 'a rather weak explanation of the present results' (p. 684).

As for *narrative* accounts of the bump, in which early adulthood is seen as a particularly identity-forming chapter in the 'life story': these again predict no difference between positively and negatively valenced emotional

memories from that period. Again, modifications to such accounts are required to accommodate Berntsen and Rubin's findings, such as—again—social censure precluding the narration of sad and traumatic events, or nostalgia for one's adolescence. 'Nostalgia,' however, Berntsen and Rubin note, 'is a description rather than an explanation' (p. 649).

The only account of the reminiscence bump which, according to Berntsen and Rubin, correctly predicts their findings without *ad hoc* modifications is what they call '*life scripts*', that is, 'normative expectations within a given culture [concerning] the patterns of individual life courses, such as the developmental changes that are expected to take place at various points in life and the different life phases that people are expected to live through at different ages' (p. 640). 'Life scripts' are to be distinguished from narratives: the former are 'generic', 'nonpersonal', involve 'cultural expectations', and are 'public knowledge'; the latter are 'concrete', 'personal', involve autobiographical memories, and are 'private knowledge that is shared with very few people' (ibid.). According to the 'life scripts' account,

> To the extent normative life scripts organize autobiographical memory, memories of the happiest and most important events should form a bump, whereas the distribution of memories for the saddest and most traumatic events should be relatively flat, because they are unlikely to be part of the life script. This pattern should be found for both voluntary and involuntary memories. (p. 638)

While this account does predict Berntsen and Rubin's results, and would seem also to underlie the 'social censure' modification of the cognitive and narrative accounts, it is not quite clear why socially mandated 'life scripts' would exclude sad and traumatic events. Indeed, one might think that unavoidable unhappy occurrences such as the deaths of close friends and family members ought to form part of a cultural life script precisely because they will happen to most people at some point(s) in their lives. Thus, a better way of framing the 'life scripts' account, which is still consonant with the flat distribution of sad and traumatic memories observed, is that 'no specific time slot [in life] is allocated to such events' (p. 640)—whereas the happy and important events in a stereotyped life script (such as graduation, first job, marriage, parenthood) do have a relatively fixed time slot, which concurs with the bump in the memory distribution.

There remain questions about the 'life script' account; for instance, what cross-cultural variation there might be in such normative expectations, and whether these are reflected in people's autobiographical memory organization. But the point is that Berntsen and Rubin's 'life script' account offers a better explanation of the 'happy memories' bump in their data than mere appeal to narrative practices. A narrative account *without* 'life scripts' as an auxiliary hypothesis cannot explain the happiness bias of the reminiscence bump.

5.2.4 The self as a product of narrative?

From the psychological evidence considered here, there emerges a patchy and diffuse picture of the role of narrative in the development and/or maintenance of a sense of self, either by shoring up autobiographical memory or more directly. Narrative practices may play a role in the developing sense of self in childhood, but it is far from clear whether only they can or do perform this role, and whether their increased sophistication in older children perhaps reflects other developmental changes. There are various interactions between narrative practices and autobiographical recollection, but none support the hypothesis that either language or narrative reasoning (unlike sensory imagery) is necessary for autobiographical memory. And a narrative account is neither the only nor the best explanation for the reminiscence bump in the temporal distribution of autobiographical memories (and thus of the temporal structuring of one's sense of self over time).

Though the *empirical* evidence in psychological studies of narrative practices is thus suggestive only of some involvement of narrative in our sense of self, some psychologists go further than this on *theoretical* grounds. Foremost and first among these is Jerome Bruner, who argues that the self is indeed a product of narrative (Bruner 1987; 1994; Bruner & Kalmar, 1998). Bruner's grounding assumptions are, briefly, these. First, stories or narratives embody a 'mode of thought' (1987, p. 11) universal among human beings that predates, in cultural evolutionary terms, what he calls 'logical or inductive' thought.[3] The evolutionary idea here is, put simply, that once

[3] This is taken up by Nelson (2003) who argues that in child development, too, narrative capacities precede 'logic and scientific theorizing' (p. 24).

human beings had developed language, they first began making sense of the world by telling stories about it. These cultural practices are then said (by followers of Vygotsky) to 'mediate thought' (Bruner, 1991, p. 3). Secondly, Bruner notes, 'We seem to have no other way of describing "lived time" save in the form of a narrative' (1987, p. 12).[4]

Taking these two assumptions together, we would seem to arrive at making sense of our lives through narratives, and since such narratives are also a 'mode of thought', they in turn 'structure perceptual experience' (p. 15). But, more than that, they supposedly produce a narratively constructed *self*. How so? Miller (1994) refers to 'self-construction' in her account of narrative practices among her subjects in South Baltimore (§ 5.2.1), but, as discussed in that context, these practices seem more a case of self-*presentation* than self-*construction*. Bruner (1994), meanwhile, at first describes the self as 'a complex mental edifice that one constructs by the use of a variety of mental processes' including remembering (p. 41), but soon emphasizes the principal role of narrative in this constructive process and concludes that 'Self is a perpetually rewritten story' (p. 53). According to this view, we are asked to take the self to be *(a)* a *construction* and *(b)* a *narrative* construction. Even so, the resulting self is patchy—because 'we tell of ourselves fitfully and patchily' (Bruner & Kalmar, 1998, p. 323)—unless 'metanarrative' construction is deployed to bring different self narratives together to a coherent whole, which occurs principally when outside (i.e. social) circumstances demand it.

The crucial features of a narratively constructed self in Bruner's sense are these. First, there is an emphasis on agency—or its opposite, which Bruner (1994) terms 'victimicy' (p. 41): stories have protagonists; thus, the narrative self is that of an agent. Secondly, since narratives are cultural practices, our 'life stories [are] highly susceptible to cultural, interpersonal, and linguistic influences' (1987, p. 14); thus a narrative self is culturally and socially, even dialogically, shaped. Thirdly, because no two instances of narration are quite alike, and the social and interpersonal requirements of each instance of self-narration admit of significant variation, the narratively constructed self is 'somewhat unstable over extended time' (Bruner & Kalmar,

---

[4] This contention is, at best, highly debatable: not all chronological descriptions are narrative in form. I discuss Bruner's claim in more detail in the next chapter (§ 6.1.3).

1998, p. 308). So, for all the consistency a narrative structure supposedly provides to otherwise disparate rememberings, the resulting self is never an authoritative version, but rather, as just mentioned, 'a perpetually rewritten story' (Bruner, 1994, p. 53)

I will discuss the difficulties faced by Bruner's account, together with those arising from other narrative accounts of the self, in the next chapter. But first, let me consider quite what is meant by 'narrative', and then introduce narrative accounts of the self proffered by philosophers.

5.3 Narrative selves?

5.3.1 Defining 'narrative'

What exactly is meant by 'narrative'? Much of the literature on narrativity and the self, whether psychological or philosophical, lacks such a definition.[5] That may perhaps be explained by noting that everyone may be presumed to have a good grasp of what a story is and of what is involved in story-telling. It may also be due to the preponderance of appeals to narrativity in a wide range of academic disciplines (see e.g. Brockmeier & Carbaugh, 2001).

Yet, the different authors cited in this chapter do use 'narrative' in subtly different ways. Bruner (§ 5.2.4) and Dennett (§ 5.3.3) overtly take a narrative to be very much like a literary novel—a view that runs into difficulties, which I will discuss in the next chapter (§ 6.1.2). Marya Schechtman's (1996; 2007) notion of narrativity (§ 5.3.4) is a looser one that allows narrative processes be 'largely implicit and automatic'. But she also identifies narrativity with certain mental activities—identifying with one's actions, awareness of their causal relations, having affective connections with one's past—that do not obviously require narrative form.

The problem with looser definitions of 'narrative' is that they easily become trivial and uninformative. As Galen Strawson (2004) puts it:

---

[5] This absence of a definition of 'narrative' is not limited to works concerned specifically with narrativity in relation to the self. Even Gregory Currie's (2010) *Narratives and Narrators: a philosophy of stories* deliberately eschews defining its first titular term.

> Well, if someone says, as some do, that making coffee is a narrative that involves Narrativity, because you have to think ahead, do things in the right order, and so on, and that everyday life involves many such narratives, then I take it the claim is trivial. (p. 439)

On the other hand, too strict a definition of what amounts to a 'narrative'—that it should have the characteristics of a *literary* narrative, say—would exclude the every-day story-telling practices discussed above (§ 5.2.1). What is required, then, is a definition of 'narrative' that is neither too restrictive in excluding such practices, nor so loose as to become trivial.

Peter Goldie's (2012) account of narrativity has some similarities with Schechtman's, as does the notion of *narrative* he employs and (unlike most others) takes some care to define:

> A narrative or story is something that can be told or narrated, or just thought through in narrative thinking. (p. 2)

The notion of 'narrative thinking' widens the applicability of 'narrative' beyond instances requiring a 'public act of narration' (p. 4). This would seem to allow Schechtman's 'implicit and automatic' narrativity. But what *is* narrative thinking? Broadly, it means thinking that is *structured* in a particular way that resembles (and, if need be, enables) overt narration. A narrative, Goldie continues,

> is more than just a bare annal or chronicle or list of a sequence of events, but a representation of those events which is shaped, organized, and coloured, presenting those events, and the people involved in them, from a certain perspective or perspectives, and thereby giving narrative structure—coherence, meaningfulness, and evaluative and emotional import—to what is related. (p. 2)

Narrative perspective, Goldie adds, can be internal or external, that is, from the point of view of a protagonist or that of a narrator or, one might add, following Schechtman (2011b), a critic.

As for the three elements of narrative structure: *coherence*, Goldie (2012) notes, is not achieved merely by linking events with causal explanations, but by selecting particular features or details in a causal history for their salience and interest. *Meaningfulness* may be obtained from both an internal and an external perspective; in the first case, by dwelling on a protagonist's 'thoughts, feelings, and actions' in relation to the events in the narrative; in

the second case, because the protagonists' thoughts and feelings 'throw light on why the narrative was related (or just through through) in that particular way' (p. 17). Finally, the *evaluative and emotional import* of a narrative, which is closely linked with its meaningfulness, consist in why and in what way the events of a narrative matter, and what feelings the narrative may precipitate.

It is important to add that, on Goldie's view, narrative structure overall 'is a property that can be possessed as a matter of degree: it can be present to a greater or a lesser extent' (p. 13). Thus, it would seem that there may be degrees of coherence, meaningfulness, and evaluative and emotional import. One might wonder how small a degree of any of these is permitted before something ceases to be a narrative, but let us assume that a narrative requires that at least *some* degree of all of these is present.

Goldie's definition seems wide enough to cover the instances of narrativity I discuss in this chapter, yet not so loose as to be open to the Strawsonian charge of being trivial. Let me attempt, next, to classify the different ways in which narrativity may be, and has been, said to relate to the self.

5.3.2 Narrativity and the self: four positions

Marya Schechtman (2011b) draws a useful 'tentative distinction' (p. 395) between two broad families of views on the relation between narrativity and the self. The first group comprises various accounts of the self as 'constituted by narratives' (ibid.), among whose proponents Schechtman lists Alasdair MacIntyre, Charles Taylor, Paul Ricœur, Jerome Bruner, and herself, as well as, in a somewhat different vein, Daniel Dennett. The second group of views 'links selfhood to the capacity to think in narrative terms and to offer narrative explanations' (p. 398). This includes Katherine Nelson's account of the development of the self (§ 5.2.1). I will here try to refine Schechtman's distinction by identifying two subdivisions in each of the two families of views.

First, then, there are what I will call *narrative constructionist* accounts of the self: these share the view that the self is, in some way, a narrative *production* or *construction*. Among them we can distinguish further between what I will call, respectively, *strong* and *weak* narrative constructionist ac-

counts. Strong narrative constructionist accounts take the self both to be a narrative construct and to have a narrative structure—in short, on such views the self *is* some kind of a narrative. They include Bruner's view discussed above (§ 5.2.4) and, I take it, Schechtman's own (1996) account, which I discuss below (§ 5.3.4). Weak narrative constructionist accounts do not take the self to be narrative in structure, but still hold that it is, in some way, a product of narrative activity. The paradigm case of such an account is Dennett's (1992) view of the self as a 'centre of narrative gravity' (see § 5.3.3). Narrative constructionist accounts of the self all accord priority to narrative processes over the self. Strong narrative constructionist views, additionally, conceive of the self as narratively structured, while weak narrative constructionist views merely take it to be a product of narrative activity.

Secondly, there is a still more diffuse group of views that, in some way, as Schechtman puts it, 'link' the self to narrative capacity, without however taking it to be a narrative construct or product. To the contrary, on these views, the self can very well be conceived of as the *producer* of narrative activity. We can call these views *narrative-capacity* views. Here, too, a further sub-distinction will be useful. One position is to take narrative capacity and activity to be *essential* for a self: something is a self only if it has the capacity to engage in narration and/or narrative thinking and, additionally, keeps exercising that capacity. An example of such a view is David Velleman's (2005) response to Dennett (§ 5.3.3). But then there are also *simple* narrative-capacity views, which hold merely that selves can and often—but not necessarily—do engage in narrative activity. My system view of the self sits on this lowest rung of the narrativity ladder.[6]

In sum, there are, broadly, four positions regarding the relation of self to narrativity. Strong narrative constructionist views take the self to be narratively constructed and narrative in structure. Weak narrative constructionist views regard the self as a product of narrative activity but not itself a narrative. Narrative-capacity views take the self to be a producer of narrative activity, but differ on whether narrativity is an *essential* activity of the

[6] I would like to claim Peter Goldie as a fellow occupant of this position, since his own (2012) view on narrativity is that while we have a 'narrative sense of self . . . the sense that one has of oneself in narrative thinking, as having a past, a present, and a future' (p. 118), he insists that this *sense* of self is neither a narrative *self*—nor the *only* sense of self we have.

self. The remainder of this chapter will be an exposition of what I take to be representative samples of three of these four positions: Dennett's weak narrative constructionist view and Velleman's development of it into an essential narrative-capacity view (§ 5.3.3), and Schechtman's strong narrative constructionist view (§ 5.3.4). Most of the next chapter will then be devoted to a sustained critique of all three views.

5.3.3 The self as fiction—or narrator

Short of the strong narrative constructionist view of the self as narrative in structure, the prominent weak narrative constructionist view treats the self as an *abstraction* from our narrative practices. This is Daniel Dennett's (1992) suggestion that the self is a 'center of narrative gravity'. Analogous to the concept of a centre of gravity in physics, the self is a theoretical, fictional abstraction—a focal point, as it were, of the every-day stories we tell about our lives. And, just as the centre of gravity of a physical object shifts when the shape of the object is interfered with, so the self *qua* centre of narrative gravity is not fixed, but gets changed and updated over time: 'We cannot undo those parts of our pasts that are determinate, but our selves are constantly being made more determinate as we go along in response to the way the world impinges on us.' (p. 110)

Thus, like Bruner's strong narrative constructionist self, Dennett's centre-of-narrative-gravity self is unstable and subject to change over time. But, whereas in Bruner's account the self *is* the story (or the metastory), here the self is something like the inferred central character of the stories we tell of ourselves. This may seem odd: if I tell a story about my life, why need I infer from my own story that someone I refer to as 'I' is its protagonist? But this is precisely Dennett's point. To be sure, there is already a persistent human animal, perhaps even a persistent experiencing subject (though Dennett would doubt this, since 'the unity of normal life is an illusion' (p. 111) given to us by our self narratives), who is telling his life story. But the 'I' or self that is the protagonist of my narrative, in so far as he has a rich psychological 'inner' life, is an abstraction from the narratives I tell. In telling stories about ourselves, we create their fictional central character, our self.

The most serious question arising from Dennett's account is this: *who* (or what) exactly is telling these stories? Even if we were to accept Dennett's contention that our perceived psychological unity is itself an artefact of our narrative activity—that there is no 'I' prior to narration—we must still ask what processes result in the production of our narratives. What is it that enables a loquacious animal to spin a narrative about itself? It is plausible to assume that, analogously to Klein's three 'enabling systems' for autobiographical memory (see § 4.2.3), something that is capable of narrating its own story with any degree of verisimilitude requires some prior self-awareness, a sense that its experiences are its own, and an ability to place events in a temporal sequence. And if such self-representational capacities are required for us to tell our tale, it seems misguided to locate the self at the end of its narrative production line. I will return to this issue in the next chapter (§ 6.2.1).

A comparatively lesser concern about Dennett's account is his characterization of the self as a fiction. For though Dennett's self is fictional, this does not mean that it is an idle concoction without any useful application. Hence the analogy with a centre of gravity in physics: 'when I say it is a fictional object, I do not mean to disparage it; it is a wonderful fictional object, and it has a perfectly legitimate place within serious, sober, *echt* physical science' (p. 104). Likewise, the self as a centre of narrative gravity has a legitimate and useful place in human psychology—elsewhere, Dennett speaks of the 'reality' of our fictional selves (1991/1993, p. 412), though it is somewhat unclear whether he simply means that they are real *qua* fictions, or something more than that.

Dissatisfied with this metaphysical nonchalance on Dennett's part, David Velleman (2005) develops Dennett's account. In his view, the self is 'both fictive and true' (p. 58)—'fictive' in the way Dennett suggests but 'real' in the sense that 'we really are the characters whom we invent' (ibid.). This reality of the narrative self is derived, on Velleman's account, from the insight that our narratives about ourselves do not merely describe or explain our behaviour, but also determine it. For instance, when we narratively state an intention that we're about to do something, we then—other things being equal—act on that stated intention. More globally, we generally attempt in our behaviour to cohere with whatever autobiographical narrative we tell of ourselves. There is, in Velleman's view, a 'feedback

loop' that ensures that our lives correspond to our life-stories both by the narration of our actions and by the acting out of our narratives.

This ability to turn narratively formed intentions into actions, Velleman insists, is what makes us 'autonomous agents'. Perhaps so. A discussion of what might be meant by 'autonomy', and in what way Velleman's might be a plausible account of it, would take me too far off course here. The emphasis on agency, however, is of some importance, because, as in Bruner's account discussed above, and of course as in Locke's account, it links the notion of self with that of action. However, unlike Locke, for whom the ownership of a past action is indicative of its being part of the same self, Velleman takes it that what makes an action mine is its being part of my narrative.

Thus the self, in Velleman's view, is not the fictional protagonist of Dennett's account, nor 'the autobiographer's reflective representation' (p. 70), i.e. the autobiographical narrative, as it is in Bruner's account. Rather, the self is the *narrator* of that narrative—and not an impartial or uninvolved narrator, but rather an 'inner locus of agential control' (p. 71), a decider and determiner of actions. It is not someone to whom things happen, but someone who makes things happen.

There is, I think, something right about this modification: it is the emphasis on the narrator as agent, which goes some way to alleviate the lack of a producer of narrative in Dennett's account. But Velleman's notion of 'agential control' goes too far. If he held that, by way of the stories we tell, we *take* ourselves to be decisive agents *seemingly* in control of our actions, so as to make sense of what we do by means of a narrative agential interpretation of diverse and often unconscious processes, that would preserve some of the essence of Dennett's account and nicely tally with empirical psychology. But he seems to be saying rather more than this, *viz*. that we are *genuinely* in control of our actions by making ourselves act in accordance with a pre-established narrative. And that, given what little insight we have into our judgements and reasons for action (Nisbett & Wilson, 1977; Carruthers, 2010), seems implausible. I'll return to this point in the next chapter (§ 6.3.1).

It is a little difficult to situate Velleman's (2005) account on the spectrum of narrative views of the self I have sketched above (§ 5.3.2). Some of his remarks—e.g., 'We invent ourselves' (p. 58)—seem to position his view

on the *narrative constructionist* side of Schechtman's distinction, among the views that take selves to be narrative constructions or productions. But his characterization of the self as *narrator*, the one who actively constructs his narratives, suggests otherwise. Indeed, Velleman explicitly disagrees with Dennett in insisting that our narrative activity requires a prior capacity, which he calls 'narrative intelligence' (p. 72). I take it, then, that Velleman's position is really a narrative-capacity view, rather than a narrative constructionist one. As to what *kind* of narrative-capacity view it is, clearly Velleman takes narrative activity to be the *essential* activity of the 'self-narrator', for it is this activity that makes him an 'autonomous agent', which is Velleman's main concern. The self, then, though not narratively *constructed*, is essentially a narrator. And that too is a problem, as I shall explain in the next chapter (§ 6.3).

Meanwhile, there is still the *strong narrative constructionist* position to consider, according to which the self is not only a narrative construct, but also narrative in structure.

5.3.4 'Narrative self-constitution'

Schechtman's (1996) 'narrative self-constitution view' (p. 93) combines two features of narrative views of the self already discussed: the idea that selves are 'self-creating' (p. 95) in some way, which is also a tenet of Dennett's account, and the idea of the self as a narrative construct, as found in Bruner's view. Moreover, Schechtman's view preserves the Lockean conflation of 'self' with 'person'.[7] According to Schechtman,

> we constitute ourselves as persons by forming a narrative self-conception according to which we experience and organize our lives. This self-conception and its operations are largely implicit and automatic. (2007, p. 162)

With this narrative self-conception in place, 'we experience the present in the context of a larger life-narrative' (ibid.), that is, the self narrative not

[7] Schechtman later (2007) resiles from the self–person conflation in response to objections, which I discuss in § 6.1.1. And in her latest (2014) monograph, even her narrative account of *persons* has been superseded by what she calls the 'person life view', according to which 'persons are defined in terms of the characteristic lives they lead' (p. 110). As this is a new account of persons rather than of selves, I forbear to discuss it here.

only serves to *organize* our experiences, but also determines *how* we experience the present. Here, something like Bruner's narrative 'mode of thought' seems to operate on our experiences. For instance, as a Schechtmanian self, I might experience the noise of strimmers and lawn-mowers in the grounds outside my window not merely as an occurrent impediment to concentration, but as an instance of a recurring motif in my life story wherein various extraneous forces (gardeners, builders, guitar-playing neighbours, children, dogs, magpies…) conspire to annoy me by making a racket whenever I'm trying to work. I will look at the possible pitfalls of such thinking in the next chapter (§§ 6.1.2, 6.3).

But what exactly does it mean to say that a person or self '*constitutes*' herself 'by forming a narrative self-conception'? What has been said so far about *having* a narrative self-conception is compatible with Goldie's (2012) view that we engage in narrative thinking about our lives and so may have a narrative *sense* of self (among other senses of self), without any of this amounting to a narrative *self*. Schechtman does not state in what sense she uses 'constitution' in her account, but a common usage in the personal-identity literature suggests that for A to 'constitute' B means that A and B are 'made of the same matter' but have different persistence conditions (Olson, 2015). On this reading, a human being 'constitutes' a person/self, in Schechtman's account, 'by forming a narrative self-conception'. Thus, it would seem that once (and for as long as) one has a narrative self-conception, one is *ipso facto* a Schechtmanian person/self.

This might suggest that Schechtman espouses an *essential narrative-capacity* view, rather than being an exponent of the *strong narrative constructionist* position. But in the original statement of her view, she also seems to imply that the narrative *is* the self. Although Schechtman (unlike Bruner) does not quite spell it out that way, in *The Constitution of Selves* (1996), in which 'personal identity' and 'self' are not distinguished, she states her position thus:

> a person's *identity* . . . is constituted by the content of her self-narrative, and the traits, actions, and experiences included in it are, by virtue of that inclusion, hers. (p. 94, original emphasis)

It is difficult to understand this except as meaning that the content of the narrative 'constitutes' the self. This would seem to make Schechtman's view a strong narrative constructionist account.

How, then, do we get a narrative self-conception? The first requirement of narrative self-constitution indeed seems a cognitive effort akin to Bruner's narrative 'mode of thought': it is 'that an individual conceive of his life as having the form and the logic of a story . . . where "story" is understood as a conventional, linear narrative' (p. 96). This conception of one's life as having a narrative form supposedly provides both coherence and intelligibility to one's experiences: experiences gain their meaning from the context of the life story in which they are situated. Obviously, in order to conceive of one's life in this way, one must have some grasp of how 'conventional, linear narratives' work. (By 'linear' I take it Schechtman means that the narrative does not involve forks or loops.) Schechtman does not discuss the acquisition of narrative capacity in development, but we must assume that the practices studied by both Fivush and Miller (§ 5.2.1) are required for anyone to develop the ability to form a narrative self-conception.

Must this narrative form necessarily be linear? Schechtman (1996) argues that it must be, since the self—in her sense of 'characterization', that is, 'which beliefs, values, desires, and other psychological features make someone the person she is' (p. 2)—must account for four features: survival, moral responsibility, self-interested concern, and eligibility for compensation. Only linear narratives, Schechtman contends, can satisfy these criteria. Indeed it seems that concern for the future and responsibility for past actions require linearity to the extent that there should be no bifurcations of self over time and that different events stand in a stable temporal relation to one another.

In seeking to give a view of the self that accounts for these four features, Schechtman's project is clearly situated in a Lockean tradition—but explicitly not in that strand of the tradition that equates Locke's account of the self with the persistence of *memory*. The extended consciousness that makes for the same self in Locke is about appropriating past actions and experiences, which happens 'if they affect *present* consciousness, causing the person pleasure or pain in the *present*' (p. 109, original emphasis). Memory will have some role in this, of course, but so have 'affective traces' (p. 110) even in the absence of explicit memory. Schechtman's narrative form is supposed to provide the framework or 'organizing principle' (p. 113) by which we make sense of those affective traces

impinging on our present consciousness, giving them meaning and coherence.

At the same time, the case for a linear narrative self is weakened somewhat by the influence of 'affective traces' of our past on present consciousness. The recall of a strong emotion is likely to be unbidden or at least unreflected, and its association with affective traces of a similar quality will mostly be out of sequence. Suppose some present event fills me with a sense of crushing disappointment and conjures up the recollection of other events in my past that gave rise to a similar affective response. It is plausible to assume that these recollections will not appear in their temporal sequence, but impress themselves on me in a seemingly random order, or perhaps in order of the magnitude of the emotion. And—assuming that my life so far has not been an unadulterated series of crushing disappointments—they will amount only to a transient, incomplete and skewed representation of myself. At that particular moment of my self-awareness, my 'disappointed self' is not a linear narrative, but a mass of negatively valenced affective traces.

Of course, I will usually get over my disappointing episode sooner or later, and in doing so may well be aided by the recollection of overcoming those previous disappointments, setting them in their temporal context and in a wider perspective which also includes other, happier incidents in my life. This, one might say in defence of Schechtman's view, is precisely where the narrative self is being deployed: by inserting the present experience into a continuing storyline, I can make sense of present events in the light of what has gone before. But that would suggest that the narrative self is a kind of cognitive repair-kit, always a step or several behind present consciousness. It does not then so much condition present experience as alter or 'remedy' it *post hoc*. And that is assuming that the narrative kit-bag already contains the requisite tools for making sense of the present. Some self narratives may not, however—I discuss that possibility below (§ 6.3.2).

But let me go back a step and consider how Schechtman's narrative self is configured. For we might wonder just how solid and detailed the narrative self-conception is by which we 'constitute' ourselves. Clearly it is not a fixed narrative—as long as one's life continues, so does the narrative receive additions and amendments. But nor is it a fully formed narrative. Narrative self-constitution, according to Schechtman (1996), does not re-

quire '[t]he formal construction of an autobiography' and 'does not have to be self-conscious' (p. 105). Here, then, is a puzzle: in what sense precisely is Schechtman's self-constitution *narrative*? The answer seems again to hark back to Bruner's notion of a narrative 'mode of thought': we supposedly think or conceive of our lives as something like a story, with certain narrative features that go beyond mere diachronicity, including one's own characterization as the story's protagonist with particular 'traits, talents, likes, and dislikes' (p. 114), the interpretation of one's actions in terms of goals and reasons, and, presumably, certain overarching themes and motifs. I will consider in the next chapter whether we do in fact construe our lives in this way, and whether and to what extent it is fruitful to do so (§§ 6.1, 6.3).

Additionally, Schechtman's narrative self is subject to what she calls the 'articulation constraint' (ibid.)—a person must 'be able to articulate her narrative locally when appropriate' (2007, p. 163). Thus, while the bulk of our life story may be 'implicit' and unconscious, we produce explicit excerpts from it when the social situation requires it and we actively narrate part of our life story. There is both a parallel with and a subtle difference from Bruner's account here. Both his and Schechtman's view involve the localized telling of (part of) one's life story. But where for Bruner the self narrative is continuously reshaped or 'rewritten' by these acts of explicit narration, it seems that for Schechtman, such narrative acts are instances of an existing implicit narrative self being made explicit.

Another constraint on our narrative self that Schechtman stipulates is the 'reality constraint': the self narrative must 'conform to what we are generally accepted to know about the basic character of reality and about the nature of persons' (ibid.). This sounds reasonable enough, but nevertheless also vague enough to admit less or more serious inaccuracies—both 'errors of fact' (1996, p. 121) and 'interpretive inaccuracies' (p. 123)—into the self narrative. And factual and interpretative inaccuracies are a problem not only for narrative constructionist views of the self, but also for narrative-capacity views that take narrative to be an essential activity of the self, as I'll discuss shortly (§ 6.3.1).

Having now presented examples of strong and weak narrative constructionist views as well as an essential narrative-capacity view of the self, it is time that I discussed their problems and shortcomings. There are enough of these to take up most of the next chapter.

Chapter Six

The author, not the tale

6.1 The self is not a story

In the title of his latest piece in a series of articles against narrative accounts of the self, Galen Strawson (2015) insists: 'I am not a story'. Nor is he. And neither am I, or you, or anyone else. Nor is anyone's *self* a story. In this section, I will argue against *strong narrative constructionist* accounts, which hold not only that the self is narratively constructed, but also narrative in structure. And I may as well begin with Strawson's criticisms, and Schechtman's responses to them.

6.1.1 Strawson v. *Schechtman*

It is a feature of strong narrative constructionist accounts that constructing a self through narrative involves that one should experience one's life in a narrative form. As Schechtman (2011b) puts it in her summary of such accounts, 'the *lives* of selves are narrative in structure' (p. 395). Thus, a narrative 'organizing principle' (1996, p. 113) supposedly operates on the way we process our experiences. According to these views, we only make sense of this or that episode in our lives by placing and understanding it in the context of a narrative structure.

Galen Strawson (2004; 2012; 2015) disagrees vehemently with this thesis:

> I have a past, like any human being, and I know perfectly well that I have a past. I have a respectable amount of factual knowledge about it, and I also remember some of my past experiences 'from the inside', as philosophers say. And yet I have absolutely no sense of my life as a narrative

> with form, or indeed as a narrative without form. Absolutely none.
> (2004, p. 433)

Strawson's testimony poses a problem for strong narrative constructionist accounts. If the contention of these accounts is that in order to have a self over time, one must conceive of one's life in a narrative way, the only conclusions available here are (i) that Strawson, given his non-narrative disposition, does not have such a self, (ii) that Strawson is somehow mistaken about his own self-conception, or (iii) that strong narrative constructionist accounts are mistaken in imputing a narrative self-conception to us all.

The first possibility here would appear to be contradicted by Strawson's assertion that he is well aware of having a past of which he has factual and experiential knowledge. Granted, in Strawson's (2009) own 'revisionary' use of the term 'self', which he reserves for any continuous period of consciousness, all this does not amount to a self. But, clearly, both his factual knowledge and 'inside' awareness of his own past are instances of self-representation over time. There is, on most reasonable definitions of 'self', a persisting self evident here.

Second, one might think that Strawson is simply in error about his own self-conception—that he *does* have a narrative self-conception without realizing it. But that would be a highly dubious move to make: Schechtman's (1996) account of the self requires that one should have a first-personal conception of one's life in narrative form. Thus, first-personal evidence such as that proffered by Strawson is perfectly relevant in this context. Indeed there is no other means currently available of deciding whether someone experiences his life in a particular way than to solicit first-personal testimony. Strawson's is unequivocal on this point. He has 'no sense of [his] life as a narrative . . . Absolutely none'. Some might still insist that Strawson is being narrative in ways he does not realize. As Valerie Hardcastle (2008) observes, 'the very descriptions that Strawson provides of himself are strikingly narrative in nature' (p. 27). That may be so, but this only shows that Strawson has narrative *capacities* and is able to deploy them when talking about himself. Even if someone's account of himself can take narrative forms locally, this does not yet mean that he conceives of his life—or his self—as a story. One may engage in narrative *practices* without having a narrative *self-conception*.

This leaves us with the third option, that strong narrative constructionist accounts overstate their case in holding that a self over time must necessarily be narrative in form. Any one counterexample—such as Strawson's—invalidates that thesis. But notice that while Strawson is emphatic in not sharing the narrative self-conception that is at the core of strong narrative accounts of the diachronic self, he is not saying that *no one* has a sense of his or her life as a narrative. On the contrary, he is perfectly aware that some do. Self-experience, Strawson (2007) notes, varies between different 'temporal temperaments' (p. 85):

> To be *Narrative* is [N] to see or live or experience one's life as a narrative or story of some sort, or at least as a collection of stories. (p. 86)

Conversely,

> To be *non-Narrative* is not to live one's life in this way; one may simply lack any Narrative tendency, or [like Strawson himself] one may have a positively anti-Narrative tendency. (ibid.)

In addition, Strawson makes a distinction between 'Diachronic' and 'Episodic' temperaments:

> If one is *Diachronic* [D] one naturally figures oneself, the self or person one now experiences oneself to be, as something that was there in the (further) past and will be there in the (further) future. (ibid.)

On the other hand,

> the defining feature of being *Episodic* is that [E] one does not figure oneself, the self or person one now experiences oneself to be, as something that there was in the (further) past and will be there in the (further) future . . . (ibid.)

Crucially, while being 'Narrative' entails that one is also 'Diachronic', it is possible to be 'Diachronic' without being 'Narrative': experiencing oneself as something with a past and future does not require narrative form. Further, being 'Episodic' and being 'Diachronic' are conceived as opposite ends of a spectrum, along which one's temporal temperament need not be fixed, but 'can vary considerably according to what one is thinking about' (ibid.) or according to one's age.

If these different temperaments reflect how different people experience themselves (perhaps at different times in their lives), it follows that a *narrat-*

ive self-conception is just one of a range of possible self-conceptions. And if one's temperament can vary situationally along the 'Diachronic'–'Episodic' dimension, so it can vary between being 'Narrative' and being 'non-Narrative'. One might, for instance, conceive of particular passages of time in one's life as having narrative coherence, while allowing that others have not.

Responding to Strawson's criticism, Schechtman (2007) goes to some length in trying to revise her narrative self-constitution view to accommodate the existence of different temporal temperaments. While accepting his self-characterization as an 'Episodic', she observes that Strawson admits that he *qua* current experiencing self stands in a special relation to other parts of his life *qua* human being. These relations 'within his human existence,' Schechtman contends, 'contain much of what is involved in having a self narrative' (p. 168). As I've already noted, Strawson's awareness of his past as being *his* is a kind of diachronicity, despite his 'Episodic' temperament. But while diachronicity is indeed something that's 'involved in having a self narrative', it is not sufficient for having one: diachronicity need not take a narrative form.

But Schechtman has a second response:

> Much of Strawson's argument against the narrative view is based on the fact that he does not experience his entire human life in narrative terms—that there are different *selves* within his human existence. (ibid.)

This, Schechtman argues, does not invalidate the narrative self-constitution view, for her view distinguishes between the human being (and its persistence conditions) and the narratively constituted self or person. Schechtman's view can thus accommodate several selves per human life 'if each self is constituted by a narrative internal to it' (ibid.).

Schechtman admits that her two responses 'are in some tension with one another' (ibid.)—one claims to detect the rudiments of narrativity in Strawson's diachronic awareness of his whole life, while the other admits of multiple narrative selves in a lifetime. She proposes to resolve this tension by introducing a new self–person distinction, whereby the *self* is 'the subject of experience, the "I"', and a *person* is 'the bearer of certain complex social capacities that carry important practical implications', 'a moral agent', 'a reasoning creature', 'a creature capable of a range of complex [social] relationships' (p. 169). *Pace* Locke and Schechtman's own earlier (1996)

account, the persistence conditions of persons and selves may thus differ (as I argued in § 2.4.1).

Now, Schechtman (2007) suggests that the narrative self-constitution view contains accounts of the narrative construction of *both* persons and selves as understood here. Her 'narrative account of *persons* (PN)' is that

> to constitute oneself as a person, one must recognize oneself as continuing, see past actions and experiences as having implications for one's current rights and responsibilities, and recognize a future that will be impacted by the past and present. One need not deeply identify with past or future actions and experiences, care about them, or take an interest in them, but one does need to recognize them as relevant to one's options in certain fundamental ways. (p. 170)

This account, Schechtman notes, is compatible with Strawson's self-experience as an 'Episodic' (indeed it seems tailor-made to accommodate Strawson's episodicity). She does however anticipate his objection that any sense in which narrativity is compatible with his notion of episodicity is trivial, like (using Strawson's example quoted in § 5.3.1) the 'narrative' of how to make coffee. Schechtman replies that a person-constituting narrative is relevantly different from a coffee-making narrative, in that 'it is an explanatory account of how actions and events lead to other[s]' (p. 172). But here too Strawson (2012) has a rejoinder: 'causal explanation, psychologically significant causal explanation, isn't always narrative' (pp. 82–83). Thus, even if one were fully to accept PN as an account of what constitutes a person, it does not necessarily amount to a *narrative* account. We can recognize the relevance of past actions and experiences to our rights and responsibilities etc. without spinning them into a narrative, even if we verbalize their relations. The conjunction 'because' does not make a story.

As to Schechtman's (2007) revised 'narrative account of *selves* (SN)', it is this:

> For an action or experience to belong to my*self* I do need to identify with it or care about or take an interest in it. Temporally remote actions and experiences that are appropriated into one's *self* narrative must . . . condition the quality of present experience in the strongest sense, unifying consciousness over time through affective connections and identification. . . . In *this* sense of narrative, actions and experiences from which I am alienated, or in which I have none of the interest that I have in my current life, are not part of my narrative. (p. 171)

Here too it seems that the conditions Schechtman lays down for past actions and experiences to be incorporated into one's present self do not necessarily require a narrative. Identifying with a past experience, appropriating a past action as one's own, caring about and having 'affective connections' with one's past may require some cognitive processing, but this does not need to take the form of a narrative.

Indeed it remains a puzzle just in what sense Schechtman's revised view is a 'narrative' account. Is it still a narrative constructionist account in either the weak or strong sense, one which requires that the self be *narratively constructed*? If so, it cannot serve as a response to Strawson's non-narrative temperament. Or are non-narrative cognitive processes that satisfy the conditions of PN and SN equally acceptable for the 'constitution' of a person or a self? In that case, it would seem odd to insist on the label 'narrative'—unless its purpose were merely to indicate that we *sometimes* make use of narrative practices in our self-conception. This would place Schechtman's revised account among the simple narrative-capacity views.

Such a view could tally with the temporal temperaments Strawson posits, for just as Schechtman's responses do not quite establish that narrativity is necessary for a self if other cognitive means are available to meet her conditions, so Strawson's anti-narrative stance does not establish that *no one* does or can resort to narrative self-conceptions. Different temporal temperaments may simply mean that different individuals have different self-conceptions over time.

It seems, however, that such a position is weaker than Schechtman would allow. In describing her remaining disagreements with Strawson, Schechtman (2007) insists that 'affective concern' for one's past can only be accomplished by including one's past in a self narrative (p. 177). I will return to the question of affective concern shortly (§ 6.1.3); for now, I take it that Schechtman's revised view, while no longer obviously a *strong narrative constructionist* account, at least accords an essential role to narrativity. Thus, she seems to have retreated to an *essential narrative-capacity* view. The difficulties facing such views will be discussed in § 6.3.

Meanwhile, strong narrative constructionist accounts face some further difficulties and objections.

6.1.2 What's the story?

If the self is a story, what sort of story is it? Just how literally—and literarily—are we to take strong narrative constructionist accounts?

Jerome Bruner, for one, unabashedly draws parallels between the life stories of his subjects and literary forms and characters. In his (1987) study of the life narratives of four members of a Brooklyn family, the adult son's story is likened to a *Bildungsroman*: 'Carl is a young Werther.' (p. 29) For his sister, meanwhile, 'the tale is more like the young Stephen Hero in the discarded early version of *Portrait*' (ibid.). Bruner's subjects, I take it, do not themselves resort to these literary comparisons in narrating their lives; rather, it is Bruner who in evaluating their tales for an academic audience finds it useful to do so. But he would be unlikely to do so in the context of introducing and defending the notion of 'life as narrative' if he did not see relevant similarities between the life stories of ordinary people and the novels of Goethe and Joyce. Dennett (1992) too likens our self narratives to literary creations when he claims that 'we are all virtuoso novelists' (p. 114).

This, Peter Lamarque (2007) argues, is a grievous mistake. Literature is a form of art, an artifice, operating on particular and distinctive principles. Thus, contrary to what Bruner might think when reading a novel, 'the role of fictional characters in literature does not closely mirror the role of real people living real lives' (p. 118). Conversely, the application of literary figures to real-life narratives has a 'distorting and pernicious effect on the self-understanding that such narratives are supposed to yield' (p. 119).

Lamarque shows the distinctness of literary from real-life narratives with reference to five principles. The first of these concerns the nature of literary characters as opposed to actual people: 'In literary works character identity is indissolubly linked to character description.' (p. 120) That is to say, the identity of a literary character is entirely dependent on the narrative and determined by how the character is described. A real person's identity is not, in Lamarque's view, derived from description, but 'constrained by factors independent of narrative' (p. 130), to wit, the *facts* of a person's life. Nor is the identity of a real person accessed in a manner comparable to the 'external perspective' a novel or its reader has on the characters.

Secondly: 'In literary works not only are characters and incidents presented to us but attention is conventionally drawn to the modes of

presentation themselves.' (p. 122) A rather wonderful example of this occurs in Nabokov's (1957) first description of Professor Pnin:

> Ideally bald, sun-tanned, and clean-shaven, he began rather impressively with that great brown dome of his, tortoise-shell glasses (masking an infantile absence of eyebrows), apish upper lip, thick neck, and strong-man torso in a tightish tweed coat, but ended, somewhat disappointingly, in a pair of spindly legs (now flannelled and crossed) and frail-looking, almost feminine feet. (1960 edn, p. 7)

Such a description does not merely introduce the character but invites the reader to make a first evaluation of Pnin that isn't given by the contents of the description but in its slightly oddly chosen, comically juxtaposed epithets (foreshadowing the character's somewhat erratic command of English) and the gentle mockery of its style. Lamarque (2007) calls this the 'Opacity Principle': 'Rather than being merely transparent vehicles for prompting imaginings the descriptions provide a more opaque kind of perspective for observing and making sense of a fictional world.' (p. 122) In real-life narratives, however, one normally aims for transparency; in such contexts the opacity that marks literary discourse 'is a weakness . . . and merely clouds personal characterisation' (p. 130).

Lamarque's third and fourth principles of literary narratives are the 'Principle of Functionality'—'It is always reasonable to ask of any detail in a literary work what literary or aesthetic function that detail is performing' (p. 123)—and the 'Teleological Principle': 'In literary works the explanation of why an episode occurs as it does and where it does often centres on the contribution the episode makes to the completed artistic structure.' (p. 126) Details matter in real-life narratives, too, but in different ways from literary narratives: they are *selected* for relevance rather than *created* for artistic effect; what matters is not their symbolic significance but their pertinence and accuracy. Ascribing symbolic functionality to details in real life is, according to Lamarque, 'misplaced', 'a kind of mysticism' (p. 131)—as is resorting to teleological explanations in real-life narratives, where '[e]xplanations for non-fictional events must stay in the realm of causes and reasons' (ibid.).

Finally, literary narratives invite thematic interpretation; their 'significance and unity' (p. 128) frequently appeal to some overarching theme. And while thematic coherence may sometimes occur in real-life narratives too, it

does so contingently. We may of course choose to recount a selection of our experiences under some thematic heading or other, as I did earlier in my example of being disturbed by outside noise (§ 5.3.4); but even if, in telling the tale of motorized gardening implements, screaming schoolchildren, and guitar-playing neighbours preventing me from working, I were to conclude it with a wry and weary 'It's the story of my life', I would (I hope) be far from assuming or asserting that my occasional exasperation over extraneous noise was anywhere close to being the *whole* story of my whole life.

We might now think that, while Lamarque may be right to point out the distinctness of literary narrative from everyday storytelling, their differences present no objection to strong narrative constructionist accounts of selfhood unless these explicitly appeal to literature as a model for self narratives. Schechtman (2007), for instance, makes clear that her narrative self-constitution view does not state that the self narrative is like a literary narrative; for instance, it does not require a 'unifying theme and direction' (p. 160) as found in Lamarque's fifth principle of literary narratives. Thus, Lamarque's argument may serve against Bruner's willingness to compare his subjects' life stories to novels, and against Dennett's assertion that 'we are all virtuoso novelists' (when one listens to the everyday tales one might hear in an English public house, it quickly becomes apparent that most of us are nothing of the sort; and there are probably no Goethes, Joyces, and Nabokovs among Miller's storytelling subjects in South Baltimore either)—but it need not invalidate a narratively constructed self on Schechtman's (original) view.

Yet, the central contention of Schechtman's account—that the self is, in her words, 'constituted' by narrative—does echo Lamarque's literary 'Character Identity' principle: in literature, a character is determined entirely by the narrative description given of him or her. If we—as selves or persons—are likewise 'constituted' by narrative, in what way do we differ from literary characters? Schechtman (2011b) replies that as characters of our self narratives, and unlike literary figures, 'we are constrained by the facts about the social and natural world in which we find ourselves' (p. 413). While this acknowledges Lamarque's point that the characters of real-life stories are determined by 'factors independent of narrative', consistency with facts is not as straightforward a requirement of

everyday narratives as it may seem. I discuss the difficulties of factual accuracy in self narratives below (§ 6.3.1).

Schechtman's reply to Lamarque goes on: we are not merely the *characters*, but also the *authors* and indeed *critics* of our self-stories.[1]

> [W]e must think of ourselves as authors of our lives insofar as we must make decisions and these must involve reasons and purposes. (pp. 413–4)

This assumption of an authorial role resembles Velleman's view of the self as narrator, as the 'locus of agential control'—though, unlike Velleman, Schechtman is commendably cautious about how much control we really have as authors of our life stories. Her point seems a more modest one: that as narratively constituted selves we do not just sit and wait to see how the story develops, but try to shape it as best we may, given the constraints imposed by the world around us and the limits of our psychological abilities.

This is complemented by our role as critics of our own narratives, which we assume when we engage in self-reflection. However, Schechtman stresses that the triad of our roles as critic, author, and character of our self narratives isn't neatly separable into its constituents; rather, the three functions interact in constructing the narrative self:

> Life is different from literature because we write it as we live it and engage in criticism as we go along rather than after the fact, and because this forces us to take on different roles and perspectives. The creative act in narrative self-constitution is thus neither to produce a tidy and meaningful story out of whole cloth nor to take accidents and contingencies and arbitrarily interpret them as meaningful. It is rather to carve out a meaningful life trajectory by appreciating the contingencies, considering how to respond to them meaningfully, and directing life so much as possible in the direction of that meaning. (p. 414)

In this passage, Schechtman stresses the differences of real-life narratives from literature in respect of Lamarque's third, fourth, and fifth principles (the functionality of details, teleology, and themes), as well as her crucial point that unlike literary characters, their creators, and their critics, we are

---

[1] Schechtman, ever a moving target, here seems to espouse something like a *weak* narrative constructionist view: while an essential narrative-capacity view would cover the 'author' and 'critic' roles, for the self to be a 'character' it has to be a narrative construct (without *being* the story, as the strong narrative constructionist position would have it).

deemed to fulfil all these functions simultaneously with respect to our narratively constructed self.

But it is striking that, even in delineating self narratives from literary narrative, Schechtman still has recourse to the terminology of literature studies when she asserts that we are 'characters', 'authors', and 'critics' of our self stories. We may grant that this usage is somewhat metaphorical—though it exemplifies the difficulties involved in referring to narratives without resorting to comparisons with literature.

Let us accept, then, that real-life narratives do not (and should not) resemble literary narratives. Even so, the claim of strong narrative constructionists that the self is story-like in structure, even in a loose way, has worrying implications concerning what it permits to be part of one's self.

6.1.3 Narrative structure is not necessary

If a self is structured like a story, this may seem to remedy two of the shortcomings of the view I criticized in ch. 4—that the self is constructed from autobiographical *memory*. Recall that a self thus constructed was found to be very 'gappy' and, additionally, totally to exclude one's early years, the period of childhood amnesia (§ 4.5.1). A *narratively* constructed self could be said to have the advantage on both these points.

First, as to the problem of the very large gaps that affect our autobiographical memory even in non-pathological circumstances, a story-like self, it would seem, can compensate for these. It does not actually fill in the gaps. But it provides narrative coherence *across* the gaps. If the isolated events and experiences we recall in autobiographical remembering are connected by some narrative arc, the gaps do not matter so much—what matters, on a narrative constructionist view, is that our islands of autobiographical memory can be part of a coherent narrative. I will later note that even autobiographical narratives can be seriously discontinuous (§ 6.3.2), but let us grant, for the moment, that a narratively constructed self has less of a problem with 'gappiness' than a self constructed purely from memory. Narrativity, we might say, papers over the cracks of our autobiographical knowledge base.

Secondly, it seems that a story-like self also does not suffer from childhood amnesia in the way a memory-based self does: narrativity can com-

pensate here, too. Now, the onset of full narrative capacities, as seen in the previous chapter (§ 5.2.1), occurs even later in child development (after five years of age) than that of autobiographical memory (at about age three). But this does not mean that a self narrative is restricted to lifetime periods in which one was oneself able to narrate. For example, I know—and could include in an autobiographical narrative—that I was born on a showery and stormy day. There is no way I could possibly remember this fact (if it is a fact)—but I can remember being told about it by my mother. Thus, parental tales of a child's early years can become included in the offspring's own self narrative. Indeed it is very likely that some of our earliest 'memories' are reconstructed from parental narratives, rather than from our own memory traces laid down at the time of the events in question. The beginning of life therefore poses no problem for narrative accounts; rather, they are at an advantage over pure memory accounts in being able to incorporate our early years.

If I am happy to grant these advantages that a *narratively* constructed self has over a self constructed from *memory*, it is because these advantages are rather insignificant in comparison to what a story-like self leaves out. Such a self, too, has serious limitations, which I will now discuss. Strong narrative constructionist accounts of the self, let us recall, hold that the self is both narratively *constructed* and also narratively *structured*. Let me now deal with these characteristics in turn.

Concerning narrative *construction*: as Dan Zahavi (2007) points out, there is a more primitive level of selfhood, which he calls the 'experiential self' (p. 185), and which precedes, and thus does not require, any narrative capacities. This bare self-consciousness, the experience of being someone distinct from one's environment, is Damasio's (1999; 2010) 'core self' that is a necessary substructure for the temporally extended consciousness of an 'autobiographical self'. It is at the level of the core self that a present experience already has the quality of 'mineness', an implicit sense

> that the objects I now perceive are being apprehended from my perspective and that the thoughts formed in my mind are mine and not anyone else's. (Zahavi, 2007, p. 185)

> The experiential core self is an integral part of the structure of phenomenal consciousness and must be regarded as a pre-linguistic presupposition for any narrative practice. (p. 191)

Narrative constructionists might rejoin that the experiential self and the sense of 'mineness' it imparts on one's experiences are subsumed into the narrative self by narrative thinking and narrative structure. They might grant these pre-narrative, primitive aspects of self-experience in the 'specious present' (James, 1890), but insist that we still require narrativity to construct a *diachronic* self from what would otherwise be disjointed, episodic experiences.

But, again, narrativity is not required for temporal consciousness. Our 'sense of personal temporality' identified by Klein and colleagues (2004) is more basic even than the contents of autobiographical memory, from which it is dissociable (see § 4.2.3). If a sense of temporality persists in the event of memory loss, it seems unlikely that narrative capacities could be required for it. The apperception of causal relations, too, does not require narrativity, as both Strawson (2012) and Goldie (2012) point out.

What, then, are we to make of Bruner's (1987) assertion, quoted earlier, that '[w]e seem to have no other way of describing "lived time" save in the form of a narrative' (p. 12)? It depends on what Bruner means by 'lived time'. Taking the expression literally, Bruner seems wrong: it is perfectly possible to describe lived time non-narratively, drawing on one's temporal consciousness, placing events in their sequential order—as Goldie (2012) reminds us, a narrative 'is more than just a bare annal or chronicle or list of a sequence of events' (p. 6)—and even identifying their causal relations. But this is not, apparently, what Bruner (1987) means:

> there are . . . other temporal forms that can be imposed on the experience of time, but none of them succeeds in capturing the sense of *lived* time: not clock or calendrical time forms, not serial or cyclical orders, not any of these. . . . Even if we set down *annales* in the bare form of events, they will be seen to be events chosen with a view to their place in an implicit narrative.[2] (p. 12, original emphasis)

So Bruner's sense of 'lived time' involves an 'implicit narrative'—which begs the question. If other forms of temporal experience are available, as he admits, why should these not suffice for having an autobiographical self? What is it that narrative structure supposedly adds to our sense of self over time?

---

[2] Bruner credits Paul Ricœur (1983–5) with this view.

Let me turn, then, to narrative *structure*, the second characteristic of a self according to strong narrative constructionist accounts. Goldie's (2012) three features of narrative structure (which, we should recall, 'can be present to a greater or a lesser extent') are 'coherence, meaningfulness, and evaluative and emotional import' (p. 2). Coherence here, as noted above (§ 5.3.1), is not merely temporal or causal coherence, but a thicker notion of events hanging together in a salient and interesting way—which, I take it, is what gives 'meaningfulness' to the narrative. But is this kind of coherence and meaning required for a diachronic self? In my discussion of biographical discontinuities and 'turning points' (§ 6.3.2) I will note that while we may acknowledge the practice reported by Bruner of constructing turning-point stories about important changes in one's life, it is also possible to accept discontinuities in our lives without a narrative arc. Sometimes, and for some, a more or less sequential ordering of events or lifetime periods is all the coherence that is required for a sense of self over time.

As mentioned earlier, Schechtman's (2007) revised narrative self-constitution view allows for narrative *selves* (unlike persons) to be of limited temporal duration relative to a person's life. Narrative coherence then operates within each self, but not across successive selves. But she also believes that 'there are advantages to making one's self narrative coincide as far as possible with one's person-narrative' (p. 176), that is by having just one (narrative) self for a lifetime, instead of several discrete ones. This is because, she suggests, merely acknowledging one's past as part of one's personal history, rather than including it in one's self narrative, 'seems a recipe for alienation' (p. 177) because of its lack of affective concern.

I have three observations on this. First, it isn't clear how or why affective concern requires narrativity (more on this shortly). Secondly, Schechtman has a point in saying that one self per life is enough to be getting on with: for reasons of cognitive economy, certainly; but, more important, because conceptually it makes little sense to impute more than one self to any one human being (§ 1.7). But—thirdly—wishing to impose a narrative structure on the self can have tremendous disadvantages too. To seek meaning and (narrative) coherence across one's life-span—linking all its general events and lifetime periods—may, besides requiring a colossal expense of cognitive energy, be a recipe for depression and despondency, when no such overarching meaning or coherence can be found, or for self-

deception, when narrative themes and meaningfulness are imposed on a recalcitrant reality (see § 6.3.2).

Let me move on to Goldie's last element of narrative structure, the evaluative and emotional import of what is being related or narratively 'thought through'. In literary narrative, the evaluative import of an episode is often conveyed opaquely (in Lamarque's terminology) by the portrayal of a scene or its characters in a manner that is geared towards evoking certain emotional responses. In what Goldie calls 'narrative thinking', different devices are deployed in creating evaluative context, such as a particular mood being associated with a particular event. Nevertheless, as pointed out in my discussion of Schechtman's replies to Strawson (§ 6.1.1), evaluative thinking about one's past is perfectly possible in other, non-narrative ways. I may, for example, recall some juvenile transgression of mine: I accept responsibility for what I did then, I now consider it a foolish act, I am faintly amused (or disgusted) by it. There's the self-identification, there's the evaluation, there's the emotional response. No narrative devices are required.

There's no denying that the emotional import of episodes in one's life is indeed something that lends itself to narrative treatment. However, the presence of certain emotions during certain events and the affective connections one has to episodes in one's past are not narrative artefacts. Emotion, on empirically plausible theories,[3] precedes any cognitive appraisal of it, including narrative thinking. While narrative thinking can of course process, evoke, and evaluate emotions, it is very far from being their root cause. Thus, as far as the self is shaped by affective concern, such concern is antecedent to any narrative gloss we may put on it.

In sum, then, strong narrative constructionist accounts of the self (though in adding narrative cohesion they may perhaps represent a slight improvement over purely *memory*-based constructionist accounts) overstate their case in several ways. First, they ignore or sideline the pre-narrative elements involved in the self: primitive self-consciousness and its attendant 'mineness', temporal consciousness, and, as just discussed, affect and emotion. Second, they are overambitious in requiring our sense of self over time to have narrative coherence and meaning when sequential coherence may

---

[3] Such as Paul Ekman's (1992) account of basic emotions; or Jesse Prinz's (2004) somatic theory of emotions, and Barlassina and Newen's (2014) correction of it.

be all that is required. Third, they are unduly prescriptive in ruling out non-narrative means of self-identification and evaluation.

We should, therefore, reject the *strong* narrative constructionist claim that selves are narrative *in structure*. The self is not a story. But some of the criticisms just discussed affect *weak* as well as strong narrative constructionist accounts: even if we dismiss the requirement for selves to have a story-like structure, the first and third points just mentioned affect *any* account of the self as the product of narrativity, whatever its resulting structure may be. Emotion, temporal awareness, self-identification over time, and evaluation of one's past may be aided by, but do not *require*, narrative capacities. Indeed, as I shall now argue, the very idea that narrative processes precede the self is misguided, back-to-front, upside-down.

6.2 Author, author!

6.2.1 The authorship problem

I come now to what I take to be the most crucial objection to narrative constructionism about the self. All narrative constructionist accounts of the self are faced with what I will call the *authorship problem*: if the self is constructed through narrative, or a 'centre of narrative gravity', or in some other way a product of narrative activity, who or what then produces the narrative? Who is the author of the tale? It cannot be the self, for the self is only produced through narrative processes, a narrative construct or fiction. But narratives have authors. Taking heed of Lamarque's warnings about confusing real-life story-telling with literary narration, when I say 'author' here I do not mean anything like a *literary* author. I do not even mean that the authorship of our real-life tales need involve a conscious or deliberate intention to concoct a story. What I do mean is that there has to be some originating process or processes for a narrative to emerge. And whatever *produces* the narrative—its 'author' in a relatively weak sense—is *not the same thing* as something produced *by* the narrative. This being so, I will in this section make the case that it is in the author, not in a product of the narrative, that we should conceptually situate the self.

But let me pause briefly to consider a possible, indeed likely, reply to the authorship problem. I can imagine a chorus of narrative constructionists intoning, 'We never said there wasn't an author! The self is the author, *too*!' Wait—the self is the producer *as well as* the product of its narratives? Indeed, one of the various permutations of Schechtman's view discussed in the preceding section had the self as the character, author, and critic of our narratives, all at once. In a similar vein, the psychiatrist James Phillips (2003) maintains: 'The whole notion of a narrative self is that the self is indissolubly narrator *and* narrative.' (p. 316) On this view, we appear to be dealing with something that creates itself. This is clearly circular, perhaps viciously so—I'm inclined to think it is, but I won't insist upon the point, as many see nothing metaphysically amiss with a self-creating entity: thus, Dennett (1984) speaks of 'self-made selves', and Schechtman (1996) cites 'a number of sources both philosophical and psychological which argue . . . that persons are self-creating' (p. 93). But even granting the possibility of such self-creation, I submit that anyone who claims that the self is 'indissolubly' both author and narrative has not been looking hard enough. That claim is simply wrong. For we *can* distinguish the author from his/her/its narrative productions. A good illustration of this, it seems to me, is provided by Dennett himself.

In his characterization of the self as a 'center of narrative gravity', Dennett (1992) invites us to consider a thought experiment involving a novel-writing machine, which creates a fictional character called Gilbert, who is the first-person narrator of the novel the machine is producing. Dennett goes on:

> So far we've imagined the novel, *The Life and Times of Gilbert*, clanking out of a computer that is just a box, sitting in the corner of some lab. But now I want to change the story a little bit and suppose that the computer has . . . wheels. It has a television eye, and it moves around in the world. It also begins its tale with "Call me Gilbert," and tells a novel, but now . . . we discover that there's a truth-preserving interpretation of that text in the real world. The adventures of Gilbert, the fictional character, now bear a striking and presumably non-coincidental relationship to the adventures of this robot rolling around in the world. If you hit the robot with a baseball bat, very shortly thereafter the story of Gilbert includes his being hit with a baseball bat by somebody who looks like you. . . . At this point we will be unable to ignore the fact that the fictional career of the fictional Gilbert bears an interesting resemblance to the "career" of this mere robot

> moving through the world. We can still maintain that the robot's *brain*, the robot's computer, really knows nothing about the world; *it's* not a self. It's just a clanky computer. It doesn't know what it's doing. It doesn't even know that it's creating a fictional character. (The same is just as true of your brain; *it* doesn't know what it's doing either.) (p. 108)

Notice how Dennett distinguishes between 'Gilbert, the fictional character' and 'this robot rolling around the world'—its author. But then, Dennett's argument is that, in all the ways that are relevant to having a self, we are just like that robot: our brains spin a story about our experiences, centred on a character we call 'I', and this fictional character is the self.

Let me unpack just what is involved in the doings of this novel-writing robot.[4] The crucial move in the scenario occurs where Dennett 'twiddle[s] the knobs on this thought experiment' (ibid.) and gives a previously stationary machine mobility and camera vision. Now, to roll around the world without bumping into the furniture or toppling down the stairs, this robot presumably has a system that integrates the input from its camera with the circuits that control the movements of its wheels—a robot's version of sensorimotor integration. To register being hit by a baseball bat, it must have the robotic equivalent of nociceptors. This does not entail that the robot *feels pain* when hit (though its narrative may put it that way, perhaps: '"Ouch," I thought, "that hurt!"'). But it needs sensors to record the impact, and—crucially—pathways for that information to reach its central processing unit, so as to be available to its narrative output. Its camera input must likewise be connected to its novel-writing processor, or it could not write a truth-preserving story of what the robot comes across in the world.

Curiously, Dennett suggests that '[w]e can still maintain that the robot's *brain*, the robot's computer, really knows nothing about the world'. I do not wish to digress into an epistemological discursion on the nature of knowledge here, but then there is no need for that: if we take Dennett's own (1981) 'intentional stance' with respect to his novel-writing robot—which allows us to attribute beliefs and desires to a system if that strategy successfully predicts its behaviour—the robot's computer jolly well *does*

---

[4] In so doing, I am merely 'following directions' (Dennett, 1991/1993, pp. 398–401)—that is, making transparent the implications of a thought experiment as it is described, rather than yielding to thought-experimenter demand by latching on to the first intuition that presents itself.

know about the world, in the sense of *processing information about the world*. More important, it processes—and integrates—information about *the robot's interactions* with the world. And so the robot registers that *it* is being hit by a baseball bat. Aided by its camera, it can move around and locate *itself* in its environment—for example, it notices *it* is trapped in a closet. Nociception, sensorimotor integration, spatial orientation—or their robotic equivalents—are all basic *self-representational capacities* (as discussed in § 1.4). And the various bits of information provided by its different self-representational capacities must be *integrated* in the robot's computer, for they all form part of its story of Gilbert the roving robot. But it is the *actual* robot (or its computer), which has the capacities just described, that is the author of that story.

But wait. It may be said at this point that I haven't quite established that the author and the character of its narrative are distinct. Surely, the robot novelist and the character Gilbert in its autobiographical narrative are the same? Well, they are in so far as this robot is writing its own story. But they are not 'indissolubly' or indistinguishably the same. They don't share all of the same properties. The systems that feed into the authorial process are features of the robot rolling around the world, not of the character. (The character Gilbert presumably does not talk in terms of his nociceptors and sensorimotor integration.) And the robot, if it is any good at novel-writing, could be telling a different story from its own. In the first version of Dennett's thought experiment, where the novel-writing machine is stationary and has no camera, we are not asked to imagine that it produces *The Life and Times of Gilbert* with a box in the corner of a lab that experiences nothing for its protagonist. We are simply asked to imagine a machine capable of writing a novel with a fictional first-personal narrator called Gilbert.

To make a little clearer how the actual robot and its character Gilbert are distinguishable, let me, too, 'twiddle the knobs on this thought experiment'—and take *narrative* capacity out of the equation. Let us imagine a like robot on wheels with a camera eye, pottering about the lab, that is *not* also a novel-writing machine, but is otherwise endowed with all the same systems and capacities as the novel-writing version. This non-narrative robot—let's call it iBot 2.0—can still register the impact of being hit with a baseball bat and connect this occurrence with its camera image of someone wielding the bat. It can still integrate its camera input with the motoric op-

eration of its wheels. It can still exercise these self-representational capacities in moving about and orientating itself in its environment. With the exception of narrative capacity, iBot 2.0 has all the same systems and capacities as the novel-writing robot in Dennett's scenario. Obviously, since it can't write, it is not an *author*—but in all other respects it is a perfect replica of the author robot.

Dennett, I take it, would object here and insist that iBot 2.0 cannot have the same integrative capacities as his novel-writing robot precisely because it cannot narrate. The level of integration of different self-representational inputs displayed by the original robot happens only, Dennett would contend, *because* a narrative is being spun. (That, after all, is the gist of his position: that there is no single unified 'I' unless and until a story is being told about such a character.)

I am not so sure. To be clear: I do not suggest that either the novelist robot or iBot 2.0 has a *self*. But each of them probably has a *self-model* (Metzinger, 2007) or the robotic equivalent of a *proto-self* (Damasio, 1999; 2010), that is, a dynamic (non-conscious) representation of itself which integrates the available distinct self-representational inputs from its robotic equivalents of vision, motor control, nociception, and so on. Why would iBot 2.0 have such integration? Because that is what allows it to move about without falling down the stairs, to register that it is locked in a cupboard, or —as the case may be—to notice thought-experimenting philosophers thwacking it with a baseball bat.[5] If Dennett's novel-writing robot has these capacities, so has iBot 2.0. Narrative capacity is an optional extra.

This, of course, is an illustration of the point I already made in my first discussion of Dennett's view (§ 5.3.3)—that it is implausible to credit narrativity with the functional integration of what are (certainly in biological organisms like ourselves, but it seems also in well-designed mobile robots) basic self-representational capacities. Language—a prerequisite for narrative capacity—is a very recent development on an evolutionary scale (as Dennett well knows). The integration of different self-representational capacities, however, is not. It is indeed essential for an organism's survival that it should be able to integrate input from different sensory modalities and to evaluate dangers and obstacles, as well as advantages and opportunities, that the world presents to *it*, with reference to how they affect *its*

[5] As a new Pet Shop Boys song has it: 'Sad robot world.' (Tennant & Lowe, 2016)

own well-being and progress. Thus, when we, now equipped with language and the ability to narrate, tell a story of the things that happen to us in the world, this narrative capacity taps into a deeper, already well-established system of self-representational capacities.

Returning, then, to the question of distinguishing the author of a self narrative from the protagonist of its/her/his tale—there really *is* no question here. Author and protagonist clearly are distinct. In the first presentation of Dennett's thought experiment, we have an immobile, unwheeled, uneyed novel-writing machine, the author, producing a story about 'Gilbert', the protagonist and I-narrator who, however, is nothing like a computer in the corner of a lab, but just a fictional character. In Dennett's second presentation, the character 'Gilbert' seems to live all the exploits of its author—now a novel-writing robot on wheels and with a camera. Dennett's point is that 'Gilbert' is *still* a fictional character, despite reflecting the experiences of the novel-writing robot. Very well. But the robot that integrates its various inputs into the narrative is not a fictional character. It is the author of the tale. Finally, iBot 2.0 is just like that robotic author, except that it doesn't write—but it still has the integrative capacities that, in the previous case, subserved the robot's narration.

Two points to note emerge from this discussion. First, there is an author–character distinction which does apply in self narratives: Dennett's novelist robot is a fairly sophisticated machine that adds the ability to narrate to its other capacities, while 'Gilbert' is the character of its narrative. Secondly, while *The Life and Times of Gilbert* is that robot's *autobiographical* narrative, this does not entail the conclusion Dennett wants us to accept—that, in the analogous case where *we* construct our life narrative in a similar vein, it is the *product* of the narrative, rather than its *author*, which we should consider 'the self'. For, as the example of iBot 2.0 illustrates, the relevant self-representational capacities that enable something to integratively process its own experiences are *prior to* narrative capacity. So, narrativity is incidental, but the integration of self-representational functions is not. We should, therefore, situate the self in the system that narrates its own story, rather than in its narrative productions.

I will shortly offer another argument in support of this. Before that, I'll attempt to explain why the contrary view is so widespread.

6.2.2 A diagnosis and a refutation[6]

Narrative views currently seem to dominate the literature on 'the self', in psychology and psychiatry as well as in philosophy (where their exponents, besides those discussed here, include MacIntyre, Taylor, and Campbell; and, in another tradition, Heidegger, Sartre, and Ricœur[7]). Not all of them are narrative *constructionist* views, but, as we have seen, such views are espoused by Bruner, Schechtman (at least originally), and Dennett, among others. If, as I have just suggested, there are good reasons to take the self to be the author, rather than the product, of our narratives, then why does the notion that we, as selves, are narrative constructs or products seem so pertinaciously attractive to so many thinkers? Put another (if less charitable) way, how can otherwise highly competent philosophers and psychologists make such an elementary mistake as to confuse the *author* of some narrative with a narrative *construction*?

To begin our diagnosis, let's admit that the narrative strategy starts from a fairly uncontroversial insight: many or most of us (perhaps even Galen Strawson) habitually tell stories about ourselves. We do so among friends and family, but also among strangers: on first dates, at job interviews, over drinks and dinner at academic gatherings. 'What's *your* story?' is a conversation opener, but more than that, it is an invitation to reveal something of *oneself* (to the extent appropriate in a given social context, of course). As discussed in the previous chapter (§ 5.2), many of our social interactions involve some exchange of autobiographical narratives (Miller, 1994), and the development of autobiographical memory in childhood benefits from narrative practices and encouragement (Fivush, 1994; Nelson, 2003; Nelson & Fivush, 2004). Let us accept, then, that modern humans, by and large, are story-tellers.

But how do we get from this insight into our narrative nature to the view that the *self* is narratively *constructed*? Merely to say that we are, in some sense, 'essentially' narrative creatures is not enough: that would give us what I've called an *essential narrative-capacity view*, but not a narrative

---

[6] I thank George Botterill for his helpful suggestions here.

[7] Strawson, 2004; Schechtman, 2011b.

constructionist one. Why, in addition, would anyone locate the self in our narrative productions?

One may, perhaps, suppose that there is something of an empiricist hangover at work here. This is a somewhat tentative diagnosis, but it goes something like this. According to the Empiricists, we could only ever have a concept—an 'idea', in Hume's (1739) terms—of something if there was a corresponding object of sensory experience—in Hume's terms, an 'impression'. Now, the *self* is not an immediate object of experience (hence Hume's failure to find one amid his perceptions). So one might be led to think that the concept of a self, which does not correspond to any object of direct experience, must somehow denote a construction, or abstraction, from things that *are* direct objects of our experience: our self narratives. The self, thus construed, is not unlike the visual field: we don't *see* a visual field, only its contents; the notion of a visual field is an abstraction from the totality of whatever we *can* see. But of course, while talk of visual fields is useful in the right context, it does not do any explanatory work about vision—which also involves the eye (which we *also* don't see: cf. Wittgenstein, 1922, § 5.663), the optic tract, the visual cortical areas.

So, narrative constructionism seems to end up with something like the inversion of the classic 'homunculus' fallacy. This fallacy is to attempt to explain some capacity by postulating a second-order entity that has that same capacity. Take again the example of vision. I can see, but (let us suppose) I know nothing about the processes that enable this capacity; I just think it's something in my brain. So I posit a little man, a homunculus, with big eyes, sitting somewhere in my brain, who does the seeing for me. (And more homunculi for the other sensory modalities, for motor control, for decision-making, etc.) Now it is clear that this strategy hasn't *explained* anything about vision—it has merely displaced the problem: how is it that my visual homunculus can see? Unless we begin explaining vision by processes other than 'seeing', an infinite regress of homunculi beckons.

Now, confusing the self with its narrative productions is structurally rather similar to this fallacy. We have ample experience of our remembering, reminiscing, narrative thinking, telling stories of ourselves. We do not —as just mentioned—have direct experience of *a* self. So, we might say, the self just *is* something produced by (or 'constituted by') our memories and narratives. For it is when I catch myself remembering my past or telling

tales about my life that I seem closest to being aware of my *self*. But now we run into the authorship problem, because I haven't accounted for *who* is remembering or telling tales. Obviously, it is *I* who recollects and narrates. But I am supposed to be a *product* of my recollections and narratives. So we end up either with the circularity of self-creating selves or, worse, something like the homuncular regress: my telling a story about myself is a narrative product telling a narrative about a narrative product telling a narrative… Either way, and worse still, I have not, on this account, said anything at all about the processes underlying narrating or remembering (just as my seeing homunculus says nothing about the processes underlying vision).

Thus, in order to *explain* what the self is, we have to put things in the right order. In the same way as the self is the constructor, rather than the construct, of autobiographical memory, so it is the constructor, and not the construct, of its life narratives. (And of course such a self can be both rememberer *and* narrator. It does not have to be exclusively one or the other.) I have already presented an *empirical* argument for the priority of self-representational capacities over both remembering (§ 4.5) and narrating (§ 6.2.1), from which arises the conceptual recommendation that it is in these prior capacities that we should situate the self, rather than in their productions, our autobiographical memories and self narratives. Philosophers with a less empirical temperament than mine might wish for something stronger. Here, then, is an attempt at a *metaphysical* argument that the self is not a narrative production.

This argument adapts Saul Kripke's *modal argument* against descriptivism about names. Kripke (1972/1981) famously argues that names are not definite descriptions, but 'rigid designators' (p. 48), because of how they behave in modal contexts. What a name designates does not change in counterfactual statements, whereas the referents of definite descriptions do. For example: Hume is the author of *A Treatise of Human Nature*. Now, 'the author of the *Treatise*' is a definite description which happens to designate Hume. But the *Treatise* might have been written by someone else, in which case the definite description 'the author of the *Treatise*' would designate *that* author, and not Hume. But 'Hume', as a name, always designates Hume: Hume might not have written the *Treatise*—he might have written *Les Sincères*[8] instead—but 'Hume' would still designate Hume (as would, in

[8] A comedy by Marivaux, premiered in 1739, then quickly forgotten (Greene, 1965).

this particular counterfactual case, the definite description 'the author of *Les Sincères*').

Something analogous seems to apply to selves and narratives. We can place the *self* in modal contexts without changing its referent, i.e. what is designated by the first-person singular pronoun. As Goldie (2012) notes, we can and routinely do put ourselves in counterfactual scenarios in every-day 'narrative thinking', such as, in his example, 'If only I hadn't dawdled, I wouldn't have missed the train' (p. 15).[9] Similarly, I might (in a fairly close possible world) have decided to pursue postgraduate studies in classics instead of philosophy. That putative, counterfactual classics postgraduate would then have been *me*—or, if you like, he would have had my *self*.

Narratives, however, do not behave in that way in modal contexts. Keeping with the example of my choice of postgraduate course, the factual (mini-)narrative here is: 'Seven years ago, Philipp went to Sheffield to do an MA in philosophy'. Now, in the counterfactual case, we end up with a *different narrative*, e.g. 'Seven years ago, Philipp went to Manchester to do an MA in classics'. Like definite descriptions, narratives are not constant across possible worlds—whereas the referent of 'I' or 'the self' is. And so, just as names cannot be definite descriptions, selves cannot be narratives.

Narrative constructionists could not disagree with my second premiss here: that narratives are not constant across different modal contexts. But they would dispute the first: that selves are. The whole point of their account, they'd say, is that the self *is* a narrative production, so a counterfactual scenario would naturally result in a *different self*—'constituted' by a different life story, or a 'centre of narrative gravity' located elsewhere, or whatever. Now I agree that, had I gone to Manchester to study classics, my *life* would, from that point onwards, have been different. I would have had different teachers, different friends, different books on my shelves. I might even, in some loose and metaphorical sense, have become a different person—one who talks freely of (or in!) the aorist optative and the genitive absolute, rather than of counterfactuals and possible worlds. But that person

[9] Whether we always *should* do so is a different matter: see § 6.3.2.

would have been *me*, that life would have been *my* life, the life of the same self that, in the actual world, didn't go on to study classics.[10]

At this point, the narrative constructionists and I could well be said to be at cross-purposes: we appear to mean something different by 'the self'. Indeed we do. But what I mean by 'the self' here is simply—*me*. In this argument, 'the self' has the referent of the first-person singular pronoun, that is, the *ordinary* meaning of 'self'. If we are asked to abandon this and adopt a different meaning of 'self', an historical-constructionist conception whereby the self is a narrative construct and has a temporal dimension, the onus is on the constructionists to make the case for that. And this, as I think I have demonstrated in the above discussion of narrative structure (§ 6.1) and of Dennett's narrative robots (§ 6.2.1), they have failed to do.

The self, in the ordinary sense, behaves like a Kripkean 'rigid designator' and so cannot be a narrative. It may have a tremendous narrative *capacity* (which, as seen here, is indeed useful in counterfactual thinking), but that does not make it a narrative *construct*. In our narrations, the self is the author, not the tale.

6.3 Are we essentially narrators?

We can now dismiss narrative *constructionism* about the self, both in its strong and its weak variant: the self is not narrative in structure or a narrative product. But we might still think that our narrative capacity is in some way essential to who we are. This is the third of the four positions I distinguished in the previous chapter (§ 5.3.2): the *essential narrative-capacity view*.

---

[10] Indeed, when we engage in counterfactual narrative thinking about ourselves, it is usually because such thinking is informative about us as we *actually* are. Take Goldie's example: 'If only I hadn't dawdled, I wouldn't have missed the train.' Having had this thought, one may well conclude that, next time, it may prove useful either not to dawdle, or to get up early enough to allow time for dawdling.

David Velleman's account of the self as *narrator* (§ 5.3.3) emphasizes the authorial role of the self in our narrative activities.[11] As I have just argued that the self is the author of our every-day stories, I obviously do not disagree with that. But Velleman (2005) also holds that our narrative activities give us 'agential control' over our doings: '[my] inner narrator is a locus of control that unifies [me] as an agent by making decisions on the basis of reasons' (p. 71). In this way, 'the process of self-narration shapes our day-to-day lives' (p. 73).

There are three claims here that need to be addressed. One is that it is only by narrating that we become unified agents with reasons for action—which suggests that, if we want to be the sort of being who acts for reasons, narrative capacity, and its deployment, are *essential*. Though I cannot venture to give a full account of *agency* here, there may be something to be said for the view that narrative practices help us understand ourselves and each other in terms of agents acting for reasons (see § 6.4). But we should separate accounting for agency in this sense from giving an account of the *self*. Recalling Galen Strawson's point that we are not all equally narrative in temperament (§ 6.1.1) and my earlier observation that many of our higher-level self-representational activities, including 'affective concern', do not require a narrative structure (§ 6.1.3), it would seem reckless to make narrative agency an essential component of the self. Furthermore, it would mean depriving of a self anyone with limited or no narrative capacity, despite their other self-representational capacities—in particular, people with autism spectrum conditions (which I discuss in the next chapter).

Velleman's second claim is that self narratives give us 'agential *control*'. In § 6.3.1, I will argue that such control is largely illusory, and point out that there are dangers inherent in an over-reliance on our self narratives. Velleman's third claim is that our lives are 'shaped' by our self narratives, which echoes Schechtman's (1996) claim that 'the lives of persons are narrative in form' (p. 93). In § 6.3.2, I will argue that, taken as a *descriptive* claim, this

---

[11] In literature, there is, of course, an important distinction between the *narrator* of a story, who is internal to the narrative and part of its artifice, and its *author*, who is not. But in the present context, that distinction is not relevant. It has already been observed that literary narratives differ from real-life stories in their level of artificiality (§ 6.1.2). And I have for present purposes defined *author* rather weakly as a producer of narratives (§ 6.2.1). Similarly, I will now take *narrator* to mean simply someone who narrates.

may be true of some people, but not others. And, taken as a *prescriptive* claim, I will note that it is dangerous.[12]

6.3.1 Ignorance, inaccuracy, and interpretation

Some mention has already been made of the phenomenon of *confabulation* associated with frontal-lobe damage, where the loss of executive function leads patients to tell completely fabricated stories and alleged 'memories' of themselves (§§ 4.4.4, 5.2.2). A similar phenomenon occurs in commissurotomy ('split-brain') patients in specific experimental conditions. Commissurotomy is the resection of the corpus callosum, the main nerve bundle linking the left and right hemispheres of the brain, performed as a (now outdated) treatment for severe epilepsy. Michael Gazzaniga (1995) reports on an experiment where instructions were delivered to split-brain patients in such a way that they would be perceived and processed only by their right cerebral hemisphere (e.g. the written instruction 'Walk!' being flashed up in the patient's left visual field, the left visual field being processed by the right cerebral hemisphere). The participants then followed these instructions, but when asked about the reasons for their behaviour, their responses—generated in the left cerebral hemisphere, where the language areas are situated—made no reference to the experimental instructions, since these data had not crossed hemispheres; the reasons participants gave were confabulated (e.g. 'I needed to stretch my legs').[13]

From this work with split-brain patients, Gazzaniga (2000) hypothesizes that the left cerebral hemisphere has an 'interpreter' mechanism that integrates monitoring our responses with available information about inputs the brain has received and processed. The 'interpreter' (which is based in cortical areas associated with language production) then produces a reason-giving narrative when prompted. But it does so even when *no data*

---

[12] Strictly, my case here is against the *descriptive* essential narrative-capacity claim. But the two claims are often interwoven in narrative accounts of the self; as Galen Strawson (2004) notes, the 'dominant view in the academy today' is that 'all normal non-pathological human beings are naturally Narrative and also that Narrativity is crucial to a good life' (p. 429). Thus, a few words on the dangers of the *prescriptive* claim also seem appropriate.

[13] I discuss split-brain cases in more detail in § 8.3.1.

are available about the actual reasons for a certain behaviour, as in the case of the split-brain patients' confabulations (whose left cerebral hemisphere receives no input from the right hemisphere). This raises an interesting and somewhat unsettling hypothesis: if the 'interpreter' merrily spins a plausible (but false) explanatory tale in the absence of any explanatory data available to it in such circumstances, might it not do the same even in a healthy brain? Could our narrative outpourings, whether oral or in 'inner speech', *all* be confabulations?

There is a considerable amount of empirical evidence to support the hypothesis that we do not ordinarily have access to the cognitive processes that inform our decisions and behaviour and that, consequently, any verbal reports we give of our reasons cannot be based on what actual cognitive processes prompted the judgement or action. Instead, as Nisbett and Wilson (1977) suggest in their seminal paper 'Telling more than we can know', our verbal reports 'are based on a priori, implicit causal theories, or judgments about the extent to which a particular stimulus is a plausible cause of a given response' (p. 231). Absent any conscious awareness of a specific stimulus, the need to stretch one's legs is a perfectly plausible reason for unexpectedly getting up from one's seat and walking about, even if, as in the split-brain example, there is a different but unobserved (by the left cerebral hemisphere) stimulus that prompts this response. Thus, what we take to be reports on our judgements and decisions are *post hoc* interpretations (Carruthers, 2010).

John Doris (2002) has argued that the plausible causes we cite as reasons for our behaviour, particularly when they refer to our presumed character traits, frequently fail to account for situational factors that may have primed our behaviour on a given occasion. A person's helpfulness towards another may owe more to a minor piece of good fortune just encountered, the fine weather, or a good lunch, than to a generally benevolent personality. However, it is possible to overstate the influence of situational factors, particularly if they result from specially concocted experimental designs or extraordinary circumstances (Scaife, 2011). And, as Nisbett and Wilson (1977) themselves make clear, there are cases where our verbal reports do pick out the relevant stimuli that prompt our responses: 'reports will be accurate when influential stimuli are (a) available and (b) plausible causes of

the response, and when (c) few or no plausible but noninfluential factors are available' (p. 253).

What are the implications of this for the self-narrator? For one thing, it seems that the reasons we give for action in our self narratives are only sometimes accurate reflections of the causes of our behaviour, often merely plausible interpretations of it, and, at worst, confabulations. But, according to Velleman (2005), '[t]he self-narrating agent is a bit like an improvisational actor, enacting a role that he invents as he goes' (p. 71). So, not only are we supposed to be narrating our reasons for *past* actions, but also narratively formulating reasons for *future* behaviour, which we then proceed to enact. But, as John Bickle (2003) notes:

> The measurable brain activity [in left-hemisphere language areas] hypothesized . . . as . . . expressing our (narrative) selves does not exert significant control over neural regions that subserve cognition and behavior . . . (p. 196)

And,

> given the limited access enjoyed by the language regions to neural networks that subserve specific cognitive and behavioral tasks, these narratives are actually outright fabrications, as is the self-in-control they create and express. (p. 201)

With respect to Velleman's notion of the narrative self as a 'locus of control', Bickle's point is well made. So is the observation that, whatever narratives we spin, they are made without introspective knowledge of the cognitive processes that really subserve our behaviour and decisions. The problem, however, is not, as Bickle suggests, that *all* our verbal reports, all our self narratives, are 'outright fabrications', but that—lacking introspective access to relevant cognitive processes—we *cannot know* which of our narrative reports are accurate and which are not. Self narratives may give us the illusion of being in control of our reasons and actions, but they are not what produces our behaviour, nor do they necessarily reflect its causes.

Some narrativists might be happy to bite that bullet. They might point out that most ordinary self narratives contain enough plausible and accurate components to be a legitimate expression of the self, whereas, as Schechtman (1996) suggests, completely delusional self narratives can be rejected outright, and partial delusions will result in defective narratives to

the extent of the delusion. While this seems fair, it flags up the need for external, non-narrative correctives. For if one is delusional, one cannot oneself judge one's delusional self narrative to be flawed. That correction needs to come from outside the person concerned, and from outside her narrative.

More trivial errors of fact in self narratives are more easily corrected. As Schechtman puts it, they differ from delusional errors in 'the typical willingness of people to revise their narratives when presented with evidence of these small errors' (p. 123). This mirrors somewhat what I said about 'benign' misrememberings and embellished memories (§§ 4.4.2–3). But again we should note that the corrective provided here is external (and, for all we know, may itself be misremembered).

Thus, self narratives need to be open to external corrections. And what if no such correctives are forthcoming? Self narratives might be said to have this advantage over autobiographical memory on its own: they can accommodate a modest number of factual inaccuracies so long as the general thrust of the narrative is not violently at odds with the facts. On the other hand, there is also a danger inherent in narrative more serious than that of occasional false memories: while one false recollection does not invalidate all memories, a self narrative built around a single false memory—like the alleged 'recovered memories' of child abuse discussed earlier (§ 4.4.5)—will thereby be wholly false, because its very foundation is an untruth. The only corrective mechanism available in such circumstances is to scrap the whole narrative and start anew.

It seems dangerous, therefore, to place too much emphasis on one's own narrative productions. Notwithstanding this danger, narrativists tend to show a certain disregard for factual accuracy:

> To some degree, and for the sake of creating a coherency to life, it is normal to confabulate and to enhance one's story. As Ricœur [1990] points out, narrative identity 'must be seen as an unstable mixture of fabulation and actual experience'. Self-deception is not unusual; false memories are frequent. (Gallagher, 2007, p. 211)

Narrative coherence, it would seem, is taken to be more important than factual accuracy. Mark Freeman (2003) makes a similar point when he contrasts '*narrative* truth' (a term he borrows from Spence) with *historical* truth. The only way to make sense of this, without stretching one's logical understanding too far by stipulating a multitude of truths, is to note again the

interpretative nature of self narratives. The criterion for 'narrative truth' then seems to be whether the self narrative provides a useful and meaningful interpretation of the events of one's life.

But what is useful and meaningful here? In Schechtman's (1996) view, considerable leeway is allowed in what interpretative stance one might take in constructing one's self narratives:

> The depressed person, the suspicious person, the optimistic person, the angry person might tell quite different narratives of the same events, but they are all comprehensible and in most cases it seems misguided to argue about which is the most accurate. (p. 128)

I'm not sure about that. Placing a particular narrative interpretation on one's life may sometimes be helpful and motivating. Some kind of story to the effect that 'it wasn't meant to be' may perhaps help one overcome the disappointment of a failed relationship, for example. But in other cases—particularly that of 'the depressed person'—an overindulgence in narrative interpretation seems a recipe for misery. Self narratives of a particular kind can be extremely noxious. I will return to this danger presently, as I now discuss whether we do (and should) shape our lives narratively.

6.3.2 Discontinuities, 'turning points', and noxious narratives

Defenders of narrative accounts of the self often refer to the coherence a self narrative provides to what may otherwise seem disparate and discontinuous experiences in one's life. Up to a point, this may be a useful cognitive strategy and serve to organize one's autobiographical knowledge base into 'general events' and 'lifetime periods', as Conway and Rubin (1993) suggest (see § 4.3.1). But there are discontinuities in most lives that aren't smoothly papered over by a continuous narrative. People change careers and social circles, move to different cities and countries, end old and form new personal relationships, take up new habits and hobbies and abandon old ones; sometimes all of these at once. The reasons for such discontinuities are frequently complex. Changes in professional and personal circumstances may depend on chance and on the other people involved. And whatever 'life-changing' decisions one may attribute to oneself in such contexts must be taken with a pinch of salt, given the lack of introspective access we have to our judgements and decisions.

Thus, it seems reasonable simply to accept discontinuities in our lives for what they are. That would be the strategy adopted by someone with an 'Episodic', non-narrative temperament like Galen Strawson—but it can also be accommodated by Schechtman's (2007) revised view, in which even a narrative self need not span a whole life. But it seems that some do sense an urge to explicate autobiographical discontinuities in narrative tropes, in what Bruner (1994) calls 'these idiosyncratic turning points that people so persistently report':

> They are vividly particular, even though they carry some affective or moral message with them. Take some instances. The football coach tells the high-school kid to 'get' the opposing end, get him out of the game. The kid turns in his uniform the next day, revolted by 'winner' morality. He gives up the adolescent ideal of being an athlete, and becomes a 'brain' instead . . . and is now an academic.[14] . . . In another turning point, a woman tells of living with an alcoholic for eight years. He comes home at all hours, threatening to beat her and sometimes carrying out the threat—but 'once too often: I remember that night; that was enough.' She leaves him.
>
> Note first that these turning points, though they may be linked to things happening 'outside', are finally attributed to a happening 'inside'—a new belief, new courage, moral disgust, 'having had enough'. They are thickly agentive. Secondly, they ride into the story on a wave of episodic memory retrieval, rich in detail and color. . . .
>
> . . . They are prototype narrative episodes whose construction results in increasing the realism and drama of the Self. In that sense, the narrative construction, whenever it actually happened, is as important as what is reported to have actually happened in the turning point episode. Turning points, in a word, construct emblems of narrative clarity in the teller's history of Self. Narrative, we know, imposes a particular structure on the 'reality' that it depicts. . . . (pp. 49–50)

Bruner's last remark is telling: these turning-point stories are not factual recollections, nor is that their function. The high-school football player must have had some aptitude for 'brain' work as well as (if not more than)

---

[14] In his subsequent discussion of this case, Bruner again evokes the literary form of the *Bildungsroman* (p. 51). And though Bruner's Goethean comparisons may be overstretched, his repeated recourse to the *Bildungsroman* form when discussing the life narratives of young American males suggests that they do have certain thematic features in common, rooted not in their individual lives, but in a shared 'cultural consciousness' (Nelson, 2003, p. 24)—which in these cases perhaps owes more to Hollywood coming-of-age movies than to *Sturm und Drang*.

for athletic activities before he handed in his kit. The turning-point story is a *post hoc* creation, an interpretation, indeed an 'idealization' (p. 51).

While the turning-point episodes Bruner cites are perhaps useful to their tellers in making sense of their lives, they are also a fine illustration of how the reconstruction of autobiographical memories is affected by the tales we tell of ourselves. Does the woman leaving her alcoholic and abusive partner really 'remember *that* night', or is her episodic memory of 'that night' an amalgamation of many distinct episodes leading up to her leaving him? We cannot know. What seems to matter here is how *she* remembers (tells) it. But, as Bruner admits, the narrative construction of a single turning-point episode is a dramatic idealization. The problem, however, is that there is no objective criterion for deciding when such idealization becomes self-deception.

Narrativists might argue that some amount of self-deception is a risk we must take in exchange for the coherence and 'meaning' a narrative self provides. That is to assume that changes and discontinuities in our lives necessitate the deployment of narrative tropes for us to make sense of them. But it is far from obvious that the only way to incorporate discontinuities—including significant, career-changing events—in one's sense of self over time is to paint over them with dramatic gloss.

For instance, I am occasionally (if mercifully rarely) asked what made me begin studying philosophy at the age of thirty, after spending the previous decade drifting from one kind of occupation, city, and country to another. Questioners perhaps expect some sort of turning-point story of the kind Bruner describes. But I have none. Though in retrospect my academic studies may with some justification be described as 'life-changing', I have no particular vivid episodic memory to mark their onset, no dramatic incident, and certainly no 'affective or moral message'. I can give a prosaic description of my circumstances at the time, theorize about what factors may have contributed to the change, and, if pressed, tell some kind of story about them—but I have no ready-made narrative up my sleeve, nor do I see why I should have one. At the same time, I happily accept both what happened before and after the change as parts of my biography: for all its *narrative* discontinuities, it is the life of a continuing self.

Recalling Strawson's notion of different 'temporal temperaments', we may perhaps say that narrativity about the self may be helpful or natural

for some, impossible or unnecessary for others. But I must go a little further and note that, for some and in some circumstances, self-narration can also be outright harmful. There are, I suggest, both forward-looking and backward-looking *noxious* self narratives (which may co-occur). They involve an increasing divergence of one's narratively constructed 'reality' from actual reality, either when the events of one's life fail to accord with a prepared (forward-looking) autobiographical narrative that is too limited in scope and thus under-equipped to deal with unexpected actualities, or when a (backward-looking) self narrative constructs one's past in a way that imposes unrealistic limits on present and future.

As an example of a forward-looking noxious narrative, let us assume that Bruner's high-school footballer does not have an epiphanic moment on the playing-field. Instead, he sticks to his idea of an athletic career. He constructs a self narrative in which he becomes first a college footballer, then a professional league footballer. He neglects his academic study in pursuit of making that story come true. Almost inevitably (I'm assuming that the statistical likelihood of an American youth becoming a professional American footballer is about as low as that of a European youth becoming a professional association footballer), he fails to get admitted to his chosen prestigious college, bides his time playing for a minor college, fails to get a degree, and ends up neither a professional footballer nor in any academic career that might have been open to him had he not followed his idealized self narrative.

Narrativists will be quick to point out that such cases, where they occur, are not paradigmatic of normal self-narration, in which the self narrative is revised in the light of the events of one's life. The point is, however, that it isn't obvious at what stage one should reasonably correct or abandon a forward-looking self story. Should the young man revise his narrative as soon as he gets rejected by his chosen college, or should he carry on 'believing in himself' (that is, in his preferred narrative version of himself), letting up other opportunities, and, if so, for how long? Disappointment ensues either way: by having to give up a beautiful story before giving it every chance to be played out, or by finding it more and more at odds with actuality.

The corollary to this scenario is a particular type of backward-looking self narrative that has the effect of unrealistically limiting one's prospects

for the future. This is the narrative *rumination* that often occurs in clinical depression. Take the (fictitious) case of Bernard, a man in his mid-forties, highly intelligent and articulate, well-educated, well-read and well-qualified. He is also chronically depressed, alcohol-dependent, and unemployed (these last three characteristics forming a causal chain, in the order named, but then looping back to the start). Bernard has a well-rehearsed, indeed over-rehearsed, self narrative involving a number of negative turning points going back nearly three decades, each of which 'explains' a particular disappointment, a particular instance when things went wrong for him, and each of which is in turn narratively construed as the inevitable consequence of the preceding one.

Strikingly, all these incidents, which Bernard recounts in ruminating detail, involve the agency of others (his parents, for instance) that is in some way blamed for the infelicitous outcome. Exemplifying what Bruner (1994) calls 'victimicy', it is never Bernard's own agency that is invoked. This denial of agential selfhood in the past also affects Bernard's sense of agency in the present: any course of action that might be taken to alter his situation can be ruled too difficult or impossible because of things that happened in the past, which happened because of what happened in the further past, and so on. If only such-and-such had been different in the past, the present too would be different, and Bernard would be better off. But such-and-such would only have been different if previous matters too had been different, which they weren't… Being clinically depressed, Bernard cannot rid himself of this noxious narrative. He cannot, for example, just start a new story. And in his persistent narrative, the only way to a better future is to change his past, which of course he cannot. Goldie's (2012) counterfactual 'narrative thinking about one's past'—while sometimes useful, in small doses—here leads to a cycle of narrative misery.

As Bernard's noxious self narrative clearly correlates with depressive illness, narrativists here have recourse to what one might call the 'pathology defence': pathological cases do not invalidate healthy cases; rather, their abnormalities illustrate the features of a normal narrative self. Bernard's example, they might say, shows that a sense of agency is required for a healthy narrative self; the dysfunctional narrative is a symptom of a defective self. They might also suggest how to repair it: the therapy required would involve changing Bernard's self narrative. Perhaps so, but then it

isn't clear which way the aetiology goes. If depression is the cause of the defective self narrative, then correcting the narrative may not get rid of the depression (or may not even be possible). Or if narrative therapy can help overcome or contain the depression, we must wonder whether having a defective self narrative is responsible for the depression in the first place, or at least for its persistence. And then it is the narrative that is doing unnecessary harm.

Thus, my point about noxious narratives, forward- *or* backward-looking, is not that the existence of harmful self narratives disproves the possibility of helpful self narratives. It is rather that, in such circumstances, *not* having a self narrative *of any kind* would be preferable to having a noxious one. Of course, it may be that holders of noxious self narratives have no choice about how narratively inclined they are, if Strawson is right about our different temporal temperaments. Yet, if our temporal temperaments are subject to change, it would also be wrong to hold up narrativity as the gold standard in what a self should be.

In sum, we need not necessarily appeal to narrativity when it comes to making sense of our lives. Discontinuities in biographies are common; change, even drastic change, is a part of a human life. One needn't always have a story to make sense of it. Thus, *descriptively* speaking, narrative activity is not essential for a sense of self over time. It would be wrong to insist that all selves are essentially narrative. And, as noxious narratives show, the wrong kind of autobiographical narrative can be seriously self-limiting. So neither should we endorse the *prescriptive* claim that our lives should be given narrative form.

6.4 A role for narrative practices

I have argued, in the first two sections of this chapter, against narrative *constructionism* about the self and, in the preceding section, against the view that narration is an *essential* activity of the self. Of the four positions one could take regarding the relation between narrative and the self, that leaves only the fourth: the *simple narrative-capacity* view—the view that selves can and often (but not necessarily) do engage in narrative activity. This is my view. For it seems that we are something of a story-telling species, and be-

sides telling stories about ourselves to others, many of us engage in what Goldie calls 'narrative thinking' about ourselves from time to time (though, in the light of my discussion of noxious narratives, we would do well not to overdo this). So let me outline, briefly, the (non-exclusive) role of narrative practices in our self-representations.

First, narrative practices can play a role in forming and reconstructing autobiographical memory. The conversational narrative practices reported by Fivush with which I began the previous chapter are an example of this, suggesting (along with Miller's research) that the onset of autobiographical memory is helped along by the onset of rudimentary narrative capacities. But there is also a role for narrativity in rehearsing and reconstructing episodic memories in adult life. The narrative 'thinking through' of an important episode is the kind of elaborative rehearsal that, according to Brown and Craik (2000), helps consolidate episodic memories (see § 4.3.1). Narrative practices also help both with the construction of our autobiographical knowledge base with its 'general events' (Conway & Rubin, 1993) and with the reconstructive remembering of these general events (Goldie, 2012). A narrative context will shape *how* an episodic recollection is reconstructed. We should note, however, that far from increasing the reliability of memories, their narrative reconstruction will obey the dictates of a particular social context or cultural trope, as in the turning-point stories Bruner discusses. Here the function of narrative practice is indeed the construction of meaning that narrativists like to emphasize—with all the dangers of interpretative largesse that entails.

Besides aiding autobiographical memory, narrative practices contribute to self-representation in other ways. Dan Hutto's (2007b) 'narrative practice hypothesis' suggests that children learn how to use and apply the concepts and tenets of folk psychology—desires and beliefs, and the way they provide us with socially acceptable reasons for action—through narrative practices: our every-day story-telling, which involves giving reasons for actions, has the function of explaining our responses to others and to

ourselves alike.[15] I said earlier that the conjunction 'because' does not make a narrative (§ 6.1.1). Nor does it: but the explanatory force of our 'becauses' is increased when they are embedded in a narrative that pads out the explanation of an individual action with additional information concerning its agent's background and/or her circumstances. Where that agent is the self, these narratives thus contribute to one's self-representation *as* an agent.[16] And, as Valerie Hardcastle (2008) puts it, our self narratives 'tie us to our social communities' (p. 48). Miller's (1994) studies of every-day story-telling illustrate this nicely.

A more controversial assessment of the role of narrativity is the theory put forward by Nelson (2003) discussed above (§ 5.2.1) that the emergence of narrative capacities around age five provides us with the 'level of consciousness' required for a full sense of self. Responding to Zahavi's (2007) point on the priority of non-narrative self-awareness over narrativity (§ 6.1.3), Schechtman (2011b) cites Nelson's theory, suggesting that our (post-narrative) self-consciousness differs qualitatively from 'brute first-personal awareness' (p. 410). But if it does, it isn't clear whether this richer kind of self-consciousness is *caused* by developing narrative capacities, merely *correlates* with their emergence, or is in turn *causing* the development of narrative capacity. On this point, Hardcastle (2008) suggests that, contrary to the 'dominant research paradigm in developmental psychology', it isn't narrativity that enables self-consciousness in the relevant sense; rather, 'the drive for selfhood pushes us along in our linguistic, cognitive, and mnemonic development instead of the other way around' (p. 52). I am not quite sure what she means by 'the drive for selfhood', but if her point is that a social demand for overt self-representation encourages the development of children's mnemonic and narrative capacities—that does seem a hypothesis worthy of investigation by developmental psychologists.

---

[15] Hutto also argues—more controversially—that it is through narrative practice that children *acquire* folk psychological categories, rather than by the activation of a theory-of-mind (ToM) module around the age of four. I do not think that the stimuli provided by narrative practices are sufficient to support that conclusion, but the point is not germane to my current concerns. (I'll discuss ToM in relation to autism in ch. 7.)

[16] This seems to be the motivation for Velleman's account.

Overall, then, narrative practices do play a role in our higher-level self-representations, both in consolidating and reconstructing autobiographical memories, and more overtly in fostering reason-giving discourse that contributes to the way we construe ourselves as agents in a social context. However, it is also worth repeating that, in childhood or in adulthood, we do not all engage in narrative practices to the same extent, nor deploy them evenly throughout life. Selves can be more or less narrative between individuals and over time within individuals. Self-representation is far from exclusively narrative.

To illustrate this, the next chapter will examine the self in autism spectrum conditions, which are often marked by a lack of narrative capacity, as well as atypical operations of autobiographical memory.

Chapter Seven

Autism spectrum conditions

7.1 Why autism?

There are several connected reasons why autism is relevant to accounts—and particularly this account—of the self. The first of these is the peculiar *self-absorbedness* of autistic individuals—a corollary of their social and communicative impairments. The earliest clinical descriptions of autistic children by Leo Kanner (1943) make frequent reference to the child's being 'self-sufficient', 'self-satisfied', or 'retire[d] within herself'.[1] Kanner repeatedly refers to autistic children's 'aloneness' as a defining feature of their condition; several of his case studies explicitly mention the difficulties experienced by parents and clinical examiners in attracting the children's attention. But their relative lack of interest in other people is often accompanied by an unusual interest in inanimate objects. Thus, there emerges a picture of individuals whose self–other and self–world relations are configured in a significantly different way from those of non-autistic ('neurotypical'[2]) individuals.

Secondly, while autistics seem excessively self-centred and self-absorbed, this does not entail an over-active self—indeed, autistic individuals seem to *lack* some typical higher-level self-representational capacities. These include the ability to ascribe mental states to oneself. As will be discussed below (§ 7.2.2), the—to date—clinically and diagnostically most successful neurocognitive theory of the symptoms of autism is that autistics lack a *theory of mind* and thus the ability to 'mentalize', i.e. ascribe mental

[1] The coinage 'autism'—from αὐτός 'self-, oneself'—obviously derives from this characteristic trait of autistics.

[2] The term recommended by people with autism to describe those without autism, avoiding the potentially offensive 'normal' (National Autistic Society, 2015).

states to others—resulting in the detached and aloof attitude to others that is characteristic of autism. Frith and Happé (1999) have argued that such lack of mentalizing ability regarding others would entail a concomitant inability to ascribe mental states to oneself and, consequently, a qualitatively quite different form of self-awareness. This, however, does not entail that the autistic person does not have a self (§ 7.3.1).

The past two decades of clinical research on autism have produced a large body of neurophysiological data on brain processes involved in mentalizing and self-awareness. It seems fair to say that, absent the clinical needs of autistic patients (and their families), far fewer neuroimaging studies would have been conducted from which we can learn about the functioning of both autistic and non-autistic brains *and* about how brain function correlates with social interaction.

Thirdly, the self-representational capacities of remembering and narrating about the self discussed in the preceding chapters show marked peculiarities in autistic individuals (§§ 7.3.2, 7.3.3). Narrative capacity may be entirely absent, as may self-referencing episodic memory. According to historical-constructionist accounts of the self, this would disqualify many individuals with autism spectrum conditions from having a self at all. On my system view, however, no such drastic consequence follows from the lack or impairment of these specific higher-level self-representational capacities. Other self-representational processes are intact in autistic individuals, and their mnemonic and narrative idiosyncrasies may be characterized as an 'alternative cognitive style' (Happé, 1991, p. 212).

Overall, for the Jamesian project of giving an account of the self by '[g]etting clear about the empirical relations between one's experience, one's brain, and one's social relations' (Barresi & Martin, 2011, p. 49), autism provides a valuable case study, precisely because the brains and social relations of autistics (and presumably their experience of themselves and the world) are so unusual. But before discussing autistic selves (§ 7.3), I need to give a brief account of the characteristics of autism spectrum conditions and the theories that have been put forward for their explanation (§ 7.2).

7.2 Characteristics and theories of autism

7.2.1 Clinical characteristics of autism

Autism is a neurodevelopmental disorder characterized by specific 'impairments in socialization, communication, and imagination' (Frith, 2001a). Though its precise aetiology is still unknown, a 'genetic basis . . . is strongly indicated from twin and family studies' (ibid.).[3] The characteristic signs and symptoms of autism, generally grouped into a 'triad' of impairments, usually become apparent between the ages of eighteen months and three years. They are:

1. *Abnormalities of social development.* Autistics fail to engage in normal social interaction, appearing aloof and self-absorbed. A clinical marker of this characteristic is their failure to monitor another's gaze and to engage in shared-attention behaviour.
2. *Abnormalities in communication*, both verbal and non-verbal. Autistic children show difficulties and significant delays in acquiring and comprehending speech. An early, non-verbal sign of communication deficit is their failure to engage in protodeclarative pointing (i.e. pointing with the intent to draw someone's attention to something). Echolalia is also observed. Where language skills approaching a normal level of competence are developed, lexical and pragmatic abnormalities persist, including difficulties with the use of personal pronouns, as well as an overly literal understanding of others' utterances.
3. A *restricted repertoire of activities and interests* and a *deficit of imagination*. These non-social characteristics of autism manifest themselves in stereotyped and repetitive behaviour—a 'desire for . . . sameness' (Kanner 1943, p. 249) in one's activities. Failure to engage in spontaneous pretend play is an early clinical marker of the imagination deficit in autistic individuals.

(Baron-Cohen, 1995, ch. 5; Bauman, 1999; Botterill & Carruthers, 1999, ch. 4; Frith, 2001a; World Health Organization, 1992/2016, F84.0).

---

[3] Thus, the description of autism as *developmental* refers to how the disorder manifests itself (i.e. in unusual psychosocial development) rather than to its being *caused* by developmental factors—though environmental factors can *modify* behaviour in autistic children (Frith, 2001a).

While this triad[4] of social, communicative, and imaginative impairments forms a consistent and diagnostically useful identifier of autism, there is considerable variation in the severity of these deficits. For this reason, autism is now regarded as a *spectrum* of more or less severe impairments in these three categories—a spectrum, moreover, that is continuous with the 'normal' population (Wing, 1996; Baron-Cohen, 1999). Thus, the distinction between 'autistic' and 'non-autistic', while a useful shorthand for clinical purposes, is no longer thought to trace a sharp dividing line between cognitively typical and atypical individuals. Nor is it always appropriate to label the presence of autistic characteristics as a *disorder*; hence, the term 'autism spectrum disorders' is giving way to the more neutral 'autism spectrum conditions' (e.g., Barnes & Baron-Cohen, 2012).

Of particular interest on the autistic spectrum are the milder forms of autism known as *Asperger syndrome* and 'high-functioning autism'.[5] Asperger cases are distinguished by normal or above-average intelligence (compared with below-average intelligence in other autism cases: Bauman, 1999; Frith, 2001a) and, by current diagnostic criteria, undelayed language development (Hill & Frith, 2003). Though socially awkward and prone to engaging in solitary and repetitive activities (thus satisfying characteristics (1) and (3) above), they often display unusual cognitive abilities, including above-average linguistic aptitude (though their speech tends to be unconversational, stilted, and/or socially inappropriate, thus satisfying the general characteristic (2) of abnormal communication) (Wing, 1996, pp. 20, 40). Such individuals can provide an insight into the autistic mind that, because of communicative impairments, cannot be obtained from more severely

---

[4] The latest version of the American Psychiatric Association's (2013) *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)* reduces this common diagnostic triad into a dyad, merging social and communicative impairments into one criterion (Hassall, 2016).

[5] 'Some confusion exists between the labels Asperger syndrome and high-functioning autism' (Hill & Frith, 2003, p. 281). Asperger syndrome has been dropped entirely as a diagnostic category in the *DSM-5* (Hassall, 2016), while the WHO's *International Statistical Classification of Diseases and Related Health Problems (ICD-10)* marks it as a 'disorder of uncertain nosological validity' (World Health Organization, 1992/2016, F84.5).

autistic cases. Evidence from Asperger cases relating to the self is discussed in § 7.3 below.

Whereas the ultimate causes of autism are still uncertain, there are now several neurocognitive theories available to account for the signs and symptoms of autism. The most diagnostically successful such account is that autistics lack a *theory of mind*. I outline this theory in § 7.2.2 and discuss its implications for an account of the self in § 7.3.1. Beyond this, the non-social characteristics of autism are theorized to derive from *weak central coherence* and/or *executive dysfunction* in the brains of autistics. I discuss these theories in § 7.2.3 and the potential implications of weak central coherence for the self in § 7.3.2. It may be that weak central coherence and executive dysfunction are not so much rival theories to theory-of-mind deficit as complements that attempt to explain those features of autism that the absence of a theory of mind cannot account for: a 'restricted repertoire of behaviours, rigidity and perseveration' (Hill & Frith, 2003, p. 284). If the theories are complementary, this raises the interesting question whether autism is, in fact, a unified condition. I cannot go into this here. But I can, at least, discuss the prominent theories put forward to account for the different symptoms of autism, and how each might affect the self.

7.2.2 Theory-of-mind deficit

In the psychology literature, *theory of mind* (hereinafter *ToM*) refers to the ability to 'mind-read' or 'mentalize', that is, to ascribe *mental states* such as beliefs to other people on the basis of their observable behaviour. This mind-reading ability allows us to engage in and navigate the complex so-

cial interactions, 'both co-operative and competitive', that are characteristic of our species (Botterill & Carruthers, 1999, p. 77).[6]

The theory that the social impairments of autism could be explained by autistic individuals' lack of a ToM was first proposed by Baron-Cohen, Leslie, and Frith (1985). The experimental paradigm they used to support this theory—the 'Sally–Anne' false-belief task—has since become a standard diagnostic test for autism in children:

> There were two doll protagonists, Sally and Anne. . . . Sally first placed a marble into her basket. Then she left the scene, and the marble was transferred by Anne and hidden in her box. Then, when Sally returned, the experimenter asked the critical Belief Question: "Where will Sally look for her marble?". If the children point to the previous location of the marble, then they pass the Belief Question by appreciating the doll's now false belief. If however, they point to the marble's current location, then they fail the question by not taking into account the doll's belief. (p. 41)

The correct ascription of a *false belief* to the doll protagonist is taken as evidence of a functioning ToM; failure to ascribe a false belief to Sally is indicative of ToM deficit. Neurotypical children generally pass the false-belief task by about three or four years of age; the vast majority of autistic children with matched mental ages do not. A corollary of this inability to ascribe false beliefs and thus recognize deception is their inability to *use* pretence and deception themselves (Baron-Cohen 1995, ch. 5; Botterill & Carruthers 1999, ch. 4). Other mind-reading tasks have been developed for older children (*faux-pas* recognition: Baron-Cohen et al., 1999) and adults (reading

---

[6] There are two main rival theories of the cognitive mechanism underlying ToM: *theory-theory* and *simulation theory*. The former is the view that our mind-reading ability derives from a core of theoretical knowledge that is implicitly deployed when we ascribe mental states to each other (and ourselves) as causal explanations of our actions. The latter theory holds that we do not draw on theoretical knowledge (however implicit) when we mind-read, but rather *simulate* others' cognitive situation in our own minds and so come to ascribe the appropriate mental states to them (Botterill & Carruthers, 1999, ch. 4). The debate between these rival theories is not critical to this discussion, nor does autism help settle it. The ability of some high-functioning autistic individuals, such as Temple Grandin (Sacks, 1995), to resort to an acquired, *explicit* ToM substitute (see § 7.3.1) suggests that at least *some* of our ToM capacities do resemble a theoretical body of knowledge, since it can be replicated by active conscious theorizing. But the remaining social awkwardness of such individuals suggests that theoretical knowledge of a repertoire of mental states is still not the whole story of our 'mentalizing' capacity.

mental states from the eyes of people in photographs: Baron-Cohen et al., 1997; 2001). In these tasks, too, mind-reading ability is consistently impaired in autistic subjects relative to controls.

Baron-Cohen (1995) hypothesizes that one of the mechanisms underlying ToM is a shared-attention mechanism (SAM), which allows us both to monitor and to direct others' attention. A functioning SAM is evidenced by behaviour such as following another's gaze, and pointing and showing gestures. Such shared-attention behaviour is also markedly absent in most autistic children and allows an even earlier diagnosis than false-belief tasks, at about eighteen months of age.

Further support of the ToM deficit theory of autism comes from autistic individuals' characteristic communicative impairments. These include both pragmatic and lexical-semantic aspects of language use (Tager-Flusberg, 2000). Pragmatic communication impairments include an overly literal use and understanding of language and a lack of conversational relevance, which suggest a failure to understand that 'communication is about . . . *intended* rather than literal meaning' (p. 126), where understanding intended meaning requires the ascription of intentions to others and therefore a functioning mind-reading ability. Narrative discourse is impoverished, with little appeal to mental states as motivators of action. Lexical deficits in cognition verbs ('think', 'know') correlate with poor performance at mind-reading tasks.

Significantly, where autistic children do pass ToM tasks, this requires a higher level of lexical ability than in neurotypical children, suggesting that the acquisition of language, if it occurs, can to some extent be used to overcome (or *mask*) ToM deficits. However, shared-attention deficits in turn make it difficult for autistic children to *acquire* language. Though the causal links between shared attention, ToM, and language acquisition are likely to be complex (Happé, 1995; Lorusso et al., 2007), there are strong correlations between impairments in all three domains.

7.2.3 Weak central coherence and executive dysfunction

While ToM deficit provides a neurocognitive theory of the social and communicative impairments of autistics, including impaired imagination in social situations, it does not explain the third characteristic of autism: ste-

reotyped and repetitive behaviour and a restricted repertoire of interests and activities, which seem indicative of a more general lack of imagination. Two further cognitive abnormalities have been suggested as explanations of the non-social impairments of autistics: *weak central coherence* and *executive dysfunction*.

The notion of central coherence is simply that normal cognitive functioning involves 'the tendency to draw together diverse information to construct higher-level meaning in context . . . For example, the gist of a story is easily recalled, while the actual surface form is quickly lost' (Frith & Happé, 1994, p. 121). Autistics, meanwhile, have a 'tendency to focus on the local, rather than global aspects of an object of interest' (Hill & Frith, 2003, p. 284), often showing a greater ability in processing and retaining details than non-autistic individuals, as exemplified by autistic 'savants' like the scientist Temple Grandin and the artist Stephen Wiltshire (Sacks, 1995). But they show a concomitant weakness in cognitive tasks that require assembling disparate but related pieces of information into a meaningful whole.

Such *weak central coherence* has been hypothesized to be the result of 'poor connectivity between more basic perceptual processes and top-down modulating processes, perhaps owing to a failure of [neural] pruning' (Hill & Frith, 2003, p. 284). This lack of normal neural pruning during development may account for the fact that 'the autistic brain . . . is on average larger and heavier than the normal brain' (pp. 282–3). More recently, weak central coherence has been associated with disrupted neural connectivity and the balance of inhibitory and excitatory processes (Zikopoulos & Barbas, 2013). However, research in this area is still somewhat inconclusive and, as yet, lacks 'well-specified mechanistic models of altered cerebral communication' in autism spectrum conditions (Vasa et al., in press).

Autistics' lack of imagination, undue perseveration, and predilection for repetitive activities also suggest impairments in 'planning working memory, impulse control, . . . and the initiation and monitoring of action' (Hill & Frith, 2003, p. 285). These executive capacities correlate with frontal-lobe activity, whose impairment is thus known as *executive dysfunction*. The main problem with advancing executive dysfunction as an explanation of specifically autistic traits is that it is neither an exclusive nor an

inclusive marker of autism, since it is a feature of psychopathologies other than autism[7] and may not be present in all autistic cases (p. 286).

7.2.4 A 'Wittgensteinian' critique of neurocognitive theories

In a recent essay, Peter Hobson (2009) advances the view that current psychological theories of autism, by focusing on neurocognitive abnormalities of the individual brain, ignore the central role of abnormal '*relations* with people and things' (p. 254) in the presentation of autism. Citing Wittgenstein's (1980) extant notes on the philosophy of psychology, Hobson suggests that the key to understanding autism lies in its sufferers' inability to join in with shared '*forms of life*' (Wittgenstein, 1953/2009, p. 226e/238e).

It may seem odd that a clinical neuroscientist should appeal to Wittgenstein in his theorizing about autism. In the context of a study of the self, however, Hobson's choice is less surprising, since Wittgenstein may be regarded as something of a (perhaps unwitting) godfather to *social* theories of the self (see e.g. Bakhurst, 1995). At any rate, he may be credited with the view that our folk-psychological concepts gain their meaning only in our shared use of these concepts.[8]

In contrast to this 'social' view of folk psychology, Hobson (2009) detects a 'strong individualistic bias' in contemporary psychological theories of autism (p. 254). This requires some elucidation. Presumably, by 'individualistic' Hobson means that neurocognitive theories of autism assume that abnormalities in autistic individuals' neurocognitive functioning will explain their abnormalities of social behaviour. In that sense, these theories are indeed individualistic. But this does not imply that they do not account for 'the nature and developmental implications of the children's *relations* with people and things' (ibid.). To the contrary, that is precisely what neurocognitive theories of autism are meant to account for. Explaining the signs and symptoms of autism in terms of ToM deficit is a way of accounting for the neurocognitive substrate (or lack thereof) that underlies our ability (or lack thereof) to engage in the shared folk-psychological discourse that shapes our relations with, and understanding of, other people. The so-

---

[7] Cf. my discussion of impaired executive function in frontal-lobe amnesia, § 8.2.2.

[8] For a more recent exposition of this view, see Kusch (1997).

cial *instantiation* of that discourse and understanding is not thereby called into question. But it is at the level of the individual brain that we must look for the underlying *capacity* to partake in that discourse.

But Hobson goes further. He suggests that the direction of theorizing employed here (individual ability/disability accounts for normal/abnormal social engagement) may be misguided. He muses 'whether much communicative, linguistic, and cognitive dysfunction in autism could prove to be the result, and not the cause, of the children's social-affective/relational impairments' (p. 255). Of course, it is entirely possible that social impairments feed back into cognitive dysfunction (difficulties in language acquisition as a result of impaired shared attention (§ 7.2.2) being a case in point). But what causes the social impairments in the first place? Given that any kind of behaviour requires an appropriate neurocognitive infrastructure that enables such behaviour, it seems reasonable to suggest that it is the absence or impairment of this neurocognitive infrastructure that causes the absence or impairment of the correlated behaviour, and not vice versa.

Another criticism for which Hobson recruits Wittgenstein's reflections on psychology is directed specifically at the ToM deficit theory. He quotes Wittgenstein (1980) on our ability to perceive others' mental states without conscious inference:

> 'We *see* emotion.'—As opposed to what?—We do not see facial contortions and *make the inference* that he is feeling joy, grief, boredom. We describe a face immediately as sad, radiant, bored, even when we are unable to give any other description of the features. (§ 570)

Thus, Hobson argues, 'we do not need to "theorize" about the nature of people's . . . feelings, intentions and the like' (p. 246).

Of course, Wittgenstein and Hobson are quite right that our ability to detect others' mental states does not, as a rule, involve *explicit* theorizing or conscious inferring.[9] But to take our 'direct perception' of mental states as evidence against ToM is to ignore that our perceptions are themselves theory-laden. Thus, if we 'see emotion' we do so because the seeing is already modulated by an active ToM (Lavelle, 2012).

---

[9] The likely exception being Asperger syndrome cases, who may employ an acquired and conscious ToM substitute to perform the same tasks (Frith & Happé 1999)—but do so precisely because they do not have access to a normal, implicit ToM (see § 7.3.1).

It may be that the term '*theory* of mind' is, as Hobson suggests, somewhat misleading in that it might be seen to imply that a functioning ToM involves explicit, quasi-scientific theorizing. But that is not how ToM is generally understood nor the best account of how it works, which is rather that ToM is an innate system of *implicit* theorizing ability (Botterill & Carruthers, 1999, ch. 4). This account of ToM is consistent with Wittgenstein's observations. The fact that we do not make conscious inferences about others' mental states but seem to perceive them directly is no argument against ToM operating at the subpersonal level and shaping our perceptions.

Overall, then, Hobson's 'Wittgensteinian' criticisms of neurocognitive theories of autism do not seem very compelling. We cannot *explain* the defect in terms of its symptoms. Hobson is right, however, to remind us that impaired social relations are a defining feature of autism. And in discussing the *implications* of autism, it may be useful to refer to a shared 'form of life' that is partly, perhaps largely, absent from the lives of autistics. I will return to this theme in discussing the 'social self' of autistic individuals (§ 7.3.4).

7.3 The self in autism

Having looked at the specific defects present in autism spectrum conditions, we can now ask how each of them affects the self. If, as I suggest, the self is a complex and dynamic system comprising different self-representational capacities (§ 1.4), we should expect the peculiarities of autistic selves to be fairly specific. Thus, I will now examine in turn the consequences for the self that may arise from ToM deficit, weak central coherence, autistics' memory organization, and social impairments. As will become apparent, the deficits discussed sometimes seem to pull in different directions when it comes to assessing how functional an autistic self might be, and caution is advised against generalizing over all conditions and individuals on the autistic spectrum.

7.3.1 Consequences of theory-of-mind deficit

How does ToM deficit affect the self? If, as I propose, the self is a system for self-monitoring, this would usually involve monitoring one's own mental

states. Uta Frith and Francesca Happé (1999) hypothesize that autistics' lack of a theory of *other* minds may imply a lack of a theory of their *own* minds. That is to say, if an individual's ToM mechanism is deficient, she can no more ascribe mental states such as beliefs to herself than she can ascribe them to others.

Support for this hypothesis comes from experimental work using a variant of the false-belief test, in which autistic children's inability to attribute false beliefs to *others* correlates with their inability retrospectively to attribute false beliefs to *themselves* ('Smarties' task, Hogrefe et al., 1986; Gopnik & Astington, 1988). It also fits Alan Leslie's (1987) theory that our representations of states of the world are kept cognitively separate from our metarepresentations of mental states (ToM), a theory supported by the ability of autistic children to reason about cause-and-effect relations (states of the world) in stories, *without* being able to comprehend motivations (mental states) of characters in those stories (Happé et al., 1996).

Frith and Happé (1999) hypothesize further that if ToM impairment implies impaired mentalizing ability with respect to oneself, it should lead to an impaired or unusual form of *self-awareness* in autistic individuals. For, without ToM, their self-awareness cannot include metarepresentations of their mental states. To support this second part of their hypothesis, Frith and Happé point to individuals with Asperger syndrome, who unlike other autistics often pass ToM tasks, but do so because they have an explicitly *learned* (rather than implicit and intuitive) ToM ability. This ToM substitute does not, however, appear to produce the self-ascription of mental states.

Thus, Hurlburt, Happé, and Frith (1994) found that self-reported self-awareness in Asperger cases differed from that of neurotypical subjects by being mostly visual, concrete, and devoid of descriptions of 'other forms of inner experience'. This Frith and Happé (1999) take to suggest that their Asperger subjects' 'unusual' ToM ability—a learned substitute for a normally intuitive process—produces qualitatively 'unusual' self-awareness, in which the ToM substitute is not activated, because it is deemed necessary only for understanding or describing others.

> This is not to say that these individuals lack mental states, but that in an important sense they are unable to reflect on their mental states. Simply put, they lack the cognitive machinery to represent their thoughts and feelings *as* thoughts and feelings. (p. 7)

'Representing one's thoughts and feelings *as* thoughts and feelings' may seem a cumbersome description of what neurotypical individuals automatically and intuitively do as part of their normal self-monitoring. How, one might ask, could we not? But that is the point: the self-experience of autistic individuals, while it presumably does involve the *experience* of cognitive and affective states, does not include their *categorization* along the familiar mental-state terms which to those gifted with ToM seems natural and automatic. Frith and Happé's formulation attempts to capture quite how different the 'inner life' of an autistic individual without ToM is from that of neurotypical individuals.

But their failure to categorize or metarepresent their mental states in ToM terms does not mean that autistic individuals are *unaware* of their thoughts and feelings. And indeed it seems that such awareness does not require ToM. Shaun Nichols and Stephen Stich (2003) theorize that, as well as—and *prior* to—ToM, we have a 'distinct mechanism that is specialized for detecting one's own mental states' (p. 163), which they call the 'Monitoring Mechanism' (MM). Evidence from developmental psychology supports the hypothesis that the MM is online in child development some time before ToM: three-year-olds perform better in detecting pretence, false beliefs, and knowledge in themselves than in others, and they also do better in difficult perspective-taking tasks when they involve shifting their *own* perspective as opposed to taking *another's* perspective. This suggests a clear dissociation between self-monitoring (via MM) and ToM capacity.

Further, Nichols and Stich note that two of the subjects in Hurlburt et al.'s (1994) study of self-reported self-awareness in Asperger cases do, in fact, use some mental-state vocabulary with reference to themselves. Overall, then, while

> the inner lives of autistic individuals differ radically from the inner lives of most of us . . . people with autism and Asperger's syndrome *do* have access to their own inner lives. They are aware of, report, and remember their own beliefs and desires as well as their occurrent thoughts and emotions. (Nichols & Stich, 2003, p. 185)

Nichols and Stich do not suggest that, in neurotypical individuals, ToM plays no role at all in the self-monitoring of mental states. Once the ToM capacity is online, it can of course be deployed *in addition to* MM. This raises the question to what extent neurotypical self-monitoring involves the

ToM capacity. Kai Vogeley and colleagues (2001) conducted a neuroimaging study to investigate whether ToM and self share the same neural mechanisms. Participants were presented with a number of stories involving either taking only someone else's perspective ('TOM'), taking only one's own perspective ('SELF'), or taking both perspectives.[10] The neural mechanisms involved in taking these perspectives were studied using functional magnetic resonance imaging (fMRI). Both TOM and SELF factors correlated with increased activity in the anterior cingulate cortex.[11] The TOM factor, but not the SELF factor, also correlated with increased cortical activity at the left temporal pole.[12] The SELF factor, but not the TOM factor, also correlated with increased cortical activity at the right temporoparietal junction.[13] TOM–SELF interaction was observed in activity in the right prefrontal cortex. In sum, the imaging results showed both an overlap and a dissociation between the neural correlates of the TOM and SELF conditions. (See fig. 7.1 for an illustration of these brain areas.)

These results suggest two things. First, the exercise of ToM and self-monitoring capacities seems to be linked by a common neural substrate.[14] Given that, in the context of the experiment, participants were asked to engage in similar activities taking their own perspective and that of others,

---

[10] The study also included two control conditions: a collection of unlinked sentences (to serve as a baseline) and stories without a 'TOM' or 'SELF' component ('physical stories').

[11] Anterior cingulate activation is consistent with other neuroimaging results finding ToM to correlate with activity in the adjoining medial prefrontal cortex including the anterior paracingulate cortex (Frith & Frith 1999; Frith 2001b; Gallagher & Frith 2003). It is also consistent with Damasio's (1999, ch. 8) identification of the cingulate cortex as a correlate of the core self.

[12] This is consistent with other neuroimaging results finding ToM to correlate with left temporopolar (periamygdaloid) cortex activation (Frith, 2001b).

[13] The right temporoparietal junction is, however, identified as involved in the ToM network by other studies (Frith, 2001b).

[14] It might be argued that ToM and self could still be separate systems whose sharing of neural correlates is entirely coincidental. However, Vogeley and colleagues' results showed *no* activation of anterior cingulate cortex in the control tasks, whereas it was consistently active in *both* TOM and SELF tasks. The linking of the systems for both provides an explanation for this activity, while the hypothesis of coincidence would itself require further explanation.

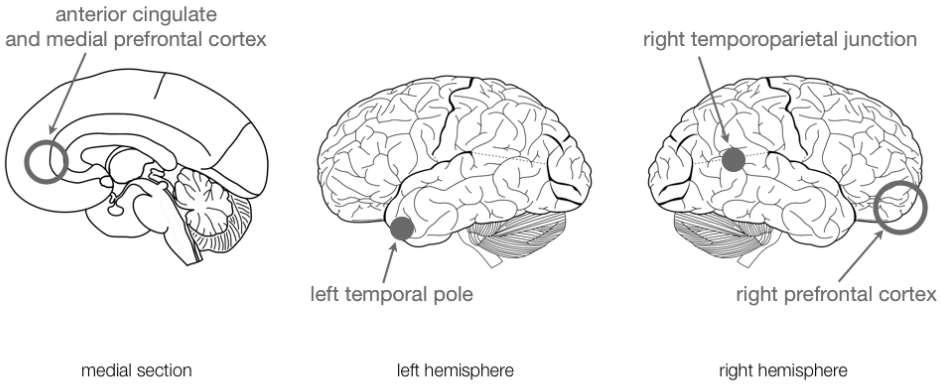

Figure 7.1 Brain areas associated with TOM and SELF factors in Vogeley et al.'s (2001) study.

this seems unsurprising. More generally, however, it is also not surprising that similar neurocognitive mechanisms should be deployed when we engage in self-monitoring and when we take another person's perspective, if both activities involve the ascription of mental states to oneself or another. So far as the self *qua* self-monitoring system involves observing one's mental life, one would expect it to be linked to our ToM capacity.

Secondly, however, Vogeley and colleagues' results also showed a differential activation of brain areas between the two conditions. Thus, even in the limited context of this experiment, with its similar tasks in both conditions, taking a self-perspective involves a neural architecture that differs from that involved in taking the perspective of others. Again more generally, this suggests that the system or systems involved in self-monitoring do not merely involve the application of socially useful capacities (like ToM) to oneself, but have their own neural channels. Vogeley and colleagues' subjects did not include individuals with autism, but the partial dissociation of TOM and SELF correlates in their findings suggests that there is a neural basis for self-monitoring without theory of mind, and thus, a neural basis for autistic selves, however different their operations may be from neurotypical ones.

Thus, it would be misguided to suggest that there is no *self* when there is no ToM capacity, which is but *one* higher-level self-representational capacity (albeit a *socially* important one). Contrariwise, if self-monitoring sys-

tems are dissociable from ToM, that is an argument against any account of the self that *requires* ToM—as would narrative constructionist and essential narrative-capacity views, since the 'reason-giving' character of self narratives would seem to require ToM capacity. Narrative capacity is also impaired in other ways in autism, as I'll now discuss.

7.3.2 Consequences of weak central coherence for narrative capacity

As weak central coherence is characterized as a deficit in 'pulling information together for higher-level meaning' (Hill & Frith, 2003, p. 284), some of its most obvious consequences are seen in peculiarities concerning narrative abilities of autistic individuals. Thus, autistic children attempting to narrate a scene with puppets that has been played before them resort to 'bizarre/inappropriate utterances', showing 'a kind of failure to comprehend the story presented or, in some cases, even to comprehend what a story is: some of the narratives produced by the autism group certainly suggest a poor awareness of the puppets as characters rather than simply objects' (Loveland et al., 1990, p. 19). In autistic adults, another peculiarity in story-telling is that they 'are biased toward providing *local* over *global* details about each element [of a story], regardless of whether the element involved mental content' (Barnes & Baron-Cohen, 2012, p. 1557). Thus, independently of ToM deficit, there are deficits in narrative comprehension and production concerning the nature and function of stories. A narrative is not understood or presented as a coherent whole involving the meaningful doings of characters; rather, specific details are picked out that may have no strong bearing on the storyline, but are dwelt on at the expense of the 'big picture' of the narrative as a whole.

What, then, are the consequences of this for the self? For defenders of *narrative constructionist* views (§ 5.3.2), autistic individuals with these deficiencies in narrative understanding and construction could not have a self, in the sense of a narrative construct, at all. Nor could autism be accommodated by *essential narrative-capacity* views, for even autistic individuals with verbal abilities may (i) struggle to construct a self narrative at all, and/or (ii) fail to understand the point of doing so, or, if a self narrative of some kind *is* constructed, (iii) produce narratives lacking in some or all of the features that we would normally expect of meaningful personal stories.

I have, of course, already rejected both these views (§§ 6.2, 6.3), so we should not take the absence or peculiarity of narrative capacity to entail a lack of self—just the lack of a particular *type* of self-representation. And where autistic individuals *do* engage in autobiographical narrating, their way of narratively self-representing is idiosyncratic.

Francesca Happé's (1991) analysis of the autobiographical writings of three adults with Asperger syndrome shows them to be examples of case (iii) above. Though the subjects are verbally highly able and clearly see some point in writing their life stories, the results diverge from other autobiographical writings in a number of ways. Besides evidence of 'social naivety' (a consequence of ToM deficit), their writings often do not follow ordinary standards of communicative *relevance*,[15] displaying, on the one hand, frequent apparently unmotivated changes of topic and, on the other hand, a tendency to perseverate. Happé notes that 'it is hard to find anything formally wrong [with these writings], rather the reader is left with an overall impression of oddness' (p. 229). Crucially, one of the writers, Barry, 'does not recognise what knowledge is shared and what is personal' (p. 214). It may seem, then, that the autobiographical writings of autistic individuals fail to adhere to certain basic narrative and communicative rules, such as maintaining relevance and a sense of what can reasonably be expected of one's audience.

But here perhaps the fault, if fault there is, lies not with the autistic writers, but with the expectations of a neurotypical reader of their writings:

> For example, it may be that the difficulty the reader has in following these writers' subject changes is in part due to the different availability of various contexts to the autistic and the normal person. Communication may break down with each party getting hold of the wrong end of the stick if the autistic speaker is intending his utterance to be processed in a context that the normal hearer does not have easily accessible at the time. (p. 229)

If, then, autistic self narratives leave an 'overall impression of oddness' with a neurotypical reader, that impression is likely to go both ways. Happé says of Temple Grandin, another of the autistic writers she studied, that she 'seems to see autism not so much as a handicap but as an alternative cognitive style' (p. 212). If that is so, autistic self-monitoring, and the auto-

---

[15] Cf. Sperber & Wilson, 1986.

biographical narratives this may give rise to, will necessarily reflect that alternative cognitive style.

While the 'cognitive style' of neurotypical individuals, for reasons of cognitive economy, tends towards generalization, focussing on salient points and discarding unnecessary details, there may, in autistic individuals, well be a 'tolerable *cost* to processing unrelated and irrelevant information'. And whereas we generally process information by reference to a given context, which in communication is a shared context, for autistic individuals, 'the most accessible *context* is idiosyncratic' (p. 237). This is also reflected in the organization of memory, and it is the peculiarities of autistic memory to which I now turn.

7.3.3 Memory: abilities and differences

Some autistic individuals have exceptional mnemonic abilities, often for numbers and dates (e.g. Scheerer et al., 1945) and/or for visual detail, as in the case of the autistic artist Stephen Wiltshire (Sacks, 1995). These 'savant' abilities are striking not only in their rapidity and accuracy of recall, but also in their isolation from other cognitive processes: they seem to bear no relation to how developed or underdeveloped the individual's cognitive functions are generally; they are usually limited to a specific domain (like numbers or images) and do not transpose to other cognitive domains; they are unbidden and untrained and do not increase with practice or decrease for lack of it.

One area of memory, however, that is generally not marked by exceptional recall in autism is autobiographical memory. Quite the reverse: recent studies note deficits in autobiographical memory in both children and adults with both 'classic' autism and Asperger syndrome. Bruck and colleagues (2007) studied autobiographical memory recall in autistic children and found them to have poorer recall than controls both of staged events and of past personal events, particularly early-life memories. Studies of adults with Asperger syndrome by Goddard and colleagues (2007) and Tanweer and colleagues (2010) both showed fewer specific memories being recalled by the Asperger groups than the control groups. Adler and colleagues (2010) found autobiographical memory impairment in individuals

with Asperger syndrome and high-functioning autism to correlate with ToM impairment.

But, given the complexity of autobiographical remembering discussed in ch. 4, it would be wrong to suggest that autobiographical memory is poorer in autistic individuals across the board. Rather, it may be that studies merely testing quality of event recall and number of events recalled obscure a more nuanced picture of autobiographical memory abilities in autistic individuals, namely a differential ability with respect to *semantic* and *episodic* memory (§ 4.2.1). For instance, Klein and colleagues (1999) report on the case of R.J., 'a high-functioning autistic individual [who] was found to have accurate knowledge of his traits, despite severely limited access to the personal experiences on which that knowledge was based' (p. 413). A study by Bowler and colleagues (2000) comparing word recognition performance between Asperger syndrome and neurotypical adults found no overall difference in performance, but Asperger subjects were more likely to respond that they 'knew' rather than 'remembered' the memorized words, which Bowler and colleagues take to indicate a moderate impairment in episodic but not semantic memory. Crane and Goddard (2008) studied autobiographical memory in autistic adults and also found a dissociation between personal episodic memory, which showed deficits in the autistic group, and personal semantic memory, which did not.

These results suggest an impaired ability in autism to recall oneself as a protagonist in a past event, as opposed to unimpaired recall of merely factual information. The study by Tanweer and colleagues (2010) is even more suggestive in this respect. One of its aims was specifically to test whether autobiographical memory in Asperger cases was processed as self-referential. Here again, besides recalling fewer specific memories, the Asperger group more often classed autobiographical memories as 'known' (noetic) than as 'remembered' (autonoetic). Thus, whereas factual recall was present, recall of the involvement of self in the process of acquiring that knowledge was missing.

And yet it may be misleading to use the label 'episodic memory' for what is impaired in autistic recall, because episodicity is precisely not the problem. Many high-functioning autistics have an exceptional capacity for recalling events in their episodic detail. As Oliver Sacks (1995) reports on Temple Grandin, the autistic animal scientist:

> I was struck both by the vividness of the re-experience, the memory, for her—it seemed to play itself in her mind with extraordinary detail—and by its unwavering quality. It was as if the original scene, its perception (with all its attendant feelings), was reproduced, replayed, with virtually no modification. This quality of memory . . . seemed to me both prodigious and pathological—prodigious in its detail and pathological in its fixity, more akin to a computer record than to anything else. Such computational analogies, indeed, are frequently brought up by Temple herself: 'My mind is like a CD-ROM in a computer—like a quick-access videotape. But once I get there, I have to play that whole part.' She could not just focus, for instance, on the cradling of an animal . . .; she had to play, in memory, the entire scene . . . (pp. 268–9)

While episodic memory is normally a reconstruction (§ 4.4), this does not seem the case here. Rather, Grandin's episodic recollection seems a faithful replay of the past event in her consciousness.[16] In a sense, we might say that Grandin is a perfect Lockean; she really is a being that can 'repeat the idea of any past action with the same consciousness it had of it at first' (Locke, 1690, II.xxxvii.10).

But Sacks notes that this unusual, 'prodigious' actualization of Lockean recall also strikes him as 'pathological in its fixity'. It may be, then, that normal *reconstructive* episodic remembering, while lacking the videographic faithfulness of Grandin's recall, provides us with a more useful tool for locating self in time. A fixed unchanging episodic memory does not, after all, take account of the passage of time. Self-representation over time is a dynamic process: perhaps it is precisely the reconstructive elements of personal remembering, the reshaping of our memories in the light of subsequent experience and of present context, that provide us with a *sense of self* over time, as opposed to the mere impersonal knowing that some event (however detailed its recollection) occurred in the past.

In sum, memory in autism shows normal or exceptional ability in certain areas, but impairments of autobiographical episodic memory, shown either by a lack of self-reference in remembering or by the abnormal 'fixity' of episodic retrieval. Thus, of the three 'enabling systems' identified by

---

[16] Grandin (2009) herself likens her mind to a 'search engine, set to locate photos. All my thoughts are in photo-realistic pictures' (p. 1437). Sacks' description of her again recalls Borges' (1942) description of the fictional Ireneo Funes, who is said to remember everything in intricate detail, while, like many autistic individuals (if not Grandin), he is unable to form abstractions and generalizations.

Klein and colleagues (2004)—self-reflection, agency/ownership, and a 'sense of personal temporality' (see § 4.2.3)—the first and third may be impaired or lacking in autistic individuals. We should note, however, that different individuals with autism may have very different memory abilities and deficits. For instance, Grandin's unusual episodic recall seems a corollary of her excellent capacity for mental visual simulation. Other people with autism lack this, and consequently do not have such videographic episodic memory, but may be able to compensate with their semantic memory abilities. There is, then, no paradigm case of autistic memory, but a variety of different abilities which may be more or less useful in supporting autobiographical memory.

7.3.4 Social aspects

An individual's self is also, in part, shaped by his social interactions.[17] It is in this area that the most obvious deficits of the autistic self appear, given the autistic characteristics of social and communicative impairments. These impairments can be categorized as either affecting *primary* or *secondary intersubjectivity*, where primary intersubjectivity concerns person-to-person relations and secondary intersubjectivity person–person–world relations (Hobson, 2011).

Primary intersubjectivity in autistic individuals is characterized by a relative lack of engagement with other people, as reported by parents of autistic children (Wimpory et al., 2000) and shown in experimental studies. These reveal a reduced propensity to offer spontaneous or prompted greetings to strangers (Hobson & Lee, 1998) and difficulties with engaging in conversation (Capps et al., 1998). There are also significant differences in affective engagement and the 'smoothness' of the interaction (García-Pérez et al., 2007), reaffirming the stereotype of the autistic individual as socially

[17] I will return to this point in discussing dementia (§ 8.2.3).

detached and awkward.[18] It must be noted, however, that since these indicators reflect *group* differences, they do not predict that any one autistic individual will perform worse at social-engagement tasks than any one non-autistic control. Autistics' difficulties in primary intersubjectivity, while a general characteristic of the condition, remain a matter of degree.[19]

Autistic impairments of secondary intersubjectivity—i.e. relations not just with another person but with another person's own relation to a shared world—are more marked. Autistic individuals show a significant lack of engagement and emotional connectedness with others' experience of the world. Thus, while they readily imitate others' goal-directed (intentional) actions (Charman & Baron-Cohen, 1994), they do not imitate others' emotional and expressive attitudes while performing those actions (DeMyer et al., 1972; Hobson & Lee, 1999; Rogers et al., 2003) and show less shared attention behaviour during imitation tasks (Hobson & Hobson, 2007).

Autistics also show a disjunction between *self*-centred emotions which, contrary to a popular conception, are unimpaired in autism (though rarely directed at other people), and *other*-centred emotions, which are impaired in autism (Hobson, 2011). For instance, pride is expressed by autistic children only as pleasure at their own accomplishment, not as anticipation of praise from another person (Kasari et al., 1993). In another experimental setup, autistic subjects showed a marked absence of 'anticipatory concern' for others' hurt feelings (Hobson et al., 2009). Thus, the anticipation of the emotional responses of others, and correlated own emotions (anticipatory pride or pity), seem impaired.

---

[18] Specific difficulties with verbal communication appear to straddle the division between primary and secondary intersubjectivity. An interesting but inconclusive case is the misuse/reversal of first- and second-person personal pronouns (Kanner, 1943; Charney, 1981). A study by Lee, Hobson, and Chiat (1994) found that *comprehension* of these pronouns was virtually unimpaired in the autistic test group, but that their *use* by autistic subjects was less frequent relative to controls in a visual-perspective task. This relative reluctance by autistics to employ first- and second-person pronouns may be connected to a deficient 'grasp of reciprocal roles in dialogue' (Hobson, 1990, p. 172) (primary intersubjectivity) and, more generally, difficulties in perspective-taking (Hobson & Lee, 1999) (secondary intersubjectivity).

[19] This applies particularly to cases of high-functioning autism and Asperger syndrome, where acquired compensation mechanisms may exist for primary intersubjective relations (§ 7.3.1).

These deficits in anticipating and mirroring others' emotions suggest a general lack of ability in autistics for *taking the perspective of others.* In an illuminating discussion of perspective-taking in Asperger syndrome, Frith and de Vignemont (2005) note that in normal intersubjectivity, we adopt both egocentric and allocentric stances towards others. From an egocentric stance, one represents others in relation to oneself. Adopting an allocentric stance, others are represented as independent of oneself. While, in non-autistic individuals, these allocentric and egocentric stances interact and feed off each other, Frith and de Vignemont hypothesize that they are disconnected and 'polarized' in Asperger cases: social relations are processed from an extreme egocentric stance that is never transposed to another's point of view. Where an allocentric stance is adopted, it is 'highly abstract' and represents social relations from neither party's point of view.

In a similar vein, Hobson theorizes that full self-awareness of a kind that is lacking in most autistic individuals involves 'identifying with the attitudes of other people' (2011, p. 582; cf. Hobson et al., 2006, pp. vii, 126). This choice of phrase recalls G. H. Mead's (1925; 1934) social behaviourist account of selfhood, in which taking the attitudes of others is a necessary condition for developing a self. Hobson's interpretation is not that far-reaching, however: he does not hold that the self is entirely constituted in one's social relations. But he notes that the development of the self usually includes the development of the ability to identify with others' attitudes. Absent this ability, one is led to conclude, the autistic self remains somehow incomplete.

However, the distinction between primary and secondary intersubjectivity shows that the social self is not monolithic but itself divides into different components. Autistic impairments in *primary* intersubjectivity concern the manner and intensity of social relations more than the *fact* of such relations, suggesting that we are dealing with an impoverished or differently configured—rather than absent—social self in autism. It is in *secondary* intersubjectivity that the social self in autism is most obviously impoverished, in that autistic individuals simply cannot, as it were, put themselves in another person's shoes: their representation of another's perspective, where it exists, remains abstract and impersonal. It may be, then, that what autistic individuals lack is a sense of being one among other *like* individuals. It is difficult to quantify how that lack of a comparative standard

affects an autistic individual's sense of 'being someone' overall. We may perhaps guess that it could be a sense of being someone who is very much alone in the world.

And yet, Hobson's above-cited injunction that we should understand autism as an inability to join in a shared 'form of life' (§ 7.2.4) falls short of recognizing that people with autism may be able to develop their *own* forms of existence, in particular in their social relations with other autistic individuals. Thus, Oliver Sacks (1995) reports on the B. family—parents and an older son with Asperger's syndrome, and a younger son with classical autism—who are able to share with each other the experience of being surrounded by a profoundly baffling social world:

> Indeed, in some autistic people this sense of radical and ineradicable differentness is so profound as to lead them to regard themselves, half jokingly, almost as members of another species ('They beamed us down on the transporter together,' as the B.s liked to say), and to feel that autism, while it may be seen as a medical condition, and pathologized as a syndrome, must also be seen as a whole mode of being, a deeply different mode or identity, one that needs to be conscious (and proud) of itself.
>
> (p. 264)

The B. family not only have a 'form of life', but a *shared* one. Within their family, they *are* selves among like selves. It is at least questionable, then, whether we should regard their differences from neurotypicals as pathological.

7.3.5 The self in autism: conclusion

I have here given an overview of how the characteristics of autism spectrum conditions may affect different cognitive processes that contribute to the self. As a result of ToM deficit, there may be no fluent and automatic self-ascription of mental states in mental-state *terms*—but there is still *awareness* of one's cognitive and affective states. Weak central coherence and different memory organization pose difficulties for narrative production and the 'autonoetic' awareness of episodic autobiographical memories. And the social impairments of autistic individuals will result in an unusual and impoverished social self.

It must not, however, be assumed that all these impairments are present in all individuals with autism spectrum conditions, or where present, always to the same degree. Autism spectrum conditions are diverse, ranging from high-functioning autistic individuals like Temple Grandin through cases of cognitively underdeveloped individuals with isolated savant abilities to individuals with classical autism, often non-verbal, who may be profoundly disabled.[20] Even among autistic individuals with similar degrees of cognitive idiosyncrasies, considerable variations exist in what cognitive capacities are particularly developed or underdeveloped—just as such variations exist between neurotypical individuals. There is, therefore, no one paradigmatic 'autistic self'. There are, rather, various self-representational capacities that may be impaired—or operate differently—in autism.

In the next chapter, I will look at other defects and disorders and their relation to memory, narrative capacity, and the self.

[20] Thus, the usefulness of 'autism' as a diagnostic category may be questioned (Hassall, 2016).

Chapter Eight

Other defects and disorders

8.1 Introduction

My target in much of this thesis has been historical-constructionist theories of the self, which attempt to define the self in terms of either autobiographical memory (ch. 4) or narrative (chs. 5 & 6). Against these, I have argued in favour of a view of the self as a complex, dynamic self-representational system. By way of illustrating my case for the system view, after autism spectrum conditions in the previous chapter, I will now discuss a number of psychopathologies that affect memory and/or self. This discussion will demonstrate the advantages of the system view over historical-constructionist accounts of the self.

Against the Lockean intuition that a loss of memory causes a loss of self, clinical case studies of memory deficiencies suggest that some elements of self persist despite the loss or impairment of explicit memory. Different memory impairments correlate with impairments of different self-representational capacities. Explicit autobiographical recollection is only one such capacity, perhaps not the most important one (§ 8.2).

Other 'self disorders'—split brains, bipolar disorder, multiple personality disorder, schizophrenia—also show disruptions of different self-monitoring processes. Tempting though it may be to speak of such conditions as resulting in multiple or dissolving selves, the self-representational capacities disturbed in each are distinct and quite specific, while other parts of the self continue to function. Moreover, these disturbances of the self are not caused by defective narrativity: narrativity may be unimpaired, or if it is impaired, this is a *consequence* of the disorder (§ 8.3).

I'll conclude by showing how the system view can account for these different psychopathologies as *partial* disturbances of the self, while ac-

knowledging their specificity and the unimpaired operations of other functions of the self (§ 8.4).

8.2 Memory defects

As noted in Chapter 4, autobiographical memory is complex and varied. Besides *episodic* recollections of the events of our lives, we have *semantic* autobiographical knowledge, both of the *facts* relating to our biographies, and of our (self-image of) personality traits (§ 4.2.5). We have, in addition, a capacity for *recognizing* ourselves in mirrors and images. Self-recognition and the various kinds of autobiographical remembering are all self-representational processes—activities of the self as system.

If one were deprived of these capacities, would one thereby lose one's *self*? There appears to be a widespread quasi-Lockean intuition that autobiographical memories are necessary for the self to persist (ch. 3). Yet, that intuition is triggered by thought experiments or experimental philosophy vignettes in which the usual persistence of memories (and self) is contrasted with a *complete* (and usually sudden) loss of memories (and sometimes all other psychological states).

It would now be useful if there were empirical studies of complete memory loss to corroborate or disconfirm the supposition that such a condition should result in a loss of self. But, as we shall see, memory loss in amnesias and dementias is never quite complete. While many amnesias are sudden in onset, their effect is mostly on a circumscribed aspect of memory; dementias are more global in their effects, but progress gradually, and may spare isolated self-representational mnemonic capacities. I will shortly look at examples of amnesias (§ 8.2.2) and dementia (§ 8.2.3) and their implications for patients' selves. But first, let me discuss an intriguing case of complete lack of *episodic* memory.

8.2.1 'Severely deficient autobiographical memory'

In April 2016, *Wired* magazine ran a story about Susie McKinnon, who has no episodic recall at all, and never has had that capacity (Hayasaki, 2016). McKinnon leads an otherwise unremarkable life: she is happily married,

has a steady job, enjoys holiday cruises, and sings in a choir. She has perfectly normal *semantic* self-knowledge: she knows who she is, her personality traits are no less stable than those of others, she has factual knowledge of things she has done in the past—but 'none of it bears a vivid, first-person stamp'. She also cannot date even recent holidays more accurately than to within a decade. Having first been made aware of her deficiency as a teenager, she has not found it a serious impediment in her personal or professional life. Indeed, she seems to have no desire to compensate for the lack of episodic memory with photography or by having a 'timeline' on social media websites: 'the life-logging impulse is lost on McKinnon'. She seems to lack what Klein et al. (2004) call a 'sense of personal temporality'. More precisely, what is missing appears to be the *constructive* capacity that underwrites episodic recollection (§ 4.4.1): just as she has no episodic memory, McKinnon also does not engage in episodic thinking about the future, or any kind of imagining: 'She does not daydream. Her mind does not wander' (Hayasaki, 2016). And thus, while she enjoys reading novels and watching films and television serials, she has no *narrative* capacity of her own. She also cannot recall the narratives of books or films she has read or seen.

McKinnon is one of only three reported cases so far with what has been labelled 'severely deficient autobiographical memory (SDAM)' (Palombo et al., 2015). None of these cases had any 'history of birth complications, seizures, stroke, traumatic brain injury or neurological disease . . ., nor was there evidence of psychological trauma', and, 'at the time of testing, there was no evidence of depression or other psychopathology' (p. 107). All three showed reduced activity in midline brain regions in fMRI scans during a personal recall task, notably in the left medial prefrontal cortex, 'associated with the mental projection of the self through time', and the right precuneus, 'associated with visual memory' (p. 111). Additionally, in structural MRI, they all showed a significant asymmetry of the hippocampus, with the right hippocampus, which has been associated with visuospatial processing, being 6–9 % smaller in volume than that of controls (pp. 109, 112, 115). The SDAM cases also performed significantly worse than controls on a visual memory task. All this suggests that normal episodic recall makes strong use of visual memory circuits. The absence of episodic memory combined with poor visual memory in SDAM thus seems the opposite of

the strongly visual cognitive style and highly detailed episodic recall of Temple Grandin discussed in the previous chapter (§ 7.3.3).

What, then, of Susie McKinnon's *self*? There is no chance that McKinnon could ever satisfy the criterion of a *Lockean* self: the re-living of past events in autonoetic consciousness. Neo-Lockeans, meanwhile, would diagnose a severe lack of 'psychological connectedness' over time, as well as seriously impaired 'psychological continuity'. As she has no narrative capacity, narrative constructionists and defenders of essential narrative-capacity views would also have to exclude her from having a self. But while the constructive capacity for episodic recollection and narration is absent, McKinnon's semantic autobiographical knowledge and her self-awareness in the present are functioning. Thus, her case illustrates perfectly how misguided the historical-constructionist project is in defining the self in terms of episodic memory and/or self narratives.

McKinnon's inability to form episodic memories is not an acquired condition; it is not a case of a previously available capacity being lost. In this respect, and in others, SDAM differs from cases of amnesia, to which I now turn.

8.2.2 Amnesias

The term 'amnesia' covers any loss of memory of any severity and duration. Amnesia usually correlates with brain lesions resulting from head trauma, stroke, tumours, infectious disease, or toxins. As the aetiologies vary, so do the affected brain areas; and with them vary the aspects of memory that are lost or impaired. More important, the presentation of memory loss varies in scope and severity, as well as in its temporal direction: retrograde amnesias are characterized by a loss of memory of events before the onset of the condition, while anterograde amnesias are marked by an inability to form new memories post-onset. Retrograde and anterograde amnesia can also co-occur.

Alan Baddeley and Barbara Wilson (1986) tested the autobiographical memory recall of patients with different forms of anterograde amnesia and found 'qualitative differences . . . that were not predictable on the basis of the more traditional memory tests' (p. 229). These differences were observed across the dimensions of fluency, episodicity, richness of detail, and

reliability. Overall, Baddeley and Wilson record four 'patterns of deficit' (p. 234): first, relatively unremarkably, a number of patients showed normal recall of autobiographical memory for events before the onset of their anterograde amnesia. A second group of patients were afflicted with what Baddeley and Wilson call 'clouding' of autobiographical memory: they were able 'to recall events in some detail on one occasion but not subsequently' (p. 235). These patients showed confusion but did not confabulate. Finally, four patients with frontal-lobe damage exhibited two further patterns of impairment present to differing degrees in different patients: non-fluency of recollection, and confabulation. While confabulation is common in frontal-lobe patients (see § 4.4.4) it is not present in all of them; nor are non-fluency and confusion present in equal measure in all cases of frontal-lobe damage.

Baddeley (1982) distinguishes between processes of *retrieval* and of *recollection* of autobiographical memories, where recollection is a second-order 'active process of setting up prospective retrieval cues, evaluating the outcome, and systematically working toward a representation of a past experience that we find acceptable' (p. 712). In most amnesics, it is the *retrieval* mechanism that is impaired, making recollection impossible: this is the case with Baddeley & Wilson's (1986) first group of patients, where the onset of anterograde amnesia marks the point beyond which no memories are retrievable. In contrast, in the frontal-lobe cases, the issues with fluency and confabulation suggest that some disorganized *retrieval* occurs, but that *recollection* (in Baddeley's sense) is disrupted, affecting 'the capacity both for directing memory retrieval and for evaluating output' (p. 249). This impairment of *executive control* over memory retrieval and assessing the veracity of reported memories is consistent with other features of frontal-lobe defects that show lack of executive control, such as distractibility and difficulties initiating and interrupting behaviour. In frontal-lobe patients, then, such disruption of self as may be inferred from the disruption of autobiographical recollection is a defect of executive control, rather than a defect of memory itself. As Baddeley and Wilson observe,

> This leads to the general question of how normal subjects verify their memories. The fact that we make mistakes indicates that this process is far from perfect, but the fact that most of our mistakes are plausible ones suggests that we do cross-check the results of our retrieval processes. (p. 249)

This is another point in favour of the priority of the self *qua* self-monitoring system over autobiographical memory. Executive control clearly is a self-monitoring process, and an important one at that: the prefrontal cortex is the closest thing to a 'central processing unit' for higher-level cognition that we have in our brains. Defects in decision-making, in planning and monitoring complex goal-directed behaviour, as well as in monitoring our memory reconstructions, are all associated with prefrontal lesions (Damasio, 1994).[1] And it is a defect in this higher-level self-monitoring which here leads to confabulation and defective memory—rather than defective memories leading to a defective self.

William Hirst's (1994) review paper 'The remembered self in amnesics' supports the view that the self depends on a number of processes whose disruptions are dissociable and not uniform. Experimental studies with anterograde amnesics show that although there is generally a failure to retrieve memories of post-onset events, memories are nevertheless being encoded and can sometimes be accessed when given the right cue, though without temporal and other contextual information. Recognition is somewhat preserved in amnesics (amnesics perform worse than control subjects but better than chance in recognition trials), as are skill learning and priming (amnesics perform as well as controls in implicit, but not explicit, word completion tasks). Additionally, amnesics retain 'islands of preserved memory' (p. 264) such as family events (perhaps for their emotional valence or because of frequent rehearsal of these memories—cf. § 4.3.1).

Hirst also notes that the *implicit* memories of amnesics are still operative in governing behaviour. For instance, Hirst mentions the case of a former executive with both retrograde and anterograde amnesia, who has no explicit recollection of his former occupation but whose demeanour remains that of one with the role and position he occupied. Thus the self, Hirst suggests, is apparent not only in individuals' explicit memories of their lives, but also 'in the amnesics' behaviour, and in their physical environment' (p. 272):

> Our patients had a sense of their place spatially, temporally, socially, professionally, and emotionally. Moreover, they could talk about what they liked to do . . . and they planned their lives around these activities. (p. 271)

---

[1] I'll return to executive control in my concluding chapter (§ 9.2.1).

Hirst also emphasizes the role played by others—by the social context—in supporting his amnesic patients' selves. He suggests that their impairment of memory functions is to some extent compensated for by a process we may call *cognitive outsourcing*: memories are 'externalized, collectivized, and eventually internalized' (p. 273), if only implicitly. (Such cognitive outsourcing also occurs for individuals with dementia—see § 8.2.3 below.)

The complexity of the effects of amnesia on the self is further illustrated by two cases from Oliver Sacks's (1986) clinical memoirs. Both are of patients with Korsakoff's syndrome, a disorder in chronic alcoholics characterized by anterograde amnesia, confabulation, and disorientation (Sutherland, 1995). One case, Mr Thompson, had profound anterograde amnesia, with recall of new experiences limited to less than a minute, for which he compensated with continuous confabulation, a 'narrational frenzy' (Sacks, 1986, p. 106) in which he mistook Sacks variously for old friends or customers of his former delicatessen, recruiting elements of pre-amnesia autobiographical knowledge apparently at random. He was able to recognize his brother but treated him with indifference, as he did the characters of his confabulations. Overall, Sacks describes him as seeming ill at ease and bewildered, except when alone in the garden: 'one never feels, or rarely feels, that there is *a person* remaining' (p. 110, original emphasis). Here then, we might say, is a case of a seriously impaired self. But what Mr Thompson lacks is not simply autobiographical memory—for his pre-onset memories are available and drawn on in his confabulations—nor narrative capacity—which if anything is present to excess. Rather, because of the anterograde amnesia there is a lack of *temporal* positioning and of the ability to *integrate* available memories coherently.

Sacks's other Korsakoff's patient, Jimmie G., presented somewhat differently. He did not confabulate, but in addition to the typical anterograde amnesia also had retrograde amnesia eclipsing about 25 years before the onset of his disorder. Of his life before that period, he had intact and detailed autobiographical memories, on which he was able to draw in conversation. He had intact spatial recognition, was able to recognize his own handwriting, and learned to recognize the nursing staff around him (though misidentifying them with people he had known in his pre-amnesia days). But as a result of his amnesia, like Mr Thompson, he was 'isolated in a single moment of being' (p. 28), without temporal positioning and having

no access to (by the time he became Sacks's patient) three decades of autobiographical memory. The question then, as Sacks himself puts it, is 'what sort of a self can be preserved in a man who has lost the greater part of his memory and, with this, his past, and his moorings in time?' (p. 22)

Sacks's own tentative first answer to this is that his patient had become a '"Humean" being' (p. 28), one whose self truly is 'nothing but a bundle or collection of different perceptions' (Hume, 1739/1978, I.iv.6). But this 'diagnosis' seems to hold only for the patient's *explicit cognitive* functions. There are *implicit* cognitive abilities—such as recognition—that are unaffected by his amnesia. And there are *affective* states and traits which may well mark one as a particular individual. Sacks (1986) quotes from his correspondence with Alexander Luria, who reminds him: 'a man does not consist of memory alone. He has feeling, will, sensibilities, moral being' (p. 32). Sacks also reports witnessing Jimmie attending divine service with 'an intensity and steadiness of attention and concentration that I had never seen before in him or conceived him capable of' (p. 36). Further, Sacks's initial description of his patient abounds with predicates of persistent (and pleasant) character traits—Jimmie G. is 'charming', 'cheerful, friendly, and warm' (p. 22), 'genial' (p. 23).

Like the other cases of amnesia discussed, the case of Jimmie G. suggests that the characteristics that make one *who* one is are not exhausted by the availability of explicit autobiographical memories. There is no denying that a large part of Jimmie G.'s autobiographical knowledge has been lost. But equally it is obvious from Sacks's clinical portrait that some elements of his patient's self—his native intelligence, his character traits, and what we may call his *affective* self—persist.

8.2.3 Dementia

Dementia is defined, generally, as 'an acquired and irreversible deterioration of intellectual function' (Marcovitch, 2010, p. 179). Thus, dementia does not merely affect memory, but (as the term's etymology implies) 'mental' capacities generally. This might suggest that *as well as* those components of self linked to autobiographical memory, *other* elements of self—such as those apparently spared in localized amnesias—should deteriorate in cases of dementia. Once again, however, the clinical literature paints a more

complex picture. I will here first consider a couple of case studies suggesting that some self-related memory functions persist even in advanced stages of dementia. Then I will discuss, more broadly, how the effects of dementia on the self may be characterized.

As discussed earlier (§ 4.2.4), knowledge of one's character traits appears to be dissociable from other semantic personal memory. As Klein, Cosmides, and Costabile (2003) note, such trait self-knowledge is 'surprisingly resilient in the face of brain damage and developmental disorders' (p. 158). Is it resilient in the face of dementia? Klein and colleagues examined this in the case of patient K.R., a 76-year-old college-educated woman with severe Alzheimer's dementia, who suffered from disorientation, had difficulties finding words and naming objects (i.e. seriously impaired semantic memory), had severe anterograde amnesia, and retained only a very sketchy knowledge of her personal past before the onset of her dementia. But when tested on questions about her own personality traits, she produced reliably consistent answers between a first test and a re-test. Though her answers did not agree with her daughter's and her carer's assessment of her *current* personality, they did agree with her daughter's ratings of K.R.'s *pre-onset* personality. Further, K.R.'s ratings of her daughter's personality 'correlated strongly with her daughter's self-ratings' (p. 161). On the other hand, her ratings of the carer's personality traits were not reliable compared with the carer's own self-ratings and those by age-matched healthy control subjects. Thus it appears that K.R. retained pre-dementia trait self-knowledge, which had, however, not been *updated* with the changes in her personality that had occurred since the onset of her dementia. Analogously, she retained pre-dementia knowledge of others' personality traits (daughter) but had not formed an accurate trait knowledge of post-onset acquaintances (carer). K.R.'s trait knowledge of self and others was thus, as it were, archived at the time of the onset of dementia and remained accessible thereafter, even as other personal memories before and after onset became inaccessible.

Hehman, German, and Klein (2005) observed a similar effect with respect to self-*recognition* in another dementia patient, P.H., an 83-year-old woman in the late stages of Alzheimer's disease (which had been diagnosed about seven years previously). P.H. was presented with 14 photographs of herself (two each from seven decades of her life, all presented in

monochrome and with context removed) on two occasions and asked whether she recognized the person in the picture. With images from her twenties and thirties, she recognized herself in 7 of 8 presentations. For the remaining decades, she only recognized herself, and with uncertainty, in 2 of 20 image presentations, and mistook a further two photos for one or other of her sisters. Hehman and colleagues conclude that, like K.R.'s trait knowledge in the other case study, 'P.H.'s self-recognition is frozen in time' (p. 121). But, unlike in P.H.'s case, the 'freezing point' of K.R.'s self-recognition appears much earlier than the onset of her dementia—it was in the oldest photographs that she most assuredly recognized herself, while there was not much difference between her responses to images from her forties and her eighties.

> Taken together, these findings demonstrate that aspects of self-knowledge (i.e., personality traits, facial appearance) that have been documented in clinical descriptions as degraded as a result of Alzheimer's disease, might degrade in such a way as to leave vestiges of earlier representational states. This idea is consistent with the proposal that Alzheimer's disease might carry with it impairment in routines that *update* various databases of self-related knowledge. (p. 122, emphasis added)

Thus, in Hehman and colleagues' analysis, while there is degradation of self-knowledge in Alzheimer's dementia concurrent with the loss of general semantic memory, *some premorbid self-knowledge persists, but is no longer updated*. In this respect, the presentation of Alzheimer's disease resembles the cases of anterograde amnesia discussed earlier.

While the cases of P.H. and K.R. suggest that memory and self-knowledge rely on dissociable components that do not deteriorate equally in dementia, the preserved elements of self-knowledge are nonetheless limited in comparison with the overall degradation of cognitive function observed, involving the progressive loss of memory, of the ability to communicate, and eventually of recognition, along with secondary effects such as changes in personality, confusion and disorientation. The experience of the outward effects of this cognitive deterioration by others, particularly close family and friends of those with dementia, has led to a popular conception of dementia as a progressive 'loss of self' (Aquilina & Hughes, 2006). And if the self is a system that recruits various cognitive capacities, a general progressive deterioration in cognitive function may well be expected to affect

most or all of them sooner or later, even if, as suggested by the studies of K.R. and P.H., this occurs at different rates for different components of the self.

But there are two qualifications still to be made here. One is that the 'loss of self' observed by others is informed by the overt behaviour of dementia sufferers, which is affected by their 'declining ability to communicate' (Small et al., 1998)—leaving the possibility that a patient's inability to communicate may mask his or her remaining cognitive capacities. The other qualification is that a dementia sufferer's self may be affected by care practices, specifically by his or her being 'positioned' (Sabat, 2006) as nothing but a helpless patient whose every behaviour is interpreted as being a result of the disease.

On the first of these points, a number of studies conducted since the 1990s show that though communication is impaired in dementia sufferers, what utterances they do make contain indicators of a functioning sense of self even in the late stages of Alzheimer's disease. Along with 'lucid periods . . . when the patient was the subject of individual attention, or . . . triggered by strong emotions in the context of, for example, music and prayer' (Aquilina & Hughes, 2006, p. 145), these indicators include the use of first-person pronouns and proper nouns, and patients' self-referential statements 'about themselves, their needs, and concerns' (Tappen et al., 1999, p. 123). Where such a sense of self can no longer be expressed verbally, it may still be apparent in patients' interactions with care staff:

> For the [care home] residents who did not use first person pronouns, self was, nevertheless, indexed in other ways. These residents were frequently involved in conflicts in which they defended their rights as an individual. In these conflicts, their awareness of and resistance to the violation of their desires by others was a clear expression of an intact self.
>
> (Small et al., 1998, p. 309)

As Steven Sabat (2005) puts it, a person with dementia thus remains a 'semiotic subject', that is, 'a person whose behaviour is driven by meaning' (p. 1030), one who has desires, acts with intention, interprets and evaluates and responds to events, situations, and others' actions. Conflicts arise because the person with dementia is unable to articulate his or her desires and intentions. Aquilina and Hughes (2006) take this as indicating

that 'there is an inner self, which cannot communicate with the carer', making dementia sufferers in effect 'mental prisoners' (p. 145).

But there is a caveat to this. If emotions, lexical and behavioural self-reference, and intentional agency are indices of a persisting self, its capacity for self-representation *over time* is nonetheless impaired by the loss of memory. For instance, Small et al. (1998) record a conversation between a member of care staff and a care home resident in which the carer makes reference to the resident's former occupation as a carpenter—'you made buildings'—whereupon the resident seems astonished and utters 'really?' (p. 300)—suggesting that his sense of self no longer, at that point, involved any such autobiographical content.

This is not to say, however, that references to a dementia sufferer's past occupations cannot be useful. And that brings me to my second qualification of the progressive loss of self in dementia stipulated earlier. Some elements of our self-representational activity emerge from our social interactions and how one is 'positioned' in interactions with others, that is, what social roles are ascribed to one by others and, interactively, by oneself. I will refer to these aspects of self-representation collectively as the *social self*. (Other terms in the literature include Rom Harré's (1991) *self*$_2$, defined as 'the selves that are publicly presented in the episodes of interpersonal interaction in the everyday world, the coherent clusters of traits we sometimes call "personae"' (Sabat & Harré, 1992, p. 445),[2] and Aquilina & Hughes' (2006) '*outer* self' (in contrast to their aforementioned 'inner self'), 'the public observable aspects of self, which depend on psycho-social structures including social relations, culture, and language' (p. 150).)

The social self depends on social interactions and these, largely, depend on interpersonal communication. Where one party's communicative abilities are impaired, as in the case of people with dementia, the onus of supporting their social selves shifts to those around them—family and friends, carers, and clinical staff. This is the common theme that emerges from studies by Sabat and Harré (1992), Small and colleagues (1998), and Tappen and colleagues (1999). In clinical and residential-care settings, people with dementia are generally treated as entirely dependent, and all

---

[2] In Sabat's (2005) revised notation, which also distinguishes between 'the self of personal identity' and 'the self of mental and physical attributes', this social self becomes 'self$_3$'.

their behaviours are interpreted as resulting from their disease. Sabat (2006) refers to this practice as 'malignant positioning' (p. 289) (where 'malignant' means 'harmful' rather than 'with malicious intent'); dementia sufferers 'can then come to see themselves in progressively more defective terms and lose a sense of self-worth' (p. 290). Besides this ethical implication, positioning the dementia sufferer as an entirely dependent patient forecloses the preservation of his or her social self.

As Small and colleagues (1998) point out, different care practices may produce different results:

> Caregivers can help preserve the personae of residents by cooperating in the co-construction of the residents' preferred personae. This will require taking a personal interest in the residents' background . . . One might argue that there is little reason for staff to talk with nonverbal residents about their (residents' or staff's) personal lives since it would be a one-way conversation. This perspective, however, fails to take into account the retained receptive abilities of many nonverbal demented residents. The fact that residents may not respond verbally does not mean that they do not understand. (pp. 312–3)

It is also possible for a still verbal dementia sufferer actively to solicit others' co-operation in supporting his social self. A rather touching example of this is provided by Sabat and Harré (1992). One of their subjects, J.B., a retired academic, still able to communicate but often finding it difficult to put into words what he was trying to say, evidently took his involvement in Sabat's research as continuous and of a kind with his own previous scientific career. It became apparent that he wished to obtain some 'tangible' recognition of his contribution. Sabat's response to this is worth quoting at some length as it illustrates what is meant by positioning a person with Alzheimer's dementia as other than just a patient, and thereby supporting his social self:

> Once the nature of J.B.'s wish to link this last spurt of academic effort with the final stage of his moral career became clear it was arranged for him to have a letter of commendation from the Dean of the College of Arts and Sciences. In that letter, J.B. was commended for giving of himself unstintingly to help with the investigation of the abilities that remain intact in spite of A.D. [Alzheimer's disease] . . . Upon receiving the letter, his wife made copies and he, then, brought a copy to the day care centre where it was read aloud to the entire group of participants. In addition, when his adult children visited, he showed them the letter with great pride, for it

> signalled that he was, indeed, doing something important. Thus, . . . the academic $self_2$ was jointly constructed once again. By virtue of the social force of the letter of commendation J.B. was positioned, not as a helpless and confused A.D. sufferer, but as one who had a contribution to make to science even in the throes of A.D. (p. 455)

Of course, different dementia sufferers will have different 'preferred personae', and, depending on individual circumstances and the progress of the disease, not all will respond in like manner to attempts by carers at reconnecting them with their occupational past (cf. the example of the carpenter in Small et al. (1998) cited earlier). In some cases, treating patients' present needs *without* regard to their past can be advantageous as well as problematic; and sometimes the preservation of a patient's personhood may occur only by proxy, through a spouse or close family member (Oppenheimer, 2006)—another example of the cognitive outsourcing described earlier (§ 8.2.2). But the point is that such social interactions as are possible in the various stages of dementia still shape and shore up a sufferer's social self. What form these interactions take makes a crucial difference—whether the person with dementia is positioned as merely a patient, or as a person (Sabat, 2006). The right kind of engagement with dementia sufferers may well amount to a 'workout' for their selves.

8.2.4 Memory defects and self: conclusions

Though the complete and immediate erasure of memory and other psychological states that is the staple of thought experiments on the persistence or otherwise of self is not encountered in actual clinical cases, the quasi-Lockean intuition was that any noticeable loss of autobiographical memory should entail a proportionate loss of self. But the loss or inaccessibility of explicit autobiographical memory in amnesias and dementia, or the lack of episodic memory in SDAM, are not the *causes* of a defective self. Rather, they are *symptoms* of the deterioration or absence of a particular self-representational capacity.

Moreover, memory is not monolithic. There are, rather, areas of memory that are preserved in these conditions: recognition including self-recognition, procedural memory, trait self-knowledge, and other 'implicit'

memories.[3] Furthermore, there are other domains of the self, such as character traits, decision-making and a sense of agency, and affective dispositions, that are not neatly classifiable as memory-related and that seem persistent even in severe memory disorders. In sum, then, the pathologies of memory once again suggest that it is the self that makes its memories, and not memory that makes a self.

8.3 Divisions, dissociations, dissolutions

While the previous section concerned the question whether disorders of memory entail a disturbance of the self, I now turn to some psychopathologies in which, conversely, a disturbance of self seems to form an inherent part of the presentation and diagnosis of the disorder. There are two caveats to this discussion. Firstly, 'disorders of self' denotes a very heterogeneous group of psychopathologies that vary according to their presentation, their severity, and their aetiology (where known). What they relevantly have in common is that their sufferers experience *some form of disturbance of self*; but their discussion under this heading should not be taken to imply a biomedical taxonomy.

Second, since the criterion 'disorder of self' allows a great many variations in its precise instantiation, there is, consequently, a large number of psychopathologies that meet this criterion. They include severe depression, bipolar disorder, borderline personality disorder (and other so-called personality disorders), post-traumatic stress disorder, depersonalization and derealization, various kinds of dissociative disorders (including dissociative identity/multiple personality disorder), and schizophrenia. For economies of space and argument, I cannot here discuss all these disorders. I will therefore limit my discussion to three of them, which have sometimes been characterized as cases of division, dissociation, and dissolution of the self:

- *Bipolar disorder*, which is characterized by the abrupt alternation of two distinct and opposite mood states;

[3] Long-term musical memory is particularly resilient in dementia (Jacobsen et al., 2015).

- *Multiple personality/dissociative identity disorder*, which has informed and inspired a number of philosophical discussions on the self; and
- *Schizophrenia*, which may appear to involve a dysfunctional self–world boundary.

But before that, let me discuss a surgically acquired condition that has for some time been something of a philosophers' favourite: *split brains*.

8.3.1 Split brains

I have already mentioned commissurotomy cases, whose 'split-brain' condition results from the resection of the corpus callosum, the main commissure linking the cerebral hemispheres (§§ 1.7, 6.3.1). Such patients exhibit laterally disconnected sensory and motor systems in experimental conditions:

> What is flashed to the right half of the visual field, or felt unseen by the right hand, can be reported verbally. What is flashed to the left half field or felt by the left hand cannot be reported, though if the word 'hat' is flashed on the left, the left hand will retrieve a hat from a group of concealed objects if the person is told to pick out what he has seen. At the same time he will insist verbally that he saw nothing. (Nagel, 1971, p. 400)

The reason why only perceptions in the *right* visual field or with the *right* hand can be reported *verbally* is that linguistic capacities are processes entirely in the *left* cerebral hemisphere, which also—contralaterally—processes sensory input and motor output to and from the *right*-hand side of the body. Lacking input from the right hemisphere and therefore the left-hand side sensory organs, the speech centres are unable to process that information, though other responses (the left hand picking up a hat) indicate that such information is nevertheless processed cognitively in the right hemisphere.

Studies of split-brain patients have two main implications. One is that the sensorimotor system of each cerebral hemisphere, and some higher-level cognitive functions (like language, in the left hemisphere), operate independently of the other hemisphere. The other is that while the cerebral hemispheres are symmetrical anatomically and with respect to sensorimotor function, they are not, in humans, *functionally* symmetrical. Certain cognitive functions are strongly *lateralized*: language and speech, hypothesis

formation, and pattern recognition are left-hemisphere functions, while the right hemisphere is specialized for face recognition, perceptual distinctions and grouping, and focusing attention (Gazzaniga, 2000).

Now, it might be suggested that split-brain patients therefore end up with, literally, two brains: a left brain that controls the sensorimotor systems for the right-hand side of the body, along with the cognitive capacities of the left cerebral hemisphere; and a right brain controlling the left-hand-side sensorimotor processes and the cognitive operations of the right hemisphere. But, although somatosensory functions are largely lateralized, split-brain patients can still execute co-ordinated movements bilaterally, e.g. walking, or waving both their arms synchronously (Gazzaniga, 2000). Nor is sensory input processed wholly contralaterally—some visual information reaches the ipsilateral visual cortex. Thus, isolating the left and right visual fields in experimental conditions in such a way that only one cerebral hemisphere receives input requires a rather complicated experimental set-up designed expressly for that purpose—a situation no split-brain subject is likely to encounter in ordinary circumstances. Thus, even in the split-brain condition it would be wrong to assert that the patient's brain had been *entirely* bisected.

Notwithstanding this, Roland Puccetti (1973) suggested not only that split-brain patients have two 'minds' and are therefore *two persons*, but that, given the cognitive specializations of the cerebral hemispheres revealed by split-brain cases but present in all humans, we are *all* 'compounds of two persons' (p. 353). He cites cases of individuals living with only one cerebral hemisphere (congenitally or postoperatively) as evidence that each hemisphere, in normal cases, already amounts to a full person's brain.

But the lateralization of brain function in humans means that our species, uniquely, does not duplicate all brain functions across both hemispheres (Gazzaniga, 2000). Therefore, a single hemisphere is precisely *not* fully equipped to discharge the functions of a whole brain. While indeed it is possible to live with only one cerebral hemisphere, patients in that condition are somewhat disabled. And where one hemisphere, post-injury or post-hemispherectomy, is able to take over some functions of the other, it can do so thanks to the remarkable neuroplasticity of our brains, which is then activated to make up for an acquired loss of function—and not because that hemisphere exercised the missing functionality all along.

Most crucially, such disconnection between cognitive processes in opposite cerebral hemispheres as is exhibited by split-brain patients is precisely what one would expect from their condition. It is because they lack the communication lines of the corpus callosum that some of their cognitive processes are not integrated in the normal way. But in non-split-brains the corpus callosum ensures that cognitive integration between hemispheres; there *is* no disconnection between the 'left brain' and the 'right brain'. Of course, such integration, even in healthy brains, is deployed *as needed*. As Thomas Nagel (1971) remarks, 'our own unity may be nothing absolute, but merely another case of integration, more or less effective, in the control system of a complex organism' (p. 410).

It is '*another* case of integration' because there are also *intrahemispheric* processes of integration at work, for instance integrating input from different sensory modes in the cortical association areas (Romero, 2006, pp. 114–116). And many other processes of integration are required for higher-level cognitive functions, if cognition is modular (Fodor, 1983). Thus, it may seem arbitrary to single out the corpus callosum as *the* defining integrative pathway: 'why then should we not regard *each* hemisphere as inhabited by several cooperating minds with specialized capacities? Where is one to stop?' (Nagel, 1971, p. 413). Indeed: once we begin considering cognitive processes, or modules, in isolation, we might think there are an indefinite number of 'minds' collaborating in one brain—echoing Hume's (1739/1978) 'republic or commonwealth' analogy (p. 261). But there is a good case for stopping at *one*, for, as Nagel suggests in the remark quoted earlier, it is in the running of *one organism* that cognitive integration and, crucially, *self-representation* are of service.

And that is as true of split brains as it is of brains with an intact commissure. For, as Gazzaniga (2000) observes, even in split-brain cases there does not seem to be an overall bisection of self-awareness:

> The most powerful impression one has when observing patients who have had their hemispheres divided is how unaffected they appear to be in their general cognitive awareness, affect and sense of self. (p. 1309)

Nevertheless, the general cognitive awareness and sense of self of split-brain patients can be faulty, because part of their self-monitoring system is of course interrupted by the severed corpus callosum. Thus, as already noted, patients cannot verbally report—verbal reporting being a left-hemi-

sphere process—on stimuli received only by the right cerebral hemisphere. Interestingly, they seem blithely unaware of this lack of input from the right hemisphere. As mentioned earlier (§ 6.3.1), when, in experimental settings, patients receive flashed instructions to the left visual field only (processed by the right hemisphere) and perform actions accordingly and are then asked for an *explanation* of their behaviour (processed by the left hemisphere), they confabulate: they give plausible, but palpably false accounts of their actions.

Gazzaniga (2000) attributes this confabulation to a left-hemisphere mechanism he calls 'the interpreter' (p. 1316), which constructs explanations and hypotheses of behaviour and perceptual information:

> The interpreter weaves together an interpretative story. Often enough, the story is actually correct, and the judgments and decisions attributed to the self are accurate. But sometimes the data are misleading or (as in the case of split-brain patients) absent altogether, and confabulation results.
>
> (Carruthers, 2010, p. 84)

According to Gazzaniga (2000), the interpreter may also be responsible for the 'feeling in all of us that we are integrated and unified' (p. 1316), including split-brain patients, who 'do not have any sense of the dual consciousness implied by the notion of having two brains' (p. 1319).

> The interpreter is the glue that keeps our story unified and creates and creates our sense of being a coherent, rational agent. . . . These narratives of our past behaviour seep into our awareness and give us an autobiography. (p. 1320)

For a defender of historical-constructionist account of the self, the interpreter would be something like the originator of the self *qua* narrative construct. On the system view, too, the interpreter clearly has an important role in sustaining one's sense of self. However, we shouldn't jump to the conclusion of identifying the interpreter with the self. For, in order to provide a veridical integrative function of cognitive self-awareness, it requires input from other self-representational processes all over the brain—half of which are what is missing in split-brain patients. So the interpreter is a component of our complex self-monitoring system—an important component in that it provides integration of information and acts as an output mechanism. But it neither *is* the self nor produces it. It is *part* of it.

As for split-brain patients, they clearly do not have two selves under either view of the self discussed here. For an historical-constructionist account, only the products of the interpreter would count as the self, and the interpreter simply ignores the processes of the right hemisphere. On my system view, the self of a split-brain patient is defective to the extent that self-representational information from the right hemisphere cannot be integrated in the left-hemisphere interpreter's productions, but, as the right hemisphere does not have an equivalent interpreter capacity, nor can there be systemic integration of this information within the right hemisphere to combine into a second, independent self-monitoring system. Thus, rather than producing two selves, the split-brain condition results in one self that is deficient to the extent that there is no interhemispheric integration of self-representational processes.

While the split-brain condition results from a clear division of the brain, other conditions in which one might be tempted to diagnose a divided self are less clear cut (to coin a phrase). They include multiple personality disorder and bipolar disorder, which I discuss next.

8.3.2 Bipolar disorder

Sufferers of bipolar disorder (more precisely bipolar *affective* disorder, sometimes also called bipolar *manic–depressive* disorder) oscillate between two extremes of mood: one being *mania*—'[p]athological over-excitement, often involving extreme agitation, excessive optimism, restlessness, flights of idea, and incoherent speech'—or *hypomania*—'in which the person is "mildly" manic; . . . too optimistic, too exuberant, too talkative, and too active' (Sutherland, 1995, pp. 261 & 212), and the other being *depression*. While 'mixed states' with both manic and depressive symptoms do occur (Anderson et al., 2012), the usual presentation of the disorder is of alternating manic/hypomanic and depressive phases, of varying duration (though usually of at least several days). The distinguishing characteristic of bipolar disorder is the spontaneous switching between these mood states, and while the precise neurobiological mechanisms of these mood switches are as yet poorly understood (Salvadore et al., 2010), their consequences are stark. A bipolar patient will alternate between a state of elevated mood and self-esteem, heightened activity, and excessive pleasure-seeking (with

sometimes adverse social and financial consequences), and a state of depressed mood and dejection, with a lack of interest in activity, fatigue, and feelings of worthlessness (Anderson et al., 2012).

What, then, are the consequences of bipolar disorder for the self? In my above discussion of amnesias, I noted that in the absence of explicit personal memories, a person's affective dispositions and character traits can remain. Let us call these dispositions and traits an individual's *affective self*. Then, it seems, a bipolar patient's affective self is fractured between the two poles of the disorder. And this fracturing obtains despite any psychological continuity that otherwise exists between a bipolar patient's depressive and manic/hypomanic phases. For although, when in a depressive phase, the patient may well *remember* being euphoric in the past and perhaps wish to regain that mood, this makes no difference to the now prevailing depressive mood. Indeed, given the reconstructive nature of *episodic* recollection (§ 4.4), it may well be impossible for a currently depressed bipolar patient to reconstruct accurately an episode from a preceding manic/hypomanic phase (and *vice versa*), for he now has no access to the affect states that accompanied the episode. He may of course remember *that* he was happy and euphoric. But the overpowering depressive mood of his current state will preclude his (as it were) 'reliving' the moment being remembered in the affective state he was in at the time.

Bipolar disorder, then, involves a fractured and discontinuous *affective* self while psychological continuity (in Shoemaker's sense) is intact. Now, it may seem very tempting to think of a bipolar patient as having two selves, particularly for the people around him or her, who cannot fail to notice the abrupt changes in personality that occur when a bipolar patient switches from one extreme mood to the other. For instance, Lloyd Wells (2003) describes the case of Dr Jones, a university professor with *rapid-cycling* bipolar disorder. One day he was depressive, pessimistic, fatigued; he found work difficult and social activities meaningless. The next day he rose early and began working immediately, was euphoric, hyperactive, a helpful colleague, and a great socializer until late at night.

> The cycle then repeated itself, every 2 days, over 8 months. Professor Jones hated his dysphoric days and loved his energetic days. . . . His wife felt differently about them. "He's not my husband," she said. "I like him fine, but he's a different person. He's so down every other day, but he's worse, for me, on the 'good' days. He's up at four in the morning, banging

> around in the kitchen and making a lot of noise. He was always very thoughtful and quiet when he got up before me in the past. And he's—he doesn't listen, and he tells jokes he would never have told before. He embarrasses me, a little, although everybody seems to think he's wonderful." (p. 298)

Dr Jones initially refused medication, as he was 'reluctant to give up on his good days'. Eventually, when, on his hypomanic days, 'he began to drink excessively at social events . . . he bought a summer house on a whim, and he nearly invested a lot of money in a scheme that would have had disastrous consequences' (ibid.), he consented to treatment with lithium, used as a mood-stabilizing drug.

> Almost immediately after therapeutic blood levels of lithium were attained, Professor Jones returned to his moderately morose, still highly achieving old self, which he thought was quite different from either of the cyclic states. He missed his high self very much, but thought that he was better off without it. He told me, "For the first time I understood Dr. Jekyll and Mr. Hyde." His wife said, "I have my husband back." (ibid.)

What is striking about this case is not merely the unusually rapid alternation between the depressed and the hypomanic mood states, but also the stark secondary effects of the mood switches on both Jones's professional and his social and marital life. So did bipolar Dr Jones have two selves?

Tempting though this way of putting things may have seemed to both the patient and his wife, such talk can only be metaphorical. Granted, the fracturing of his affective self was significant enough for both him and those around him to conclude that he embodied, if not strictly two persons, then two very distinct personalities (or three if we include his pre-onset and post-treatment 'moderately morose old self'[4]). However, Jones suffered no amnesia between alternating mood states, and was well aware of his sudden shifts of mood: there was a continuing self—disturbed and disrupted though it was in its affective states. That is not to say that Jones's self was not impaired by his bipolar disorder. But it was not split in two.

Further, few cases of bipolar disorder exhibit as rapid shifts between extremes of mood as Dr Jones's, and manic and depressive episodes of

---

[4] It is also remarkable how quickly mood-stabilizing medication appears to have put an end to his bipolarity (cf. § 8.3.5).

longer duration may alternate with more 'neutral' mood states, too. While in such cases bipolar patients' *affective* selves are disrupted and discontinuous when their mood state enters a new (hypomanic or depressive or 'neutral') episode, there is nonetheless a persisting self that *experiences* these mood changes, even if the *presentation* of a bipolar self to others may seem like two or three distinct personalities.

The supposed presence of two or more conflicting personalities in one individual is even more marked in cases of multiple personality or dissociative identity disorder, which I consider next.

8.3.3 Multiple personality/dissociative identity disorder

Multiple personality or dissociative identity disorder (MPD/DID)[5] involves the apparent alternating presence, in one individual, of a number of different 'personalities' or 'identities'. These alternative personalities, or '*alters*', exhibit various 'cognitive, sensory, and physiological differences' (Braude, 1995, p. 48), the last including differing facial expressions and postures, in some cases even the presence or absence of allergic reactions. Alters are often characterized by particular traits or personality types—'persecutor', 'helper', 'recorder or memory' (p. 40); frequently they involve at least one child alter, and in about one in two cases, a personality of the opposite sex. The number of alters an individual may have varies between about half a dozen and two dozen (Braude, 1995; Hacking, 1995a) but in some cases has

---

[5] The naming of the disorder is of some interest. The American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* substituted the term 'dissociative identity disorder' for 'multiple personality disorder' in its fourth edition (*DSM-IV*) in 1994 (Hacking, 1995b). While reasons for this change were mostly 'clinical and pragmatic rather than philosophical' (Braude, 1995, p. 249), they reflect a change in therapeutic approach (Hacking, 1995a, p. 54) laying more emphasis on a patient's singleness (however dissociated) than on the presentation of multiplicity. What the exact psychiatric distinction between 'identity' and 'personality' might be is rather more murky, and the current (*DSM-5*) edition's diagnostic criteria for 'dissociative identity disorder' continue to refer to 'distinct personality states' (American Psychiatric Association, 2013, p. 292). Meanwhile, the WHO's *International Statistical Classification of Diseases and Related Health Problems* retains 'multiple personality' in its current version (*ICD-10*), as a subcategory of 'other dissociative [conversion] disorders' (World Health Organization, 1992/2016, F44.8). I will here acknowledge the currency of both labels by using the somewhat ungainly shorthand 'MPD/DID'.

been numbered in the hundreds, though '[m]any of these are perhaps better described as personality *fragments*' (Braude, 1995, p. 41, original emphasis).

The onset of MPD/DID can usually be traced to childhood with the 'initial split' occurring 'between the ages of four and six' (p. 48). And psychological trauma or abuse in childhood seems to be one of 'two main causal determinants' of the condition, the other being 'a capacity for profound dissociation' or 'self-hypnotizability', which provides a coping mechanism for the psychological trauma suffered: 'through dissociation, the subject is able to avoid experiencing or dealing with an intolerable episode by turning it over to an alternate personality (or *alter*) who undergoes those experiences in his place' (p. 37).[6] The dissociation may then become habitual: once successfully deployed as a coping strategy, the mechanism may be deployed again and again, thus leading to a proliferation of alters.

Switching between alters can take 'one or two seconds to several minutes, and it can be voluntary or involuntary' (p. 41). The process of switching is characterized in the literature as one alter gaining 'executive control' over the patient's behaviour or body (p. 42). But what exactly is meant by one particular personality being 'in control' *of* the patient? Short of resorting to mediaeval notions of possession or dualistic theories of mind, it can only mean that the patient then exhibits the traits and characteristics, psychological and physiological, that are attributed to that particular alter. Rather than speaking of another alter *gaining control* of the patient's behaviour, it is perhaps better to describe the process as a *reconfiguration* of executive control, such that a different set of traits and characteristics is brought to the fore.

The occurrence of *voluntary* personality-switching supports this point, for it raises the question *whose* volition it is that brings about the switch. Is it that of the alter that is about to be displaced by another? This may seem plausible in cases where the displacing personality is of the 'helper' type,

---

[6] As some MPD/DID patients tend to reveal their 'multiplicity' under hypnosis, it has been suggested that the condition as a whole is iatrogenic, i.e. brought about by therapeutic intervention. But other MPD/DID cases have presented in the absence of hypnosis and therapeutic encouragement (Braude, 1995, p. 62). Nevertheless, a therapist's expectations and classifications may affect and change the presentation of the disorder (Hacking, 1995a, p. 21).

but hardly likely if the newly installed alter is of the 'persecutor' type. Nor does it sound altogether plausible in such a case that the 'persecutor' personality actively *takes over* executive control, for in order to do so it would already have to *have* some executive control—but we're told that the process of switching is precisely about a *change* in executive control. Voluntary switching between alters thus cannot readily be explained by means of the terminology of alters 'gaining executive control'. Rather, it suggests some underlying *unified* executive control.

This brings me to the crucial question: Does an MPD/DID sufferer have multiple *selves*? The first point to note here is that dissociation of personality need not *per se* entail a plurality of selves. Many of us dissociate some aspects of our personalities in certain circumstances: one might suppress some undesirable character traits in a professional context and others on a social occasion. These are instances of what Erving Goffman (1959) calls the '*presentation* of self', and they depend on social circumstances. In MPD/DID, too, it is often the social and professional context that prompts a switch from one alter to another:

> The more functional multiples often distribute activities between alters in creative ways, or at least in ways that do not command attention. For example, a college student might divide course work among personalities with the appropriate abilities (e.g., the alter good at math will take the statistics course). (Braude, 1995, p. 45)

It may seem, then, that the switching of personalities in MPD/DID is merely a particularly striking example of an individual's adopting different 'masks' (Goffman, 1959) for different purposes.

But the dissociation in MPD/DID goes somewhat deeper, in that different alters frequently do not share the same (explicit) memories. Indeed, the whole point of adopting different personalities for coping with traumatic experiences is precisely that one alter need not share the disturbing memories of another. On a quasi-Lockean account of selves, this would suggest a plurality of selves, each defined by the particular strands of personal memory it retains. This, however, would imply multiplicity of selves in other individuals who have temporarily forgotten an autobiographical episode. I have of course rejected the memory-based account of the self, but some of the reasons for doing so might still support a plurality of selves in MPD/DID. For instance, I mentioned the retention of character traits as in-

dicative of a continuing self in amnesia and dementia. But character traits are precisely what is dissociated between the different alters of an MPD/DID patient. However, as Stephen Braude (1995) points out, such dissociation is not complete. Alters overlap in their implicit and procedural memories, such as linguistic and other motor skills, and in some of their cognitive capacities, too:

> Generally speaking, then, what distinguishes one alter from another, and what gets dissociated in cases of MPD, is not a trait shared with no other alter or split-off completely from the rest of the multiple's activities. Rather, it is a distinctive *combination* of traits, any one of which might be shared with other alters. (p. 106, original emphasis)

All this suggests some underlying unity in the MPD/DID sufferer, a stock of common capacities to which different alters have selective access. Are these shared capacities sufficient for a single self? There remains the defining feature of MPD/DID: a 'marked discontinuity in sense of self and sense of agency' (American Psychiatric Association, 2013, p. 292). This, it seems, is no mere presentational exercise, but rather a case of experiencing oneself as *someone else* whenever one's executive control functions are configured for a different alter. Here it might seem as if narrative accounts of the self are, after all, suited to explaining this phenomenon.

In particular, Dennett's (1992) view of the self as a 'center of narrative gravity' (see § 5.3.3) readily accommodates multiple personalities as multiple *selves*. When an MPD/DID patient switches to a different alter, her narrative centre of gravity shifts, as it were, to a different protagonist. For this fictive account of the self, therefore, MPD/DID apparently poses no problem. For what I have called strong narrative constructionist accounts, the picture is slightly different: if the self is 'constituted' by a narrative, and if, in the case of an MPD/DID patient, each alter has its own narrative, there would indeed seem to be multiple selves—but all of them defective and incomplete, because each alter's self narrative would only cover the periods when it was 'in control' of the patient's behaviour.

But there remains the *authorship problem* (§ 6.2), which concerns both Dennett's account and strong narrative constructionism: *who*, in fact, is spinning the narrative? For narrative constructionists, the question is, presumably, empty: according to these accounts, there is no 'who' that can have authored the narrative until some initially disorganized strands of

narrative activity produce either a fictive protagonist or a coherent story that qualifies as a 'self'. No prior *single* self seems required. But consider this *cri de cœur* of an MPD/DID sufferer, Mary, reported by Lloyd Wells (2003): '"It is really hell to wake up every morning and hate the person who's there, who is not me."' (p. 301) Who in this case is 'me'; who is doing the hating? Narrative constructionists might respond that it is simply another centre of narrative gravity or rudimentary self narrative. But there is another and, given the apparent strong affect reported in this testimony, more plausible option, namely, an underlying self.[7]

Stephen Braude (1995, ch. 7) proffers a number of good reasons for hypothesizing an underlying self in cases of MPD/DID. First, as already mentioned, the alters of MPD/DID patients overlap to some extent in their traits and capacities. Secondly, he suggests that an underlying 'synthesizing self' (p. 167) is required to *maintain* the dissociation of personalities. This may seem counterintuitive. Braude does not dwell on this point, but what I take him to mean is this: a single monitoring system is required to *keep track* (consciously or unconsciously) of the roles of the multiple alters deployed in MPD/DID (e.g. Daniel is the 'helper' personality, Jerry the 'persecutor', John the 'memory', Carl drives the car, David is the one with artistic sensibilities, Stephen is good at maths, etc.). Thirdly, Braude makes the important point that MPD/DID patients frequently report conflicts between their alters. For this to occur, the alter currently online must have awareness of the alters that are currently offline, which again suggests that some (unconscious) integration between alters is required.

Considering the onset and aetiology of MPD/DID, Braude observes, fourthly, that the initial dissociation in childhood presupposes prior unity of self, and fifth, that dissociation as a coping strategy for dealing with unbearable psychological trauma 'seem[s] to make sense only with respect to a single agent' (p. 175), one who both experienced the trauma and deploys

---

[7] Another theoretical possibility might be that MPD/DID sufferers have a 'primary' personality which would be the one that, in this case, hates the personality that is currently online. But, as Braude (1995, § 2.5) notes, the recent MPD/DID literature makes little reference to 'primary' personalities, and nor is it clear whether such a personality would be understood as *historically* or *functionally* primary. Either way, for such a primary personality to be able to hate others it would have to be online persistently, along with whatever *alter* is currently online—thus, the notion of a primary personality would collapse into that of an underlying persistent self.

personality dissociation in order to cope with it. The self-monitoring system 'decides', as it were, that the trauma is too much to deal with, and institutes a compartmentalization—a dissociation—that segregates one's intolerable memories of, say, childhood abuse, from one's day-to-day cognitive operations. Further, '[i]t appears . . . as if alter-formation is intelligently guided, at least in the early stages before it becomes habitual' (p. 177). This observation is akin to the authorship problem with narrativity—here, the question is *who* is doing the alter-forming, and the acceptable answer is, once more, a persistent self. (As more splits occur habitually and successively over prolonged periods, it is not sufficient for there to be a single self only before the *initial* split.)

How, then, is the presence—indeed the necessity—of an underlying self to be reconciled with the experience of MPD/DID sufferers of themselves as different personalities? Braude suggests that the alters of an MPD/DID patient are distinct 'apperceptive centers' (p. 78), meaning that when acting as a particular alter, the patient does not believe or experience the psychological states of her other alters to be her own. This accounts for the distinctness, in the MPD/DID sufferer's experience, of her different personalities. What accounts for the unity required by the above considerations is, in Braude's view, a 'synthesizing self', by which he means essentially Kant's (1787) 'transcendental I', which though 'a precondition for the functional distinctness of alters', is not itself an *object* of experience but a *formal property* of it, whereas 'on a level of experience much closer to the surface of awareness, alters are distinct' (Braude, 1995, p. 188).

Unhelpful though Braude's Kantian approach and terminology may be, what I take to be the gist of his ideas fits very well with my system view. For all the dissociation between alters in MPD/DID, a persistent *single* self-monitoring system is required to account for (1) the overlap of capacities shared between alters, (2) the consistent tracking of different personalities, (3) conflicts between alters, (4) the initially undissociated self, (5) the efficacy of dissociation as a coping mechanism, and (6) the creation of new alters. The nature of the dissociation in MPD/DID, then, is not a fragmentation of this self-monitoring system as such, but rather its churning out, simultaneously or alternately, distinct *partial* self-monitoring 'readouts' that correspond to the different personalities MPD/DID sufferers experience themselves as having or being.

But these 'personalities' are fundamentally incomplete, because none has all the capacities required to account for all the patient's doings. They are, as the narrativists would have it, distinct 'self narratives'—which may have deleterious implications for the ability of an MPD/DID patient to act as a unified agent in her social context.[8] They are not, however, distinct selves *qua* distinct authors: the same (faulty) self-monitoring system is the source of them all.

8.3.4 Schizophrenia

Schizophrenia is not, as its name implies and as is often popularly assumed, a 'split-mind' condition. Rather, it is characterized by hallucinations, delusions, 'thought disorder' ('patterns of reasoning which appear to us to be odd, circuitous, or nonsensical'), along with flattened affect, alogia ('reduced fluency of thought') and avolition ('difficulty initiating things by one's own will power') (Nettle, 2006, pp. 90–91). It is with the symptoms of delusion and hallucination that schizophrenia comes within the purview of a study of the self: schizophrenics typically suffer from *auditory verbal hallucinations* (AVH), i.e. the experience of 'hearing voices', and *thought insertion*, where a patient experiences a thought as not being his or her own, sometimes as being that of a quite specific other person.

Assuming that the thoughts schizophrenics experience as voices or inserted thoughts are, in fact, their own, but are not experienced as such, one obvious characterization of what has gone wrong here is that it is a failure to distinguish between oneself and the external world, a dissolution of the self–world boundary. This view, which is attributed to Freud (Hoerl, 2001), seems initially plausible. But a functioning self–world boundary is one of the most fundamental of our cognitive capacities (Dennett, 1991, ch. 7; Churchland, 2002a, ch. 3; Metzinger, 2004), without which an organism would not be capable of survival. One might of course suggest that the dissolution of that boundary is *partial* in schizophrenia—leaving intact the ability to distinguish one's body from the environment, say, but impairing the patient's ability to distinguish his own conscious cognitive operations from the outside world.

---

[8] For a treatment of the ethical implications of dissociation, see Radden (1996).

But that suggestion still does not quite capture the phenomenology of AVH and thought insertion.[9] In a study of AVH by McCarthy-Jones and colleagues (2014), more subjects reported the 'location' of their voices as 'inside the head' than 'outside the head'. As to thought insertion, that phenomenon does not involve the simulacrum of an auditory perception, but the patient's impression that 'thoughts which are *not his own* are intruding *into his* mind' (Wing et al., 1974, p. 160, emphasis added). For both AVH and thought insertion, the essential feature of the experience seems not to be that the schizophrenia patient confuses the external world with the 'inner world' of his own mind, as the Freudian theory of self–world boundary dissolution would have it. Rather, both kinds of phenomena are or can be experienced as occurring *within* oneself, without however being *one's own*, having been put there by some outside agency.

The breakdown, then, is not in the schizophrenic's self–world boundary, but rather in his 'subjective unity' (Bayne, 2013) or, more precisely, his sense of *agency* (Stephens & Graham, 2000). On this view, as Christoph Hoerl (2001) characterizes it, schizophrenics suffering from AVH or inserted thoughts 'lack a sense of active participation in the occurrence of certain thoughts' (p. 189). Put another way, the defect concerns not 'where the thought occurs' (p. 191) but rather 'who is . . . *the author* of that thought' (p. 192, emphasis added). It is, then, a failure of the self-monitoring system at the stage where the system would normally keep track of what are one's own cognitive processes. An AVH or inserted thought is correctly monitored as occurring *to* a schizophrenia sufferer, but fails to be registered as *his own* thought (cf. Frith, 1992).

The effect of these phenomena on schizophrenia patients' sense of self is stark. Wells (2003) reports on a young woman with schizophrenia, Joanne, who described her situation thus:

> "I know I'm still myself, but it doesn't feel that way. Where I was is filled with noise and voices, and there's—it's a small area, the brain, but there's a huge emptiness there that I used to fill." (p. 299)

---

[9] Perhaps surprisingly, qualitative studies of the phenomenology of hallucinations in schizophrenia are quite a recent development in psychiatric research. A commendable multidisciplinary research project in this area, 'Hearing the Voice', is currently underway at Durham University (Woods et al., 2014; <http://hearingthevoice.org>).

Despite her *epistemic* self-awareness still functioning ('I *know* I'm still myself'), this patient finds her *phenomenal* sense of self seriously impaired ('it doesn't *feel* that way'). This is consonant with how Louis Sass and Josef Parnas (2001) characterize the phenomenology of schizophrenia: a 'diminishment in the sense of existing as a subject of experience' (p. 352) and a 'lack of continuity in the sense of self and a sense of alienation from one's own body, face, and thoughts' (p. 353). As the normal processes of self-monitoring in the present are malfunctioning, it becomes impossible for the patient to sustain a coherent sense of self over time.

What exactly is going wrong when a schizophrenia sufferer experiences his own thoughts as inserted or voiced by someone else? A plausible explanation first put forward by Irwin Feinberg (1978) is that the phenomenon results from a faulty or absent *efference copy* (also known as 'corollary discharge') of the 'motor mechanisms of thought' (p. 639). When the brain initiates a motor action, along with the efferent neural signal for making a movement, it also generates a *copy* of the efferent signal which 'loops back to the sensory systems identifying you as the source of the . . . movement' (Churchland, 2013, p. 210). Efference copy is thus a very useful self-monitoring device for keeping a record of one's motor signals (and, incidentally, the reason why one cannot tickle oneself).

Feinberg's suggestion is that efference copy also accompanies thought. For verbal thought, that is plausible, since in 'inner' or 'covert speech' the motor areas responsible for language production are active. Thus, as Pat Churchland describes it:

> When you merely *think*, "I need to buy milk,' . . . a movement-planning signal informs the sensory brain about the source of the covert speech—*me*. Except sometimes the mechanism is bungled . . . So sometimes a person may *think*, "I need to buy milk," but because there is no efference copy signal or none with the right timing, he may fail to realize that his thought is actually *his* thought. (ibid., original emphasis)

There is evidence that efference copy is indeed dysfunctional in schizophrenics, whether or not they experience AVH (Ford & Mathalon, 2005).

Another potentially promising approach to explaining the anomalies of self-monitoring in schizophrenia may be found in recent research on the brain's default-mode network (DMN). The DMN is characterized by activity in the cortical midline regions that is hypothesized to be the brain's 'de-

fault mode' or 'baseline activity'. This activity is 'suspended during specific goal-directed behaviors' (Raichle et al., 2001, p. 676) but 'engaged when individuals are left to think to themselves undisturbed' (Buckner et al., 2008, p. 1). A meta-analysis of neuroimaging studies of self-specific activity suggests that DMN activity is indeed related to self-monitoring (Qin & Northoff, 2011).

Recent studies suggest that the DMN is both hyperactive and hyperconnected in schizophrenia patients and fails to deactivate, as it normally would, when they are engaged in a task (Whitfield-Gabrieli et al., 2009; Pomarol-Clotet et al., 2008). This is consistent with the hyperreflexivity—excessive self-focused attention—that is a feature of schizophrenia (and other mental disorders) (Pérez-Álvarez, 2008). But how would a hyperactive and hyperconnected DMN be linked with the disturbed self-monitoring that is evident in schizophrenics' hallucinations? Buckner and colleagues (2008) note that 'there appears to be dynamic competition between the default network and brain systems supporting focused external attention', an interaction which is normally subject to a control mechanism. They suggest: 'The complex symptoms of schizophrenia could arise from a disruption in this control system resulting in an overactive (or inappropriately active) default network' (p. 27). Qin and Northoff (2011) suggest that 'our sense of self may result from *a specific kind of interaction* between resting-state [i.e. DMN] and stimulus-induced activity' (p. 1221, emphasis added). It may thus be a defect in the normal control of this interaction that accounts for both the hyperactive DMN and the erroneous readings of the self-monitoring system in schizophrenia.

Whether AVH and thought insertion are explained by defective efference copy or by an abnormal DMN, or both (a hyperactive DMN could in some way be linked to the dysfunction of efference copy), in any case the disturbed self of schizophrenics seems to result from the malfunction of a rather basic self-monitoring mechanism in the brain.

8.3.5 Self disorders and narrativity

Let me conclude this section with some remarks on how these self disorders fit or do not fit with the narrative accounts I discussed in Chapters 5 and 6.

For what I have termed *simple* narrative-capacity views, according to which the self may engage in narrative thinking and narrative practices to a greater or lesser extent, the self disorders discussed here pose no difficulty. A person with MPD/DID will have a number of different, incomplete, discontinuous, though somewhat overlapping, self narratives. The 'protagonist' of each narrative will be the alter or personality from whose perspective the narrative is told, the narrative being illustrative of that alter's particular personality traits.

For a bipolar person, self narratives may similarly seem somewhat discontinuous in that they will be coloured by the patient's current mood state —(hypo)manic, neutral, or depressed. For instance, in the depressed mood state, the self narrative is unlikely to contain any positive planning for the future and likely to edit out any personal memories of a positive valence, dwelling instead on memories of disappointments and failure. Meanwhile, in a hypomanic state, the self narrative likely includes exaggerated plans and schemes and a similarly exaggerated emphasis on past accomplishments. It's important to note, however, that the discontinuities of a bipolar self narrative are discontinuities in the selection of life events dictated by the current mood; they are not complete temporal breaks as might be found in MPD/DID self narratives. For, as mentioned above, the psychological continuity of a bipolar patient is maintained despite the shifting mood states, so his self narratives, however varied in mood, retain the same protagonist, who *undergoes* those mood alterations.

In considering self narratives in schizophrenia, it is worth recalling a point from Rubin and Greenberg's (2003) review of studies of memory disorders cited earlier (§ 5.2.2), to wit, that our capacity for 'narrative reasoning' (where present in the first place) is very difficult to disrupt even in severe psychopathologies. Thus, even as severe a disruption of self-monitoring as that encountered in schizophrenia need not affect patients' ability to think narratively. Statements by schizophrenia patients such as that by Wells's (2003) patient Joanne quoted above seem to bear this out. Her almost poetic description of her predicament—'there's a huge emptiness there that I used to fill'—exemplifies the patient's ability to think narratively and historically precisely in describing the condition that has so seriously impaired her self.

Indeed, of the disorders considered here, schizophrenia is the one that most obviously contradicts narrative *constructionist* accounts of the self. For it is primarily in the *phenomenology* of schizophrenic symptoms—AVH, thought insertion, Joanne's 'huge emptiness'—that the disturbance of self is manifest, which does not require any narrative thinking or overt narration to take place. And if, as seems likely, schizophrenia is a neurological disorder, its deleterious effects on self-monitoring are rooted in deep-seated neurological defects (such as the lack of efference copy and/or defect in DMN control referred to above), not in defective narrativity.

MPD/DID might, superficially, seem a condition more supportive of narrative constructionist accounts: multiple self narratives with multiple protagonists, indeed multiple 'internal' narrators,[10] might suggest a multiplicity of narratively constructed selves. But although '[t]he narrative unity falls apart entirely' (Hardcastle & Flanagan, 1999, p. 652) in cases of MPD/DID, that does not mean that the *author* is disunited, even if the narratives' protagonists and internal narrators are. For, as pointed out earlier, some continuous self-monitoring system is required to keep track of the multiple personalities of the MPD/DID patient. Perhaps this business of keeping track is not unlike the task of a novelist keeping track of the different personalities, styles of speech, and biographies of the various characters in a novel: but the novelist is a single author. The disrupted, discontinuous self narratives of an MPD/DID patient are indeed indicative of a defective self, but they are not evidence of a multiplicity of selves.

In bipolar disorder, it is even less obvious how narrativity is supposed to account for the self, since what disruption of self there is in this disorder is clearly attributable to shifts in mood, which must have a neurochemical cause—otherwise the administration of mood-stabilizing medication would not, as in Dr Jones's case, have such rapid beneficial effects. As Melvin Woody (2003) puts it in discussing Dr Jones, the patient with rapid-cycling bipolar disorder mentioned above:

> Surely, it is not for lack of a narrative that his sense of self becomes fragile and the continuity of his identity becomes tenuous, but because the quality of his experiences changes so drastically from day to day. (p. 334)

---

[10] That is, here: a narrator internal to the narrative, as opposed to the *author* of the narrative.

Again it is the phenomenology of the individual that characterizes the condition, not the shape, presence or absence of a self narrative.

This is not to say that narrative practices and thinking do not play a role in psychiatric patients' being able to *make sense* of their conditions. A patient may well have an 'illness narrative' (Phillips, 2003) that both charts her experience of psychopathology and 'contributes to the experience of symptoms' (p. 320). But this is an example of a self engaging in narrative thinking. It is not a case of narrative self-construction as suggested by narrative constructionist accounts, for the illness narrative does not exhaust the self, nor is the self wholly defined by the illness: 'A disease is something that one has, not that one is.' (Ghaemi, 2013, p. 67)

In sum, while narrativity has a role to play in describing and responding to pathologies of the self, it does not *explain* them. Their causes run deeper, to affective, pre-reflective, neurological disturbances, all of which are prior to any reflective narrative thinking. Defects of narrativity, where they occur, are a *symptom* of such underlying disorders.

8.4 Defects and disorders of self and memory as specific system malfunctions

In describing the various defects and disorders we have encountered in this and the previous chapter, it is easy to slip into metaphorical talk about selves: the autistic child who is 'self-absorbed', the person with SDAM having 'no self', the dementia sufferer 'losing his self', the bipolar patient having 'two selves', the MPD/DID case with 'multiple selves', the schizophrenic experiencing a 'dissolution of the self'… But such talk, while it may perhaps occasionally be socially useful, is unhelpful scientifically and philosophically (and, I submit, ultimately unhelpful for the sufferers of these conditions and those around them), because it does nothing to explain or manage the defects and disorders discussed. For the 'selves' that are supposedly impaired, absent, lost, multiplied, or dissolved are all different.

On my system view of the self, we can make sense of these conditions by noting how each involves a failure or malfunction of a *distinct* self-representational capacity, which explains *in what way* an affected person's self

is affected. In autism spectrum conditions, a number of self-representational processes seem to operate differently from those in neurotypicals, affecting memory operations and self-monitoring of one's cognitive and affective states. In SDAM, there is an absence of self-representation through 'mental time travel', both backwards and forwards, possibly linked to a defect in constructive visual memory and simulation. In frontal-lobe amnesia, self-representation is disrupted at the point of *recollecting* and cross-checking autobiographical memories because the self-monitoring process of executive control is faulty. In other amnesias, different neurological aetiologies produce different impairments affecting the encoding and retrieval of autobiographical memory. In dementia, there *is* progressive deterioration of *all* cognitive functions including higher-level self-representational capacities, but its progress and effect varies greatly between different capacities, like explicit autobiographical remembering, self-recognition, trait self-knowledge, and other implicit memories. In split brains there is a straightforward surgically acquired disruption of interhemispheric communication affecting any self-representational process that normally makes use of the corpus callosum. Bipolar disorder seems to result from faulty subpersonal self-monitoring at the neurochemical level, mood-altering neurotransmitters being abnormally released or withheld. MPD/DID, meanwhile, seems to be a state of altered self-representation (including a disrupted sense of agency and overt self-presentation) at the personal level, induced by psychological trauma. And schizophrenia involves the breakdown of the processes normally subserving the self-ascription of one's own cognitive processes.

There are many more conditions resulting from abnormal or defective self-representational processes that I could have discussed at some length: anosognosia (an organism's failure to self-monitor illness or injury, resulting in patients' insisting that they are well when they are not), phantom limbs in amputees and the converse case of alien-limb syndrome (a severe defect in proprioception resulting in the belief that one of one's limbs is not one's own), out-of-body experiences, and, most strangely perhaps, Cotard delusion (whose sufferers believe that they are dead). But I think my point is clear: while each of these defects and disorders affects the self, and some aspect of the *sense* of self, of its sufferers, each is traceable to a different malfunctioning self-representational process or capacity. The diversity of these

conditions supports my view that the self is a *complex* system comprising a whole range of processes and capacities, each of which can malfunction. But a malfunctioning system is still that same system. And so, though bits of it may go awry, there still is a self.

It remains for me to try and explicate quite what we should take that system to be. After a brief summing up, that is what I shall do in the concluding chapter that follows.

Chapter Nine

Review and conclusions

9.1 Review: the system view and its advantages

In the face of the wide variety of conceptions of the self in the philosophical and psychological literature, my recommendation in this thesis has been that we should conceive of the self as a complex self-monitoring system deploying our higher-level self-representational capacities (ch. 1). The first advantage of this system view is that it puts the self on an objective empirical grounding. Rather than seeking the self by introspecting—the method by which Hume (1739) failed to find it—the system view situates the self in nature. Basic self-representational processes in nervous systems are important evolved capacities of complex organisms. And they provide a 'neural platform' (Churchland, 2002b) for our higher-level self-monitoring processes, such as action planning, impulse control, and autobiographical memory.[1]

The self, understood as the functional system responsible for these processes, has the persistence conditions of the living brain in which it is realized. It does not bifurcate or multiply or get replaced during its owner's lifetime. That is the second advantage of the system view, which distinguishes it from the views of Locke (1690/1706) and the neo-Lockeans (ch. 2). These views mistake the self for (some of) its productions, such as autobiographical memories, allowing all manner of theoretical but otherworldly scenarios like fissured, fused, and disembodied selves. Furthermore, the Lockeans conflate the self with the attributes of personhood—which are a separate philosophical problem. Nevertheless, a quasi-Lockean

[1] As Pat Churchland has pointed out to me, the human self thus exploits capacities we share with many other animals (see § 9.3).

view of the self as defined by the retention of personal memories seems to have some intuitive popular appeal (ch. 3).

It has been worthwhile, therefore, to look more closely into the nature of autobiographical memory (ch. 4). Here it soon became apparent that what we call autobiographical memory is, again, a diverse and complex matter, involving both semantic and episodic memory and requiring different 'enabling systems' (Klein et al., 2004) for its operations. In particular, the reconstructive nature of episodic recollection means that it is misguided to assume the self to be constructed from autobiographical memories. Rather, it is the self that constructs autobiographical recollections.

Narrative practices, whether in overt narration or 'narrative thinking' (Goldie, 2012), help rehearse and consolidate autobiographical memory (ch. 5). But this by no means entails that we construct our *selves* through narrative practices. Narrative structure is not necessary for self-representation, and our higher-level self-representations are far from being exclusively narrative (ch. 6). It is misguided, therefore, to take narrative capacity to be *essential* for a self. Further, narrative *constructionist* accounts of the self face the problem of authorship: who or what is telling the tale? As it happens, the integration of our self-representational processes precedes narrative ability. Our self-monitoring system is the 'author' of our self narratives—rather than our narratives being the origin of the self.

The third advantage of the system view, then, is that it puts things in the right order, empirically and logically. The self is not constructed from autobiographical memories or narratives, as claimed by what I have called historical-constructionist (Lockean and narrativist) views. Rather, the self is the system that, among other things, constructs autobiographical memories and narratives: these temporal constructions are just the consciously available products of a complex self-monitoring system that is constantly running in the background of our cognitive activities. By constructing memories and narratives we situate ourselves in time. But self-monitoring involves more than conscious remembering and narrating: monitoring one's affective and cognitive states, planning behaviour, controlling one's impulses, and so on (see § 9.2.1 below).

Finally, the system view does a better job than its rivals in accounting for deficiencies and disorders of the self. In autism spectrum conditions, some self-monitoring capacities are weakened or absent (central coherence;

theory of mind), but this does not mean that autistic individuals do not have other self-monitoring capacities (such as monitoring their own cognitive and affective states or, in some cases, exceptional episodic recollection) (ch. 7). Their selves may be configured differently from neurotypical selves, but they are still self-monitoring systems. Other defects and disorders of the self-monitoring system, as found in memory deficiencies and personality disorders, are usually circumscribed to a particular self-representational capacity and do not (with the possible exception of late-stage dementia) affect other or indeed all self-representational functions (ch. 8). In a complex system like the self, many different processes can go wrong. Such system malfunctions are usually specific rather than global.

In sum, the system view (1) naturalizes the self, (2) avoids bizarre bifurcations and multiplications of selves and does not conflate the self with the qualities of personhood, (3) does not confuse the self with its productions, in the form of autobiographical memories and self narratives, and (4) accounts for deficiencies and disorders of the self as specific malfunctions in the self-monitoring system. But what exactly, it may now be asked, *is* that system? That is the topic I'll now address.

9.2 The self: the brain, the whole brain, and nothing but the brain?

'It is a category mistake to start looking around for the self in the brain,' says Daniel Dennett (1992, p. 109). If by that he means we should not look for the self in one specific area of the brain, that is surely right—if somewhat trivial. Though some brain functions can be localized neuroanatomically in a fairly precise manner (language production, for instance), the multiplicity of its components makes it very unlikely that this could be true of something as complex as the self-monitoring system.

But if Dennett means that looking around in the brain yields no answers at all with regard to the self (which in the context of his narrative constructionist account of the self seems his likely meaning), that is surely wrong. Clearly, the self-representational capacities that I have characterized as making up the self system are capacities of our brains. The processes exploited by the self-monitoring system are *brain* processes. So—in a loose

and somewhat metaphorical sense—we might say that the self *is* the brain. But—less loosely—is it the *whole* brain and *nothing but* the brain? These questions concern whether we can distinguish between core and peripheral processes of self-representation within the brain, and whether it might make sense to speak of an 'extended self' beyond the brain. I will now outline some of the correlates of self-representation in the brain and sketch answers to those questions.

9.2.1 Where and how does the brain make a self?

Although the self is not going to be associated with activity in a single brain area, there are a number of brain regions whose activity correlates fairly reliably with self-representational processing. They include the thalamus, the insulae, and cortical midline regions such as the medial prefrontal cortex, the anterior and posterior cingulate cortices, and the precuneus; further, in self–non-self distinction tasks, the temporoparietal junction and temporal pole (illustrated in fig. 7.1) (Damasio, 1999; Legrand & Ruby, 2009; Christoff et al., 2011). The anterior and posterior cingulate, the temporoparietal junctions, and the insula have also been associated with the default-mode network (Qin & Northoff, 2011; see § 8.3.4 above). However, these brain areas are not *exclusively* involved in self-representational processing but are also active in other tasks, such as inferential reasoning. Thus, as Kalina Christoff and colleagues (2011) note, 'describing these regions (singly or collectively) as self-specific could be unwarranted' (p. 104).

This note of caution echoes a comprehensive review by Seth Gillihan and Martha Farah (2005) of some three dozen studies of the neural correlates of various self-representational capacities. They note that the brain regions involved in different kinds of self-related processing are highly diverse, and that 'there is generally little clustering even with specific aspects of the self' (p. 94). A striking example of this is provided by their review of a dozen studies of self-face recognition, which used a variety of experimental designs and subjects and, between them, associated the sole task of recognizing images of one's own face with activity in a vast array of brain areas (fig. 9.1). If just one type of self-representational activity yields such a diversity of neural correlates—none of which seems to be uniquely *specialized* for that particular self-representational capacity—it seems likely that

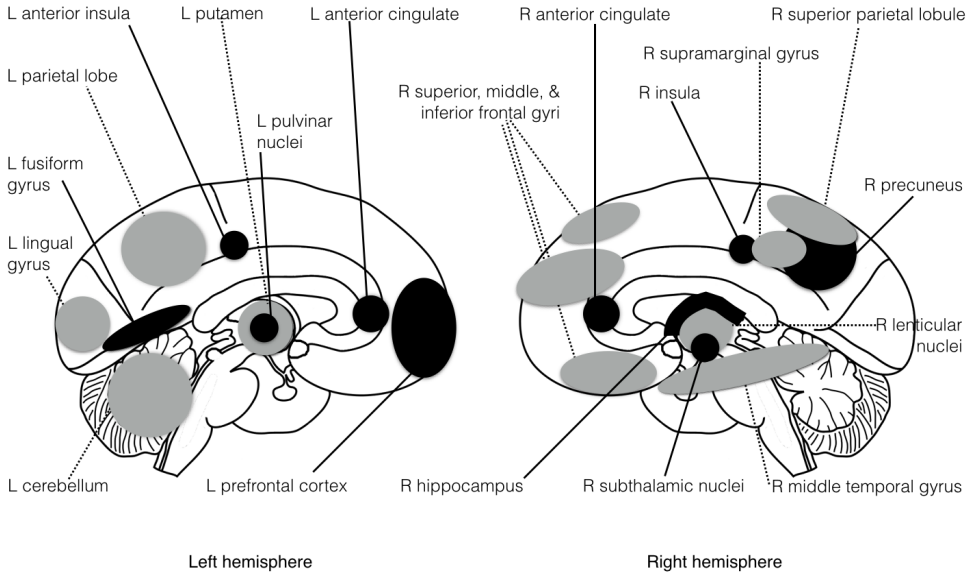

Figure 9.1 Brain areas associated with self-face recognition in studies reviewed by Gillihan & Farah (2005).

much of the brain (subcortical as well as cortical areas, plus the cerebellum) is involved, at different times, in some kind of self-representation or other. The self, then, is not a *neuroanatomically* circumscribed system. Where the neural correlates of self-representational processes are concerned, it seems that the self might really need the whole brain.

Let us then take one step back from neuroanatomy and regard the self, once more, as a complex *functional* system. At the beginning of this thesis, I suggested that we should understand the self as the system of potentially conscious higher-level self-representational capacities (§ 1.4). As we have seen, these are quite diverse. Klein (2010) lists as components of the self our capacities for various semantic and episodic kinds of autobiographical self-knowledge, the sense of agency and ownership, the ability to self-reflect, the sense of personal continuity over time, and self-recognition. To these we must add monitoring and self-ascription of our affective, proprioceptive, and cognitive states, executive control of our behaviour and cognitive functions, evaluative and output functions like Gazzaniga's interpreter mechanism (§ 8.3.1), and probably others. Are any of these capacities and processes more central or essential than others? And is there, perhaps, a

hierarchy of self-representational processes, such that some processes integrate and bundle others?

As for a central or essential self-representational process, Antonio Damasio's (1999; 2010) *core self* looks a credible candidate. This is a transient but already conscious self-representation of the organism in its interactions with its environment at any given time. It is, in other words, a process resulting in what one might call 'bare' or 'raw' self-awareness in the present. (Crucially, this 'core consciousness' is not impaired in conditions like Alzheimer's dementia or schizophrenia.) The core self does not in itself involve any *continuing* self-monitoring. But it is plausible that this capacity *subserves* any higher-level self-monitoring processes, in that self-awareness at any given moment seems a necessary condition for self-awareness over the time it takes, for instance, to plan and execute complex behaviour while monitoring one's affective and cognitive states.

And it is in *executive control*—strongly associated with the frontal lobes of the brain—that we find an integration of different self-representational capacities. Decision-making, the planning and execution of goal-directed behaviour, and controlling our impulses require input from a variety of self-monitoring functions: monitoring and, if necessary, controlling one's affective states (where they might interfere with a goal-directed action), keeping track of one's cognitive states (so as not to lose sight of one's goal), monitoring other body states—hunger, thirst, fatigue, injury—that bear on the success of one's actions, remembering past actions that are relevant to the present goal, and so forth. Executive control, then, seems a strong candidate for a self-monitoring process that integrates subsidiary self-representational capacities. One could not, for instance, be a person in Frankfurt's sense (see § 2.4.1) without executive control. People with damaged prefrontal cortices, like the unfortunate Phineas Gage whose frontal lobes were transpierced by a tamping iron, show erratic and goalless behaviour and are unable to control their impulses in the pursuit of any targeted action (Damasio, 1994). They become akratic, as philosophers like to put it, or 'scatter-brained', as the popular idiom has it. Loss of executive control, as incurred in frontal-lobe lesions, is a serious loss of an important integrative self-monitoring function.

That said, we should not identify the integration of the self-monitoring system with executive control alone. People lacking executive control do

not usually lack self-awareness or their sense of self, which already integrates various somatic, affective, mnemonic and other cognitive self-representations. Nor do we lose our sense of self when we are not currently engaged in goal-directed behaviour. Thus, executive control is not the only process that integrates other self-representational processes. What other such processes does the brain deploy?

A promising proposal for the integration of self-representational functions in self-awareness is Thomas Metzinger's (2003) 'self-model theory of subjectivity'. The core of this theory is the *phenomenal self-model* (PSM), which integrates the consciously accessible self-representational processes instantiated in the brain, from one's body image through affective states to whatever cognitive processing is currently being consciously experienced. The PSM is *phenomenal* in that its contents are available consciously. It is a *dynamic* model, as its contents are continuously changing. (It is also, by default, *transparent*: its contents do not include the information *that* it is a model.) Metzinger suggests that the PSM, integrated with a similar transparent dynamic phenomenal model of an organism's relations with the world, is what gives us not only our sense of self, but our subjective first-person perspective on the world.[2] Metzinger's PSM is, as yet, a 'hypothetical notion' (p. 299), but it is a plausible account of why we seem to have global self-awareness and a sense of self that integrates our multiple distinguishable self-representational processes. More empirical work is needed to test his hypothesis (see § 9.3 below). But if our brains do produce a PSM, that conscious global self-representation accounts for our sense of being a distinct individual in the world, our sense of self. It would be the most fundamental integrative production of the self-monitoring system.

9.2.2 Embodied, extended, and experiential selves

The self, as the brain-based self-representational system of a living organism, is obviously *embodied*. Except in the imagination of certain philosophers, active brains do not dwell in vats, but in living organisms. Many of our self-representational capacities serve to monitor various states of the body,

---

[2] Metzinger goes further, claiming that his theory naturalizes not only subjectivity, but also intentionality. Tantalizing though it is, I cannot go into this second claim here.

from homeostatic regulation and sensorimotor integration to higher-level conscious representations of the organism in its environment. It is for these bodily purposes that self-representational capacities have evolved. There is thus no need of a prolonged argument in support of the view that the self is embodied. As Quassim Cassam (2011) notes, 'the fantasy of a disembodied self is just that: a fantasy' (p. 154). But, more than that, it is by its embodiment in a persisting organism that the self, too, persists over time (Fuchs, 2016).

But then, should we not say that the self is the whole organism, instead of identifying it with the self-representational system in its brain?[3] After all, our self-representations, though generated by the brain, are instances of the *organism* representing itself and not of the *brain* representing itself. Indeed, as I have just said, the evolutionary function of the self-representational system is to monitor the whole organism. However, there are conditions in which the persistence of a living organism and the persistence of a self in terms of potentially conscious higher-order self-representation come apart. An organism in a persistent vegetative state lacks self-awareness and higher-order self-monitoring. It retains only basic self-representational capacities—homeostatic functions—in the brainstem. It has, in Damasio's (1999) terms, a 'proto-self', but it no longer has a self: for it can no longer, as Locke (1690/1706) put it, 'consider itself as itself' (II.xvii.9). Now it is true that the thing that I consider *myself* extends to the whole organism that I am. But my ability to do so is dependent on the functioning of higher-order self-representations generated by my brain. It seems more fruitful, therefore, to think of *the self* as the system that gives me the capacity to perceive myself as a distinct individual than to take it to be the whole organism, which might after all come to lack that capacity—while otherwise remaining the same organism. Nor would it make much sense to identify 'the self' with the referent of each and every ordinary use of 'me' or 'myself', which often is not the *whole* organism, as when someone, having cut a finger, says 'I hurt myself'—and sometimes may not be the organism at all, if I were to say, for instance, 'I put a lot of myself into this thesis'.

Does the self, then, perhaps extend *beyond* the organism? In their 'extended mind' thesis, Andy Clark and David Chalmers (1998) suggest that cognition extends beyond 'the boundaries of skin and skull' (p. 7), in-

[3] I thank Albert Newen for this challenge.

volving external cognitive aids such as, in their example, a notebook used as an *aide-mémoire* by a dementia sufferer, Otto. I cannot evaluate the extended mind thesis *per se* here. But Clark and Chalmers suggest further that such an extended mind implies an *extended self*:

> Most of us already accept that the self outstrips the boundaries of consciousness; my dispositional beliefs, for example, constitute in some deep sense part of who I am. If so, then these boundaries may also fall beyond the skin. The information in Otto's notebook, for example, is a central part of his identity as a cognitive agent. What this comes to is that Otto *himself* is best regarded as an extended system, a coupling of biological organism and external resources. (p. 18)

Now it might at first seem plausible that the self, as a self-representational system, can stretch beyond the organism to any external representations of oneself, like photographs, diaries, and other cognitive aids like Otto's notebook. But, as Eric Olson (2011) points out, such an extended self would be a rather strange kind of entity: thinking beings, on this view, would not be biological organisms, but 'bundles of mental states and processes' (p. 495). For a naturalistic account of the self, that does not seem acceptable. Nor does the extended mind thesis *entail* an extended self, for even if one's cognitive processes transcend the boundaries of the organism, it does not follow that the cognizing *subject* does, too: 'Otto's memories may or may not extend into the notebook he keeps in a drawer. But either way, Otto himself is right here.' (ibid.) So, though we use external cognitive aids that may serve to shore up our sense of self over time, such items aren't *part of* the self. They are *used by* the self. The self-representational *processing* that is triggered or aided by external items still happens in the brain. The self here is the actor, not the prop.

The final challenge to identifying the self with the brain's self-representational capacities that I wish to consider is of a more methodological bent. For it might be said that a naturalistic account of the self such as the one I have proposed here leaves out the *experiential* aspect of the self. Thus, Stanley Klein (2012b) suggests a distinction between what he calls (in a somewhat non-standard way) the 'epistemological self' and the 'ontological self'. The first is 'a collection of diverse neural components that provide us with our beliefs, memories, desires, personality, emotions, etc.' (p. 474). This sounds very much like the self-monitoring system I have sketched

here. It is, in Klein's words, 'the self of science' (p. 508). The other, 'the self of experience' (ibid.), is our 'subjective, unified awareness, a point of view in the first person' (p. 474). The two selves, or aspects of the self, are further distinguished in that the 'self of science' is an object, and the 'self of experience' a subject. The suggestion is that the methods of science have no traction on the latter: in Klein's view, 'it is important not to . . . reduce the conscious self to the self of empirical exploration' (p. 508).

I find this distinction, and the implied irreducibility of the subjective 'self of experience' to the objective 'self of science', deeply misguided. For the whole point of the scientific project of identifying and studying the complex self-representational processes in the brain which 'provide us with our beliefs, memories, desires, personality, emotions, etc.' is to give an account of how we come to experience ourselves as unified subjects of awareness. The self of experience is not a different thing from the self of science. They are, as Thomas Fuchs (2016) puts it, 'two aspects of one and the same life process'. Our having 'a point of view in the first person' is precisely what, for example, Metzinger's neuroscientifically informed self-model theory is designed to explain: one's first-person perspective on the world is a result of the integration of self-representational processes in the brain. Scientific exploration does not take away the experiential qualities—the 'what it is like'—of our subjective awareness. On the contrary: it adds to our understanding of how we experience ourselves in the world in the ways that we do. Thus, erecting conceptual barriers between selves of science and of experience is not helpful.

There is much more work to be done in investigating self-representation in the brain, and I will outline some directions for future research in the next, final section of this thesis. Before doing so, let me briefly summarize the points I have made in this section. First—*pace* Dennett—we *should* look around for the self in the brain, though we should not expect to find it in a particular brain region or a circumscribed neural network. The neural correlates of self-monitoring are widely distributed across the brain. Secondly, it is not very promising to look around for the self *outside* the brain, still less outside the organism: while the objects of neural self-representational activity are wherever the nervous system reaches, and whereas we make use of external cognitive aids in our higher-level self-representations, the processing and integration of these representations are done by

the brain. Finally, there is no unbridgeable metaphysical *or* epistemological gap between the self we *experience* 'from the inside' and the self we investigate *scientifically* by means of, say, an MRI scanner. They are the same self. Let us then see what else we may learn about it empirically.

9.3 An interdisciplinary research programme

Research into selves as self-representational systems and their components already straddles many disciplines. Philosophy has a role here in offering conceptual clarifications (as I hope I have done in this thesis) and, experimentally, in studying popular intuitions about the self. Disciplines conducting empirical self-related research include psychology, psychiatry, cognitive neuroscience, evolutionary biology and different branches of zoology, and the already interdisciplinary research areas of artificial intelligence and robotics. I here outline some self-related topics of research that are currently being investigated or should receive researchers' attention in the near future.

Does the magpie making a racket outside my window have a self? It seems that she may, at least, have the capacity to recognize herself in a mirror (Prior et al., 2008). Evidence of mirror self-recognition has been noted in a variety of species, including chimpanzees (Gallup, 1970), bottlenose dolphins (Reiss & Marino, 2001), and, most recently, giant manta rays (Ari & D'Agostino, 2016). Meanwhile, it has been suggested that my magpie's corvid cousins may engage in a form of mental time travel (Clayton & Dickinson, 1998), as may rats (Roberts & Feeney, 2009). There are live disputes over whether all cases of contingency checking in a mirror amount to genuine self-recognition (Gallup et al., 2011) and in what ways rats' and corvids' mnemonic capacities really resemble human mental time travel (Suddendorf & Corballis, 2007; 2010). But such disputes are evidence that there is a flourishing research programme underway concerning self-representational capacities in non-human animals, from which we may in time draw conclusions about how our own self-representational system has evolved, and why there seems to be convergent evolution of these capacities in quite distantly related species: there clearly appear to be adaptive advantages to an organism's having complex self-representational capacities.

We should not assume that we humans are the only animals around that have such higher-level capacities.

For our own species, there is much research to be done in seeking to identify the neural correlates of higher-level self-representations, such as Metzinger's (2003) phenomenal self-model. As mentioned earlier (§ 9.2.1), these correlates are likely to be distributed across different brain regions, so the focus will have to be less on picking out specific neuroanatomical locations than on the processes that activate particular networks and combinations of brain areas. Open questions here concern the precise role the default-mode network (§ 8.3.4) plays in self-related processing, what its activation and deactivation mechanisms are, and what other functions these might serve. A related topic of interest is Damasio's (1999; 2010) suggestion that the emergence of a core self and that of consciousness are inextricably linked. Since basic self-representational processes (homeostasis, sensorimotor integration) happen non-consciously, it seems clear that self-representation *per se* precedes consciousness, and it is therefore unlikely that there could be consciousness without its involving any form of self-representation. It is less clear whether consciousness is a by-product of the integration or recursion of self-representational processes or a different process entirely. The balance of the neural evidence so far suggests the former (with 'core consciousness' and the core self sharing a neural substrate), but more empirical work on the boundaries of consciousness (in the sleep–wake cycle, or in psychopathologies involving unusual disturbances of consciousness) will be useful here.[4]

Metzinger's self-model theory is an account of our first-person perspective on the world, an attempt to explain subjectivity by showing how the brain models the relations between the self and objects in its environment. On this point, Dorothée Legrand and Perrine Ruby (2009) suggest that our best chance of finding neural mechanisms that are truly self-specific (rather than shared with other neurocognitive functions) will consist precisely in studying the neural processing of self–object relations. They hypothesize that the first-person perspective arises from the sensorimotor integration of efference copy of the brain's motor commands (see § 8.3.4) with sensory re-afference, that is, the perceived feedback on the effects of our

[4] It may also help put to rest the 'new mysterianism' about consciousness propounded by the likes of Thomas Nagel (1974) and David Chalmers (1995).

motor actions in the world. This is an ongoing research project worthy of detailed investigation—particularly of how, where, and how soon efference copy and re-afference signals are integrated. On a related note, one may wonder whether and how self-representations and the representation of self–world interactions fit with the recent model of the brain as a predictive processing engine, defended notably by Andy Clark (2013). Here, Anil Seth (2013) suggests that the self, as an embodied mechanism, does indeed operate as an inferential, predictive system. This too is very much a project for future research, including the role of particular neural populations, neurotransmitters, and higher-level cognitive processes in the error correction mechanisms posited by the predictive-processing thesis.

What we can learn about self-representation and the representation of self–object relations in the brain has applications in robotics. The issue here is not to build *self-aware* robots (though it is an intriguing philosophical question what level of sophistication and integration of robotic self-representational capacities may be said to constitute self-awareness[5]). Rather, current robotics research is concerned with endowing robots with a predictive self-model and a brain-like sensorimotor control architecture that will allow them to adapt to and interact with different environments and to plan and guide their activities (Schilling & Cruse, 2012; Prescott et al., 2014). Robots with these capacities will be useful—and needed—in areas as diverse as therapeutic settings and personal care, infrastructure repairs, and planetary exploration.

Returning to the human self, there are many open questions concerning its subsidiary processes of memory that I have discussed in this thesis. Just how much autobiographical remembering is reconstructive? Is 'severely deficient autobiographical memory' (§ 8.2.1) a freak condition or one end of a spectrum ranging from a complete absence of episodic recollection to the videographic memories of autistic Temple Grandin (§ 7.3.3)? More basically, how does memory retrieval actually work in neural terms? We have a fair amount of knowledge of the neural correlates of *encoding* memory (most prominently, the hippocampus and adjoining cortical regions), but so far very little is known about the neural basis of *retrieving* memories. It seems reasonable to expect that some more of our pretheoret-

[5] A charming exploration of this question and its ethical implications can be found in the *Star Trek: The Next Generation* episode 'The Quality of Life' (Frakes, 1992).

ical assumptions about remembering will be challenged as cognitive neuroscience makes progress on these questions. And, concerning those pretheoretical assumptions, I plan to conduct an experimental philosophy study of whether there is a popular causal conception of remembering, according to which we assume memories to be causally linked, in the 'right' way, to the events of which they are memories (§ 4.4.1).

Finally, there are open questions for the psychiatry of the self, particularly with respect to autism spectrum conditions and schizophrenia. Though the diagnostic criteria of autism in behavioural experiments continue to evolve, we still do not know quite how the neural architecture of people with autism differs from that of neurotypicals, or indeed whether there is a single neurological cause of the symptoms of autism. Another urgent question is what precisely has gone wrong, and from what causes, with self-representational processing in schizophrenia. An account of the neural causes of the breakdown of the self in schizophrenia could point the way towards preventing or curing this distressing condition. Above all, we should hope that a conceptualization of the self as a complex system with many different components may forestall the all-too-easy but, to those concerned, detrimental assumption that people with autism spectrum conditions, with schizophrenia, or with dementia have no self. Their self-monitoring system may be differently configured (in autism) or defective (in schizophrenia and dementia)—but it is not absent.

The self—its evolution, its functioning in the brain, its links with consciousness, its malfunctions in psychiatric disorders, its possible replication in robots—promises to be the subject of a diverse interdisciplinary research programme for some time yet. I like to think that David Hume, were he alive today, would be fascinated by this—and eager to join in.

References

Adler, N., B. Nadler, Z. Eviatar, and S. G. Shamay-Tsoory. 2010. The relationship between theory of mind and autobiographical memory in high-functioning autism and Asperger syndrome. *Psychiatry Research* 178: 214–216.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)*. Arlington, Va.: American Psychiatric Association.

Anderson, I. M., P. M. Haddad, and J. Scott. 2012. Bipolar disorder. *British Medical Journal* 345: e8508.

Anscombe, G. E. M. 1975. The first person. In S. Guttenplan (ed.), *Mind and Language: Wolfson College Lectures, 1974*, Oxford: Oxford University Press.

Ari, C., and D. P. D'Agostino. 2016. Contingency checking and self-directed behaviors in giant manta rays: Do elasmobranchs have self-awareness? *Journal of Ethology* 34: 167–174.

Aquilina, C., and J. C. Hughes. 2006. The return of the living dead: agency lost and found? In Hughes, Louw, and Sabat, 2006, pp. 143–161.

Baddeley, A. D. 1982. Domains of recollection. *Psychological Review* 89 (6): 708–729.

Baddeley, A. D., M. W. Eysenck, and M. C. Anderson. 2009. *Memory*. Hove and New York: Psychology Press.

Baddeley, A. D., and B. Wilson. 1986. Amnesia, autobiographical memory, and confabulation. In D. C. Rubin (ed.), *Autobiographical Memory*, Cambridge and New York: Cambridge University Press, pp. 225–252.

Bakhurst, D. 1995. Wittgenstein and social being. In D. Bakhurst and C. Sypnowich (eds), *The Social Self*, London: Sage, pp. 30–46.

Barclay, C. R. 1994. Composing protoselves through improvisation. In Neisser & Fivush, 1994, pp. 55–77.

Barlassina, L., and A. Newen. 2014. The role of bodily perception in emotion: In defense of an impure somatic theory. *Philosophy and Phenomenological Research* 89 (3): 637–678.

Barnes, J. L., and S. Baron-Cohen. 2012. The big picture: Storytelling ability in adults with autism spectrum conditions. *Journal of Autism and Developmental Disorders* 42: 1557–1565.

Baron-Cohen, S. 1995. *Mindblindness: An essay on autism and theory of mind*. Cambridge, Mass.: MIT Press.

——1999. The extreme male-brain theory of autism. In H. Tager-Flusberg (ed.), *Neurodevelopmental Disorders*, Cambridge, Mass.: MIT Press, pp. 401–429.

Baron-Cohen, S., A. M. Leslie, and U. Frith. 1985. Does the autistic child have a 'theory of mind'? *Cognition* 21: 37–46.

Baron-Cohen, S., M. O'Riordan, V. Stone, R. Jones, and K. Plaisted. 1999. Recognition of faux pas by normally developing children and children with Asperger Syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders* 29: 407–418.

Baron-Cohen, S., S. Wheelwright, and T. Jolliffe. 1997. Is there a language of the eyes? Evidence from normal adults and adults with autism or Asperger's syndrome. *Visual Cognition* 4: 311–331.

Baron-Cohen, S., S. Wheelwright, J. Hill, Y. Raste, and I. Plumb. 2001. The 'reading the mind in the eyes' test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry* 42: 241–251.

Barresi, J., and R. Martin. 2011. History as prologue: Western theories of the self. In Gallagher, 2011b, pp. 33–56.

Bartlett, F. C. 1932/1995. *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.

Bauman, M. L. 1999. Autism: clinical features and neurobiological observations. In H. Tager-Flusberg (ed.), *Neurodevelopmental Disorders*, Cambridge, Mass.: MIT Press, pp. 383–399.

Bayne, T. 2013. The disunity of consciousness in psychiatric disorders. In K. W. M. Fulford, M. Davies, R. G. T. Gipps, G. Graham, J. Z. Sadler, G, Stanghellini, and T. Thornton (eds), *The Oxford Handbook of Philosophy and Psychiatry*, Oxford: Oxford University Press, pp. 673–688.

Beauchamp, T. L. 1999. The failure of theories of personhood. *Kennedy Institute of Ethics Journal* 9 (4) 309–324.

Berkeley, G. 1732/2008. Alciphron: or the minute philosopher. In *Philosophical Writings* (ed. D. Clarke), Cambridge: Cambridge University Press. As cited in Strawson, 2011a.

Bernecker, S. 2009. *Memory: A philosophical study*. New York: Oxford University Press.

Berniūnas, R., and V. Dranseika. 2016. Folk concepts of person and identity: a response to Nichols and Bruno. *Philosophical Psychology* 29 (1): 96–122.

Berntsen, D., and D. C. Rubin. 2002. Emotionally charged autobiographical memories across the life span: the recall of happy, sad, traumatic, and involuntary memories. *Psychology and Aging* 17 (4): 636–652.

Berryhill, M.E., L. Picasso, R. Arnold, D. Drowos, and I.R. Olson. 2010. Similarities and differences between parietal and frontal patients in autobiographical and constructed experience tasks. *Neuropsychologia* 48: 1385–1393.

Bickle, J. 2003. Empirical evidence for a narrative concept of self. In Fireman et al., 2003, pp. 195–208.

Blok, S., G. Newman, and L. J. Rips. 2005. Individuals and their concepts. In W.-K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, and P. Wolff (eds), *Categorization Inside and Outside the Laboratory*, Washington, D. C.: American Psychological Association, pp. 127–49.

Boeker, R. (forthcoming). Locke on personal identity: A response to the problems of his predecessors. *Journal of the History of Philosophy*.

Borges, J. L. 1942/1944/1974/1997. Funes el memorioso. In *Ficciones*, Madrid: Alianza.

Botterill, G., and P. Carruthers. 1999. *The Philosophy of Psychology*. Cambridge: Cambridge University Press.

Bowler, D. M., J. M. Gardiner, and S. J. Grice. 2000. Episodic memory and remembering in adults with Asperger syndrome. *Journal of Autism and Developmental Disorders* 30 (4): 295–304.

Braude, S. E. 1995. *First Person Plural: Multiple personality and the philosophy of mind* (revised edn). Lanham, Md.: Rowman & Littlefield.

Brockmeier, J., and D. Carbaugh (eds). 2001. *Narrative and Identity: Studies in autobiography, self and culture.* Amsterdam/Philadelphia: John Benjamins.

Brown, S. C., and F. I. M. Craik 2000. Encoding and retrieval of information. In Tulving & Craik, 2000, pp. 93–107.

Bruck, M., K. London, R. Landa, and J. Goodman. 2007. Autobiographical memory and suggestibility in children with autism spectrum disorder. *Development and Psychopathology* 19: 73–95.

Bruner, J. 1987. Life as narrative. *Social Research* 54 (1): 11–32.

——1990. *Acts of Meaning*. Cambridge, Mass.: Harvard University Press. As cited in Nelson, 2003.

——1991. The narrative construction of reality. *Critical Inquiry* 18 (1): 1–21.

——1994. The 'remembered' self. In Neisser & Fivush, 1994, pp. 41–54.

Bruner, J., and D. A. Kalmar. 1998. Narrative and metanarrative in the construction of self. In M. Ferrari and R. J. Sternberg (eds), *Self-awareness: Its nature and development*, New York: Guildford Press, pp. 308–331.

Buckner, R. L., J. R. Andrews-Hanna, and D. L. Schacter. 2008. The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences* 1124: 1–38.

Butler, J. 1736. Dissertation I: Of personal identity. In *The Analogy of Religion, Natural and Revealed*, London: Knapton, pp. 301–308.

Capps, L., J. Kehres, and M. Sigman. 1998. Conversational abilities among children with autism and children with developmental delays. *Autism* 2: 325–344.

Carruthers, P. 2010. Introspection: divided and partly eliminated. *Philosophy and Phenomenological Research* 80 (1): 76–111.

Cassam, Q. 2011. The embodied self. In Gallagher, 2011b, pp. 139–156.

Chalmers, D. J. 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2 (3): 200–219.

Charman, T., and S. Baron-Cohen. 1994. Another look at imitation in autism. *Development and Psychopathology* 6: 403–413.

Charney, R. 1981. Pronoun errors in autistic children: support for a social explanation. *British Journal of Disorders of Communication* 15: 39–43.

Christoff, K., D. Cosmelli, D. Legrand, and E. Thompson. 2011. Specifying the self for cognitive neuroscience. *Trends in Cognitive Sciences* 15 (3): 104–112.

Churchland, P. S. 2002a. *Brain-Wise: studies in neurophilosophy*. Cambridge, Mass.: MIT Press.

——2002b. Self-representation in nervous systems. *Science* 296: 308–310.

——2013. *Touching a Nerve: The self as brain*. New York: Norton.

Clark, A. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36: 181–204.

Clark, A., and D. Chalmers. 1998. The extended mind. *Analysis* 58 (1): 7–19.

Clayton, N. S., and A. Dickinson. 1998. Episodic-like memory during cache recovery by scrub jays. *Nature* 395: 272–274.

Conway, M. A., and D. C. Rubin. 1993. The structure of autobiographical memory. In A. F. Collins, S. E. Gathercole, M. A. Conway, P. E. Morris (eds), *Theories of Memory*, Hove: Lawrence Erlbaum Associates, pp. 103–137.

Cooper, J.M., F. Vargha-Khadem, D.G. Gadian, and E.A. Maguire 2011. The effect of hippocampal damage in children on recalling the past and imagining new experiences. *Neuropsychologia* 49: 1843–1850.

Crane, L., and L. Goddard. 2008. Episodic and semantic autobiographical memory in adults with autism spectrum disorders. *Journal of Autism and Developmental Disorders* 38: 498–506.

Cullen, S. 2010. Survey-Driven Romanticism. *Review of Philosophy and Psychology*, 1: 275-296.

Currie, G. 2010. *Narratives and Narrators: A philosophy of stories*. Oxford: Oxford University Press.

Damasio, A. 1994/2006. *Descartes' Error: Emotion, reason and the human brain*. London: Vintage.

——1999/2000. *The Feeling of What Happens: Body, emotion and the making of consciousness*. London: Vintage.

——2010. *Self Comes to Mind: Constructing the conscious brain*. London: Heinemann.

DeMyer, M. K., G. D. Alpern, S. Barton, W. E. DeMyer, D. W. Churchill, J. N. Hingtgen, C. Q. Bryson, W. Pontius, and C. Kimberlin. 1972. Imitation in autistic, early schizophrenic, and non-psychotic subnormal children. *Journal of Autism and Childhood Schizophrenia* 2: 264–287.

Dennett, D. C. 1976. Conditions of personhood. In A. O. Rorty (ed.), *The Identities of Persons*, Berkeley and Los Angeles: University of California Press, pp. 175–196.

——1981/1997. True believers: The intentional strategy and why it works. In J. Haugeland (ed.), *Mind Design II*, Cambridge, Mass.: MIT Press.

——1984. *Elbow Room: The varieties of free will worth wanting*. Oxford: Clarendon Press.

——1991/1993. *Consciousness Explained*. London: Penguin.

——1992. The self as a center of narrative gravity. In F. Kessel, P. Cole and D. Johnson (eds), *Self and Consciousness: Multiple perspectives*, Hillsdale, N.J.: Erlbaum, pp. 103–115. Online version: <http://cogprints.org/266/1/selfctr.htm> (accessed 10 June 2014).

Descartes, R. 1641/2000. Meditations on First Philosophy. In *Meditations and Other Metaphysical Writings* (trans. D. M. Clarke). London: Penguin.

Doris, J. 2002. *Lack of Character*. Cambridge: Cambridge University Press.

——2009. Skepticism about persons. *Philosophical Issues* 19, Metaethics: 57–91.

——2015. *Talking to Our Selves: Reflection, ignorance, and agency*. Oxford and New York: Oxford University Press.

Eder, R. A. 1994. Comments on children's self-narratives. In Neisser & Fivush, 1994, pp. 180–190.

Ekman, P. 1992. An argument for basic emotions. *Cognition and Emotion* 6 (3–4): 169–200.

Feinberg, I. 1978. Efference copy and corollary discharge: implications for thinking and its disorders. *Schizophrenia Bulletin* 4 (4): 636–640.

Fernández, J. 2008. Memory and time. *Philosophical Studies* 141 (3): 333–356.

Fireman, G. D., T. E. McVay, Jr, and O. J. Flanagan (eds). 2003. *Narrative and Consciousness: Literature, psychology, and the brain*. New York: Oxford University Press.

Fitzgerald, J. M. 1988. Vivid memories and the reminiscence phenomenon: the role of a self narrative. *Human Development* 31: 261–273. As cited in Robinson, 1992.

Fivush, R. 1994. Constructing narrative, emotion, and self in parent–child conversations about the past. In Neisser & Fivush, 1994, pp. 136–157.

Fodor, J. A. 1983. *The Modularity of Mind: An essay on faculty psychology*. Cambridge, Mass.: MIT Press.

Ford, J. M., and D. H. Mathalon. 2005. Corollary discharge dysfunction in schizophrenia: can it explain auditory hallucinations? *International Journal of Psychophysiology* 58: 179–189.

Frakes, J. (dir.) 1992. The quality of life. *Star Trek: The Next Generation*, season 6, episode 9. Los Angeles: Paramount Television.

Frankfurt, H. G. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 68 (1): 5–20.

Freeman, M. 2003. Rethinking the fictive, reclaiming the real: autobiography, narrative time, and the burden of truth. In Fireman et al., 2003, pp. 115–128.

Frith, C. D. 1992. *The Cognitive Neuropsychology of Schizophrenia*. Hove: Lawrence Erlbaum.

Frith, C. D., and U. Frith. 1999. Interacting minds—a biological basis. *Science* 286: 1692–1695.

Frith, U. 2001a. Autism. In R. A. Wilson and F. Keil (eds), *The MIT Encyclopedia of the Cognitive Sciences*, Cambridge, Mass.: MIT Press, pp. 58–60.

——2001b. Mind blindness and the brain in autism. *Neuron* 32: 969–979.

Frith, U., and F. de Vignemont. 2005. Egocentrism, allocentrism, and Asperger syndrome. *Consciousness and Cognition* 14: 719–738.

Frith, U., and F. Happé. 1994. Autism: beyond 'theory of mind'. *Cognition* 50: 115–132.

——and——. 1999. Theory of mind and self-consciousness: what is it like to be autistic? *Mind and Language* 14 (1): 1–22.

Fuchs, T. 2016. Self across time: the diachronic unity of bodily existence. *Phenomenology and the Cognitive Sciences* (published online 16 January 2016), doi: 10.1007/s11097-015-9449-4.

Gallagher, H. L., and C. D. Frith. 2003. Functional imaging of 'theory of mind'. *Trends in Cognitive Sciences* 7 (2): 77–83.

Gallagher, S. 2007. Pathologies in narrative structures. In Hutto, 2007a, pp. 203–224.

——2011a. Introduction: A diversity of selves. In Gallagher, 2011b, pp. 1–29.

——(ed.). 2011b. *The Oxford Handbook of the Self*. Oxford: Oxford University Press.

Gallup, G. G., Jr. 1970. Chimpanzees: self-recognition. *Science* 167 (3914): 86–87.

Gallup, G. G., Jr., J. R. Anderson, and S. M. Platek. 2011. Self-recognition. In Gallagher, 2011b, pp. 80–110.

García-Pérez, R. M., A. Lee, and R. P. Hobson. 2007. On intersubjective engagement in autism: a controlled study of nonverbal aspects of conversation. *Journal of Autism and Developmental Disorders* 37: 1310–1322.

Gazzaniga, M. 1995. Consciousness and the cerebral hemispheres. In M. Gazzaniga (ed.), *The Cognitive Neurosciences*, Cambridge, Mass.: MIT Press, pp. 1391–1400.

——2000. Cerebral specialization and inter-hemispheric communication: does the corpus callosum enable the human condition? *Brain* 123: 1293–1326.

Gergen, K. J. 1977. The social construction of self-knowledge. In T. Mischel (ed.), *The Self: psychological and philosophical issues*, Oxford: Basil Blackwell, pp. 139–169.

Ghaemi, S. N. 2013. What is me? What is bipolar? *Philosophy, Psychiatry, & Psychology* 20 (1): 67–68.

Gillihan, S. J., and M. J. Farah. 2005. Is self special? A critical review of evidence from experimental psychology and cognitive neuroscience. *Psychological Bulletin* 131 (1): 76–97.

Goddard, L., P. Howlin, B. Dritschel, and T. Patel. 2007. Autobiographical memory and social problem-solving in Asperger syndrome. *Journal of Autism and Developmental Disorders* 37: 291–300.

Goffman, E. 1959/1990. *The Presentation of Self in Everyday Life*. London: Penguin.

Goldie, P. 2012. *The Mess Inside: Narrative, emotion, and the mind*. Oxford: Oxford University Press.

Gopnik, A., and J. Astington. 1988. Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development* 59: 26–37.

Grandin, T. 2009. How does visual thinking work in the mind of a person with autism? A personal account. *Philosophical Transactions of the Royal Society B* 364: 1437–1442.

Greene, E. J. H. 1965. *Marivaux*. Toronto: University of Toronto Press.

Greenwald, A. 1980. The Totalitarian Ego: fabrication and revision of personal history. *American Psychologist* 35 (7): 603–618.

Grice, H. P. 1941. Personal identity. *Mind* 50: 330–350.

——1961. The causal theory of perception. *Proceedings of the Aristotelian Society* supp. vol. 35: 121–153.

Hacking, I. 1995a. *Rewriting the Soul: Multiple personality and the sciences of memory*. Princeton, N.J.: Princeton University Press.

——1995b. Why multiple personality tells us nothing about the self/mind/person/subject/soul/consciousness. In Bakhurst & Sypnowich, 1995a, pp. 159–179.

Happé, F. G. E. 1991. The autobiographical writings of three Asperger syndrome adults: problems of interpretation and implications for theory. In U. Frith (ed.), *Autism and Asperger Syndrome*, Cambridge: Cambridge University Press, pp. 207–242.

——1995. The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development* 66: 843–855.

Happé, F., S. Ehlers, S. Fletcher, U. Frith, M. Johannsson, C. Gillberg, R. Dolan, R. Frackowiak, and C. Frith. 1996. 'Theory of mind' in the brain. Evidence from a PET scan study of Asperger syndrome. *Neuroreport* 8: 197–201.

Hardcastle, V. G. 2008. *Constructing the Self*. Amsterdam/Philadelphia: John Benjamins.

Hardcastle, V. G., and O. Flanagan. 1999. Multiplex vs. multiple selves: distinguishing dissociative disorders. *The Monist* 82 (4): 645–657.

Harré, R. 1991. The discursive production of selves. *Theory & Psychology* 1 (1): 51–63.

Hassall, R. 2016. Does everybody with an autism diagnosis have the same underlying condition? In K. Runswick-Cole, R. Mallett, and S. Timimi (eds), *Re-thinking Autism: Diagnosis, identity and equality*, London: Jessica Kingsley.

Hayasaki, E. 2016. In a perpetual present: The strange case of the woman who couldn't remember her past—and can't imagine her future. *Wired*, April 2016, <http://www.wired.com/2016/04/susie-mckinnon-autobiographical-memory-sdam/> (accessed 20 April 2016).

Hehman, J. A., T. P. German, and S. B. Klein. 2005. Impaired self-recognition from recent photographs in a case of late-stage Alzheimer's disease. *Social Cognition* 23 (1): 118–123.

Hill, E. L., and U. Frith. 2003. Understanding autism: insights from mind and brain. *Philosophical Transactions of the Royal Society B* 358: 281–289.

Hirst, W. 1994. The remembered self in amnesics. In Neisser & Fivush, 1994, pp. 252–277.

Hobson, J. A., R. Harris, R. García-Pérez, and R. P. Hobson. 2009. Anticipatory concern: a study in autism. *Developmental Science* 12: 249–263.

Hobson, J. A., and R. P. Hobson. 2007. Identification: the missing link between imitation and joint attention? *Development and Psychopathology* 19: 411–431.

Hobson, R. P. 1990. On the origins of self and the case of autism. *Development and Psychopathology* 2: 163–181.

——2009. Wittgenstein and the developmental psychopathology of autism. *New Ideas in Psychology* 27: 243–257.

——2011. Autism and the self. In Gallagher, 2011b, pp. 571–591.

Hobson, R. P., and A. Lee. 1998. Hello and goodbye: a study of social engagement in autism. *Journal of Autism and Developmental Disorders* 28: 117–127.

——and——1999. Imitation and identification in autism. *Journal of Child Psychology and Psychiatry* 40: 649–659.

Hobson, R. P., G. Chidambi, A. Lee, and J. Meyer. 2006. Foundations for self-awareness: an exploration through autism. *Monographs of the Society for Research in Child Development* 71 (2): 1–165.

Hoerl, C. 2001. On thought insertion. *Philosophy, Psychiatry & Psychology* 8 (2–3): 189–200.

Hogrefe, G.-J., H. Wimmer, and J. Perner. 1986. Ignorance versus false belief: a developmental lag in attribution of epistemic states. *Child Development* 57: 567–582.

Hood, B. 2012. *The Self Illusion: Why there is no 'you' inside your head*. London: Constable & Robinson.

Hopkins, R. (forthcoming). Imagining the past: on the nature of episodic memory. In F. Dorsch and F. Macpherson (eds), *Perceptual Imagination and Perceptual Memory*, Oxford: Oxford University Press.

Hudson, J. A. 1990. The emergence of autobiographical memory in mother–child conversation. In R. Fivush and J. A. Hudson (eds), *Knowing and Remembering in Young Children*, New York: Cambridge University Press, pp. 166–196. As cited in Nelson, 1993.

Hughes, J. C., S. J. Louw, and S. R. Sabat (eds). 2006. *Dementia: Mind, meaning, and the person*. Oxford: Oxford University Press.

Hume, D. 1739/1888/1978. *A Treatise of Human Nature* (ed. L. A. Selby-Bigge, 2nd edn. rev. P. H. Nidditch). Oxford: Oxford University Press.

Hurlburt, R. T., F. Happé, and U. Frith. 1994. Sampling the form of inner experience in three adults with Asperger syndrome. *Psychological Medicine* 24: 385–395.

Hurley, N.C., E.A. Maguire, and F. Vargha-Khadem. 2011. Patient HC with developmental amnesia can construct future scenarios. *Neuropsychologia* 49: 3620–3628.

Hutto, D. D. (ed.). 2007a. *Narrative and Understanding Persons* (Royal Institute of Philosophy Supplement 60). Cambridge: Cambridge University Press.

——2007b. The narrative practice hypothesis: origins and applications of folk psychology. In Hutto, 2007a, pp. 43–68.

Jacobsen, J.-H., J. Stelzer, T. H. Fritz, G. Chételat, R. La Joie, and R. Turner. 2015. Why musical memory can be preserved in advanced Alzheimer's disease. *Brain* 138: 2438–2450.

James, W. 1890/1918. *The Principles of Psychology*. London: Macmillan.

Kanner, L. 1943. Autistic disturbances of affective contact. *Nervous Child* 2: 217–250.

Kant, I. 1781/1787/1966. *Kritik der reinen Vernunft*. Stuttgart: Reclam.

Kasari, C., M. D. Sigman, P. Baumgartner, and D. J. Stiper. 1993. Pride and mastery in children with autism. *Journal of Child Psychology and Psychiatry* 34: 352–362.

Klein, S. B. 2010. The self: as a construct in psychology and neuropsychological evidence for its multiplicity. *WIREs Cognitive Science* 1: 172–183, doi: 10.1002/wcs.25.

——2012a. 'What is the self?': approaches to a very elusive question. *Social Cognition* 30 (4): 363–366.

——2012b. The self and its brain. *Social Cognition* 30 (4): 474–518.

——2013a. The complex act of projecting oneself into the future. *WIREs Cognitive Science* 4: 63–79, doi: 10.1002/wcs.1210.

——2013b. Making the case that episodic recollection is attributable to operations occurring at retrieval rather than to content stored in a dedicated subsystem of long-term memory. *Frontiers in Behavioral Neuroscience* 7, doi: 10.3389/fnbeh.2013.00003.

——2013c. The sense of diachronic personal identity. *Phenomenology and Cognitive Science* 12: 791–811.

Klein, S. B., R. L. Chan, and J. Loftus. 1999. Independence of episodic and semantic self-knowledge: the case from autism. *Social Cognition* 17 (4): 413–436.

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

Klein, S. B., T. P. German, L. Cosmides, and R. Gabriel. 2004. A theory of autobiographical memory: necessary components and disorders resulting from their loss. *Social Cognition* 22 (5): 460–490.

Klein, S. B., and S. Nichols. 2012. Memory and the sense of personal identity. *Mind* 121 (483): 677–702.

Knobe, J., and S. Nichols. 2008. An Experimental Philosophy Manifesto. In J. Knobe and S. Nichols (eds), *Experimental Philosophy*, New York: Oxford University Press, pp. 3–16.

Kringelbach, M. L. 2005. The human orbitofrontal cortex: linking reward to hedonic experience. *Nature Reviews Neuroscience* 6: 691–702.

Kripke, S. A. 1972/1981. *Naming and Necessity*. Malden, Mass., and Oxford: Blackwell.

Kusch, M. 1997. The sociophilosophy of folk psychology. *Studies in History and Philosophy of Science* 28 (1): 1–25.

Lamarque, P. 2007. On the distance between literary narratives and real-life narratives. In Hutto, 2007a, pp. 117–132.

Lavelle, J. S. 2012. Theory-theory and the direct perception of mental states. *Review of Philosophy and Psychology* 3 (2): 213–230.

Lee, A., R. P. Hobson, and S. Chiat. 1994. I, you, me, and autism: an experimental study. *Journal of Autism and Developmental Disorders* 24: 155–176.

Legrand, D., and P. Ruby. 2009. What is self-specific? Theoretical investigation and critical review of neuroimaging results. *Psychological Review* 116 (1): 252–282.

Leslie, A. M. 1987. Pretence and representation: the origins of 'theory of mind'. *Psychological Review* 94: 412–426.

Lewis, D. 1976. Survival and identity. In A. O. Rorty (ed.), *The Identities of Persons*, Berkeley and Los Angeles: University of California Press, pp. 17–40.

Locke, J. 1690/1706/1997. *An Essay Concerning Human Understanding* (ed. R. Woolhouse). London: Penguin.

Loftus, E. F. 1974/1975. Reconstructing memory: The incredible eyewitness. *Psychology Today* 8 (7): 116–119. Reprinted in *Jurimetrics Journal* 15 (3): 188–193.

——1996. Memory distortion and false memory creation. *Bulletin of the American Academy of Psychiatry and the Law* 24 (3): 281–295.

——1997. Creating false memories. *Scientific American* 277 (3) (Sep. 1997): 70–75.

Lorusso, M. L., R. Galli, L. Libera, C. Gagliardi, R. Borgatti, and B. Hollebrandse. 2007. Indicators of theory of mind in narrative production: a comparison between individuals with genetic syndromes and typically developing children. *Clinical Linguistics and Phonetics* 21 (1): 37–53.

Loveland, K. A., R. E. McEvoy, B. Tunali, and M. L. Kelley. 1990. Narrative story telling in autism and Down's syndrome. *British Journal of Developmental Psychology* 8: 9–23.

McCarthy-Jones, S., T. Trauer, A. Mackinnon, E. Sims, N. Thomas, and D. L. Copolov. 2014. A new phenomenological survey of auditory hallucinations: evidence for subtypes and implications for theory and practice. *Schizophrenia Bulletin* 40 (1): 225–235.

Mackie, J. L. 1976. Personal identity. In *Problems from Locke*, Oxford: Clarendon Press, pp. 173–205.

Madell, G. 2015. *The Essence of the Self: In defense of the simple view of personal identity*. New York and Abingdon: Routledge.

Maguire, E.A., F. Vargha-Khadem, and D. Hassabis. 2010. Imagining fictitious and future experiences: evidence from developmental amnesia. *Neuropsychologia* 48: 3187–3192.

Marcovitch, H. (ed.). 2010. *Black's Medical Dictionary* (42nd edn.). London: A & C Black.

Markowitsch, H. J. 2000. Neuroanatomy of memory. In Tulving & Craik, 2000, pp. 465–484.

Matthen, M. 2010. Is memory preservation? *Philosophical Studies* 148: 3–14.

Mead, G. H. 1925. The genesis of self and social control. *International Journal of Ethics* 35: 251–273.

——1934. *Mind, Self, and Society* (ed. C. W. Morris). Chicago: University of Chicago Press.

Metzinger, T. 2003. *Being No One: The self-model theory of subjectivity*. Cambridge, Mass.: MIT Press.

——2004. The *subjectivity* of subjective experience: a representationalist analysis of the first-person perspective. *Networks* 3–4: 33–64.

——2007. Self models. *Scholarpedia* 2(10): 4174, <http://www.scholarpedia.org/article/Self_models> (accessed 9 May 2016).

——2009/2010. *The Ego Tunnel: The science of the mind and the myth of the self*. New York: Basic Books.

Miller, P. J. 1994. Narrative practices: their role in socialization and self-construction. In Neisser & Fivush, 1994, pp. 158–179.

Morcom, A. M. 2014. Re-engaging with the past: recapitulation of encoding operations during episodic retrieval. *Frontiers in Human Neuroscience* 8: 351.

Murphy, K., and P. Naish. 2004/2006. Learning and memory. In K. Murphy, P. Naish, and D. Nettle, *Learning and Language* (SD226 Biological psychology: exploring the brain, Book 5), Milton Keynes: The Open University, pp. 1–48.

Nabokov, V. 1957/1960. *Pnin*. London: Penguin.

Nagel, T. 1971. Brain bisection and the unity of consciousness. *Synthese* 22: 396–413.

——1974. What is it like to be a bat? *Philosophical Review* 83 (4): 435–450.

National Autistic Society. 2015. How to talk about autism. <http://www.autism.org.uk/news-and-events/media-centre/how-to-talk.aspx> (accessed 29 September 2015).

Neimeyer, G. J., and A. E. Metzler. 1994. Personal identity and autobiographical recall. In Neisser & Fivush, 1994, pp. 105–135.

Neisser, U., and R. Fivush (eds). 1994. *The Remembering Self: Construction and accuracy in the self-narrative.* Cambridge: Cambridge University Press.

Nelson, K. 1993. Explaining the emergence of autobiographical memory in early childhood. In A. F. Collins, S. E. Gathercole, M. A. Conway, P. E. Morris (eds), *Theories of Memory*, Hove: Lawrence Erlbaum Associates, pp. 355–385.

——2003. Narrative and the emergence of a consciousness of self. In Fireman et al., 2003, pp. 17–36.

Nelson, K., and R. Fivush. 2004. The emergence of autobiographical memory: a social cultural developmental theory. *Psychological Review* 111 (2): 486–511.

Nettle, D. 2004/2006. Schizophrenia. In F. Toates, B. Mackintosh, and D. Nettle, *Emotions and Mind* (SD226 Biological psychology: exploring the brain, Book 6), Milton Keynes: The Open University, pp. 89–115.

Nichols, S. 2004. Folk concepts and intuitions: from philosophy to cognitive science. *Trends in Cognitive Sciences* 8 (11): 514–518.

——2008. Imagination and the *I*. *Mind & Language* 23 (5): 518–535.

Nichols, S., and M. Bruno. 2010. Intuitions about personal identity: an empirical study. *Philosophical Psychology* 23: 293–312.

Nichols, S., and S. P. Stich. 2003. *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds*. Oxford and New York: Oxford University Press.

Nisbett, R. E., and T. D. Wilson. 1977. Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review* 84 (3): 231–259.

Olson, E. T. 1997/1999. *The Human Animal: Personal identity without psychology*. New York: Oxford University Press.

——1998. There is no problem of the self. *Journal of Consciousness Studies* 5 (5–6): 645–657.

——2011. The extended self. *Minds & Machines* 21: 481–495.

——2015. Personal identity. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2015 Edition), <http://plato.stanford.edu/archives/fall2015/entries/identity-personal/> (accessed 5 February 2016).

Oppenheimer, C. 2006. I am, thou art: personal identity in dementia. In Hughes, Louw, and Sabat, 2006, pp. 193–203.

Oxford University Press. 2015. self, pron., adj., and n. *OED Online*, December 2015, <http://www.oed.com/view/Entry/175090> (accessed 8 February 2016).

Palombo, D. J., C. Alain, H. Söderlund, W. Khuu, and B. Levine. 2015. Severely deficient autobiographical memory (SDAM) in healthy adults: A new mnemonic syndrome. *Neuropsychologia* 72: 105–118.

Parfit, D. 1984/1987. *Reasons and Persons*. Oxford: Oxford University Press.

——2011. The unimportance of identity. In Gallagher, 2011b, pp. 419–441.

——2012. We are not human beings. *Philosophy* 87 (1): 5–28.

Pérez-Álvarez, M. 2008. Hyperreflexivity as a condition of mental disorder: A clinical and historical perspective. *Psicothema* 20 (2): 181–187.

Perry, J. 1972. Can the self divide? *Journal of Philosophy* 69 (16): 463–488.

Phillips, J. 2003. Psychopathology and the narrative self. *Philosophy, Psychiatry, & Psychology* 10 (4): 313–328.

Pillemer, D. B., and S. H. White. 1989. Childhood events recalled by children and adults. In H. W. Reese (ed.), *Advances in Child Development and Behavior*, New York: Academic Press, vol. 21, pp. 297–340. As cited in Nelson, 1993.

Pomarol-Clotet, E., R. Salvador, S. Sarró, J. Gomar, F. Vila, Á. Martínez, A. Guerrero, J. Ortiz-Gil, B. Sans-Sansa, A. Capdevila, J. M. Cebamanos, and P. J. McKenna. 2008. Failure to deactivate in the prefrontal cortex in schizophrenia: dysfunction of the default mode network? *Psychological Medicine* 38: 1185–1193.

Prescott, T. J., N. Lepora, and P. F. M. J. Verschure. 2014. A future of living machines? International trends and prospects in biomimetic and biohybrid systems. *SPIE Proceedings Vol. 9055: Bioinspiration, Biomimetics, and Bioreplication*; doi: 10.1117/12.2046305.

Prinz, J. J. 2004. *Gut Reactions: A perceptual theory of emotion*. Oxford and New York: Oxford University Press.

——2008. Empirical philosophy and experimental philosophy. In J. Knobe and S. Nichols (eds), *Experimental Philosophy*, New York: Oxford University Press, pp. 189–208.

Prior, H., A. Schwarz, and O. Güntürkün. 2008. Mirror-induced behaviour in the magpie (*Pica pica*): evidence of self-recognition. *Public Library of Science Biology* 6/8: e202.

Puccetti, R. 1973. Brain bisection and personal identity. *British Journal for the Philosophy of Science* 24: 339–355.

Qin, P., and G. Northoff. 2011. How is our self related to midline regions and the default-mode network? *NeuroImage* 57: 1221–1233.

Radden, J. 1996. *Divided Minds and Successive Selves: Ethical issues in disorders of identity and personality*. Cambridge, Mass.: MIT Press.

Raichle, M. E., A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman. 2001. A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America* 98 (2): 676–682.

Reid, T. 1785/2002. *Essays on the Intellectual Powers of Man* (ed. D. Brookes). Edinburgh: Edinburgh University Press. As cited in Strawson, 2011a.

Reiss, D. and L. Marino. 2001. Mirror self-recognition in the bottlenose dolphin: a case of cognitive convergence. *Proceedings of the National Academy of Sciences of the United States of America* 98 (10): 5937–5942.

Ricœur, P. 1983–1985. *Temps et récit* (3 vols). Paris: Seuil.

——1990. *Soi-même comme un autre*. Paris: Seuil.

Roberts, W. A., and M. C. Feeney. 2009. The comparative study of mental time travel. *Trends in Cognitive Sciences* 13 (6): 271–277.

Robinson, J. A. 1992. Autobiographical memory. In M. Gruneberg and P. Morris, *Aspects of Memory, vol. 1: The practical aspects* (2nd edn), London and New York: Routledge, pp. 223–251.

Rogers, S. J., S. L. Hepburn, T. Stackhouse, and E. Wehner. 2003. Imitation performance in toddlers with autism and those with other developmental disorders. *Journal of Child Psychology and Psychiatry* 44: 763–781.

Romero, I. 2004/2006. The human nervous system. In F. Toates, I. Romero, and S. Datta, *From Cells to Consciousness* (SD226 Biological psychology: exploring the brain, Book 1), Milton Keynes: The Open University, pp. 85–137.

Ross, M. 1989. Relation of implicit theories to the construction of personal histories. *Psychological Review* 96 (2): 341–357.

Ross, M., and R. Buehler. 1994. Creative remembering. In Neisser & Fivush, 1994, pp. 205–235.

Rubin, D. C., and D. L. Greenberg. 2003. The role of narrative in recollection: a view from cognitive psychology and neuropsychology. In Fireman et al., 2003, pp. 53–85.

Rubin, D. C., S. E. Wetzler, and R. D. Nebes. 1986. Autobiographical memory across the lifespan. In D. C. Rubin (ed.), *Autobiographical Memory*, Cambridge: Cambridge University Press, pp. 202–221.

Sabat, S. R. 2005. Capacity for decision-making in Alzheimer's disease: selfhood, positioning, and semiotic people. *Australian and New Zealand Journal of Psychiatry* 39: 1030–1035.

——2006. Mind, meaning, and personhood in dementia: the effects of positioning. In Hughes, Louw, and Sabat, 2006, pp. 287–302.

Sabat, S. R., and R. Harré. 1992. The construction and deconstruction of self in Alzheimer's disease. *Ageing and Society* 12 (4): 443–461.

Sacks, O. 1985/1986. *The Man Who Mistook His Wife for a Hat*. London: Picador.

——1995. *An Anthropologist on Mars: Seven paradoxical tales.* London: Picador.

Salvadore, G., J. A. Quiroz, R. Machado-Vieira, I. D. Henter, H. K. Manji, and C. A. Zarate Jr. 2010. The neurobiology of the switch process in bipolar disorder: a review. *Journal of Clinical Psychiatry* 71 (11): 1488–1501.

Sass, L. A., and J. Parnas. 2001. Phenomenology of self-disturbances in schizophrenia: some research findings and directions. *Philosophy, Psychiatry & Psychology* 8 (4): 347–356.

Scaife, R. K. 2011. *Agency and Freedom of the Will: The challenge from psychology*. PhD thesis, University of Sheffield.

Schacter, D. L., and D. R. Addis. 2007. Constructive memory: The ghosts of past and future. *Nature* 445: 27.

Schacter, D. L., D. R. Addis, D. Hassabis, V. C. Martin, R. N. Spreng, K. K. Szpunar. 2012. The future of memory: remembering, imagining, and the brain. *Neuron* 76: 677–694.

Schechtman, M. 1990. Personhood and personal identity. *Journal of Philosophy* 87 (2): 71–92.

——1996. *The Constitution of Selves*. Ithaca, N. Y.: Cornell University Press.

——2007. Stories, lives, and basic survival: a refinement and defense of the narrative view. In Hutto, 2007a, pp. 155–178.

——2011a. Memory and identity. *Philosophical Studies* 153: 65–79.

——2011b. The narrative self. In Gallagher, 2011b, pp. 394–416.

——2014. *Staying Alive: Personal identity, practical concerns, and the unity of a life*. Oxford and New York: Oxford University Press.

Scheerer, M., E. Rothmann, and K. Goldstein. 1945. A case of 'idiot savant': an experimental study of personality organization. In J. F. Dashiell (ed.), *Psychological Monographs*, vol. 58 (4), Evanston, Ill.: American Psychological Association.

Schilling, M., and H. Cruse. 2012. What's next: recruitment of a grounded predictive body model for planning a robot's actions. *Frontiers in Psychology* 3: 383.

Schooler, J. W., and E. Eich. 2000. Memory for emotional events. In Tulving & Craik, 2000, pp. 379–392.

Seth, A. K. 2013. Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences* 17 (11): 565–573.

Shoemaker, S. 1970. Persons and their pasts. *American Philosophical Quarterly* 7 (4): 269–285.

——1984. Personal Identity: A materialist's account. In S. Shoemaker and R. Swinburne, *Personal Identity*, Oxford: Basil Blackwell.

——2011. On what we are. In Gallagher, 2011b, pp. 352–371.

Small, J. A., K. Geldart, G. Gutman, and M. A. Clarke Scott. 1998. The discourse of self in dementia. *Ageing and Society* 18: 291–316.

Sorabji, R. 2006. *Self: Ancient and modern insights about individuality, life, and death*. Oxford: Clarendon Press.

Sosa, E. 2008. Experimental philosophy and philosophical intuition. In J. Knobe and S. Nichols (eds), *Experimental Philosophy*, New York: Oxford University Press, pp. 231–240.

Sperber, D., and D. Wilson. 1986. *Relevance: Communication and cognition*. Oxford: Blackwell.

Squire, L. R., and E. R. Kandel. 2009. *Memory: From mind to molecules* (2nd edn). Greenwood Village, Colo.: Roberts & Co.

Stephens, G. L., and G. Graham. 2000. *When Self-Consciousness Breaks: Alien voices and inserted thoughts*. Cambridge, Mass.: MIT Press.

Strawson, G. 2004. Against narrativity. *Ratio* (new series) 17: 428–452.

——2007. Episodic ethics. In Hutto, 2007a, pp. 85–115.

——2009. *Selves: An essay in revisionary metaphysics*. Oxford: Clarendon Press.

——2011. *Locke on Personal Identity: Consciousness and concernment*. Princeton, N. J., and Oxford: Princeton University Press.

——2012. 'We live beyond any tale that we happen to enact'. *Harvard Review of Philosophy* 18: 73–90.

——2015. I am not a story. *Aeon*, 3 September 2015, <https://aeon.co/essays/let-s-ditch-the-dangerous-idea-that-life-is-a-story> (accessed 17 September 2015).

Strawson, P. F. 1962. Freedom and resentment. *Proceedings of the British Academy* 48: 1–25.

Strickland, B., and A. Suben. 2012. Experimenter philosophy: the problem of experimenter bias in experimental philosophy. *Review of Philosophy and Psychology* 3: 457–467.

Suddendorf, T., D. R. Addis, and M. C. Corballis. 2009. Mental time travel and the shaping of the human mind. *Philosophical Transactions of the Royal Society B* 364: 1317–1324.

Suddendorf, T., and M. C. Corballis. 2007. The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences* 30: 299–313.

——and——2010. Behavioural evidence for mental time travel in nonhuman animals. *Behavioural Brain Research* 215: 292–298.

Sutherland, S. 1995. *The Macmillan Dictionary of Psychology* (2nd edn). Basingstoke: Macmillan.

Sutton, J. 2010/2016. Memory. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2016 edn), <http://plato.stanford.edu/archives/spr2016/entries/memory/> (accessed 22 March 2016).

Tager-Flusberg, H. 2000. Language and understanding minds: connections in autism. In S. Baron-Cohen, H. Tager-Flusberg, and D. J. Cohen (eds), *Understanding Other Minds: Perspectives from developmental cognitive neuroscience* (2nd edn), Oxford: Oxford University Press.

Tanweer, T., C. J. Rathbone, and C. Souchay. 2010. Autobiographical memory, autonoetic consciousness, and identity in Asperger syndrome. *Neuropsychologia* 48: 900–908.

Tappen, R. M., C. Williams, S. Fishman, and T. Touhy. 1999. Persistence of self in advanced Alzheimer's disease. *Clinical Scholarship* 31 (2): 121–125.

Tennant, N., and C. Lowe. 2016. Sad robot world. London: Cage Music/Kobalt Music Publishing. Pet Shop Boys, *Super*, x2 0008 CD1, LC 30596, track 9.

Tulving, E. 1972. Episodic and semantic memory. In E. Tulving and W. Donaldson (eds), *Organization of Memory*, New York and London: Academic Press, pp. 381–403.

——1983. *Elements of Episodic Memory*. Oxford: Clarendon Press, and New York: Oxford University Press.

——1985. Memory and consciousness. *Canadian Psychology* 26: 1–12.

Tulving, E., and F. I. M. Craik (eds). 2000. *The Oxford Handbook of Memory*, Oxford: Oxford University Press.

Valberg, J. J. 2007. *Dream, Death, and the Self.* Princeton, N.J., and Oxford: Princeton University Press.

Vasa, R. A., S. H. Mostofsky, and J. B. Ewen (in press). The disrupted connectivity hypothesis of autism spectrum disorders: Time for the next phase in research. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* (available online 2 March 2016), doi: 10.1016/j.bpsc.2016.02.003.

Velleman, J. D. 2005. The self as narrator. In J. Christman and J. Anderson (eds), *Autonomy and the Challenges to Liberalism: New essays*. New York: Cambridge University Press.

Vico, G. 1744/1948. *The New Science* (trans. T. G. Bergin, M. H. Fisch). Ithaca, N.Y.: Cornell University Press.

Vogeley, K., P. Bussfeld, A. Newen, S. Herrmann, F. Happé, P. Falkai, W. Maier, N. J. Shah, G. R. Fink, and K. Zilles. 2001. Mind-reading: neural mechanisms of theory of mind and self-perspective. *NeuroImage* 14: 170–181.

Vogeley, K., and S. Gallagher. 2011. Self in the brain. In Gallagher, 2011b, pp. 111–136.

Wagenaar, W. A. 1994. Is memory self-serving? In Neisser & Fivush, 1994, pp. 191–204.

Wells, L. A. 2003. Discontinuity in personal narrative: some perspectives of patients. *Philosophy, Psychiatry, & Psychology* 10 (4): 297–303.

Wheeler, M. A. 2000. Episodic memory and autonoetic awareness. In Tulving & Craik, 2000, pp. 597–608.

Wheeler, M. A., D. T. Stuss, and E. Tulving. 1997. Toward a theory of episodic memory: the frontal lobes and autonoetic consciousness. *Psychological Bulletin* 121 (3): 331–354.

Whitfield-Gabrieli, S., H. W. Thermenos, S. Milanovic, M. T. Tsuang, S. V. Faraone, R. W. McCarley, M. E. Shenton, A. I. Green, A. Nieto-Castanon, P. LaViolette, J. Wojcik, J. D. E. Gabrieli, and L. J. Seidman. 2009. Hyperactivity and hyperconnectivity of the default network in schizophrenia and in first-degree relatives of persons with schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America* 106 (4): 1279–1284.

Wilkes, K. V. 1988. *Real People: Personal identity without thought experiments*. Oxford: Clarendon Press.

Williams, B. 1970. The self and the future. *Philosophical Review* 79: 161–180.

Wimpory, D. C., R. P. Hobson, J. M. Williams, and S. Nash. 2000. Are infants with autism socially engaged? A study of recent retrospective parental reports. *Journal of Autism and Developmental Disorders* 30: 525–536.

Wing, J. K., J. E. Cooper, and N. Sartorius. 1974. *Measurement and Classification of Psychiatric Symptoms*. Cambridge: Cambridge University Press.

Wing, L. 1996. *The Autistic Spectrum: A guide for parents and professionals*. London: Constable.

Wittgenstein, L. 1922/2003. *Logisch-philosophische Abhandlung (Tractatus logico-philosophicus)*. Frankfurt am Main: Suhrkamp.

——1953/2009. *Philosophical Investigations* (revised 4th edn, trans. G. E. M. Anscombe, P. M. S. Hacker, and J. Schulte). Oxford: Blackwell.

——1974. *Philosophical Grammar* (ed. R. Rhees, trans. A. Kenny). Oxford: Blackwell. As cited in Sutton, 2010.

——1980. *Remarks on the Philosophy of Psychology*, vol. 2 (G. H. von Wright and H. Nyman, eds; trans. C. G. Luckhardt and M. A. E. Aue). Oxford: Basil Blackwell.

Woods, A., N. Jones, M. Bernini, F. Callard, B. Alderson-Day, J. C. Badcock, V. Bell, C. C. H. Cook, T. Csordas, C. Humpston, J. Krueger, F. Larøi, S. McCarthy-Jones, P. Moseley, H. Powell, A. Raballo, D. Smailes, and C. Fernyhough. 2014. Interdisciplinary approaches to the phenomenology of auditory verbal hallucinations. *Schizophrenia Bulletin* 40: S246–S254.

Woody, J. M. 2003. When narrative fails. *Philosophy, Psychiatry, & Psychology* 10 (4): 329–345.

World Health Organization. 1992/2016. *International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10)* (Version: 2016). Geneva: World Health Organization. <http://apps.who.int/classifications/icd10/browse/2016/en> (accessed 18 April 2016).

Zahavi, D. 2005. *Subjectivity and Selfhood: Investigating the first-person perspective*. Cambridge, Mass.: MIT Press.

——2007. Self and other: the limits of narrative understanding. In Hutto, 2007a, pp. 179–201.

——2011. Unity of consciousness and the problem of self. In Gallagher, 2011b, pp. 316–335.

Zikopoulos, B., and H. Barbas. 2013. Altered neural connectivity in excitatory and inhibitory cortical circuits in autism. *Frontiers in Human Neuroscience* 7: 609.