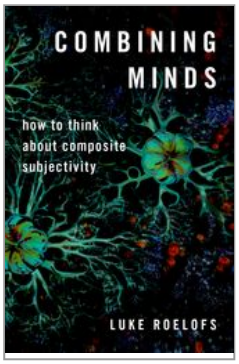

Combining Minds: How to Think about Composite Subjectivity
Luke Roelofs

Print publication date: 2019
Print ISBN-13: 9780190859053
Published to Oxford Scholarship Online: February 2019
DOI: 10.1093/oso/9780190859053.001.0001

Composite Subjectivity and Psychological Subjects

Luke Roelofs

DOI:10.1093/oso/9780190859053.003.0007

Abstract and Keywords

This chapter is about how to combine subjects of experience as they are understood by the psychological theory of personal identity (Neo-Lockeanism). On this theory subjects are not the systems which generate mental states, but are instead constructs defined by the patterns of continuity among mental states. This requires considering how component and composite subjects can be individuated from one another, how they can develop self-consciousness, and how they can display agency. This results in a combinationist account of what is going on in everyday experiences of inner conflict and in dissociative identity disorder—an account which can recognize the conflicting or dissociated parts as subjects in their own right, but also as forming a composite subject with a greater or lesser degree of unity.

Keywords: personal identity, subject of experience, Neo-Lockeanism, self-consciousness, agency, inner conflict, dissociative identity disorder

INNER CONFLICT IS the first thing that comes to mind for many people when they hear the phrase “parts of the mind”: experiences where it feels as though one part of you is struggling against another, as though each is seeking to fulfill its own goals by any means possible, as though each was an agent in its own right. In extreme dissociative cases, they even seem to speak for themselves as separate agents. What are these “parts”? They are probably not the kind of thing covered in divisions 2 and 3—not microsubjects or particular brain regions. So what are they?

In chapter 2 I distinguished a metaphysical and a psychological conception of subjects, and the past four chapters have worked with the former. In this chapter I switch to the psychological conception, outlining “psychological combinationism,” a theory of how human psychological subjects (or “personas”) can be divided and combined. The two conceptions might be seen as rival theories of what subjects are or as complementary perspectives that bring to prominence different aspects of reality. If they are rivals, then psychological combinationism is an alternative to panpsychist combinationism and functionalist combinationism, a conflicting view of what subjects are and how they combine. If the metaphysical and psychological conceptions are complementary, **(p.230)** then psychological combinationists can also be functionalist combinationists or panpsychist combinationists, regarding psychological combinationism as illuminating the world of composite subjectivity from a different angle. They can think that the world contains streams of experiences connected in certain ways and that both the underlying systems which generate these experiences and the structured patterns that coalesce out of these experiences have some claim to being “us.”

Section 7.1 outlines psychological combinationism and explores the complexities of the relationship between component personas and component substrates, using as examples variations on the fictional case of Jekyll and Hyde. This involves considering what psychological combinationism can say about the five internal problems for combinationism considered in chapter 2. In the next two sections I then consider two bridging problems that arise specifically for combining *persons*, subjects with self-consciousness and agency. Section 7.2’s problem concerns self-consciousness: Are parts of a self-conscious composite subject self-conscious in their own right, each knowing itself distinct from the other parts? Section 7.3’s problem concerns agency: generally, what someone else does is not my own action, even if they use my body to do it, so if my parts are themselves agents, I seem to be in competition with them for agency.

My solution to both problems turns on the idea that when the wills of two subjects are related in the right way, they will each experience the other not as distinct but as an extension of themselves, and the actions of both will be attributable to each. When the component subjects within a composite subject are related in these ways, the whole will be conscious of itself as a single entity and will be responsible for certain of the actions taken by its parts. In my final section I consider how the preceding accounts of component personas, and of properly related wills, can cast light on the human experience of inner conflict and dissociation.

7.1. Psychological Combinationism and the Psychological Conception of Subjects

According to what chapter 2 called the “psychological conception of subjects,” we conscious beings are not strictly identical with brain, bodies, organisms, or any other object within which consciousness arises. Rather, we are “personas,” beings constituted by the interplay between different experiences and psychological states. Can such beings be combined with each other into composite personas?

(p.231) 7.1.1. The Neo-Lockean Account of Personal Identity

To answer this question we first need to get clearer on what personas are. My starting point is the popular Neo-Lockean analysis of personal identity, which many philosophers have defended in very similar terms, a sample of which I quote:

> Two soul-phases belong to the same soul . . . if they are connected by a continuous character and memory path. . . . Two soul-phases are directly continuous . . . if the character revealed by the constituents of each is closely similar, and if the later contains recollections of some elements of the earlier. Two soul-phases are . . . connected by a continuous character and memory path if there is a series of soul-phases [between them] all of whose members are directly continuous with their immediate predecessors and successors. (Quinton 1962, 398)

> What I mostly want in wanting survival is that . . . my present experiences, thoughts, beliefs, desires, and traits of character should have appropriate future successors. . . . Change should be gradual rather than sudden, and (at least in some respects) there should not be too much change overall . . . [and] such change as there is should conform, for the most part, to lawful regularities concerning the succession of mental states . . . [so that] each succeeding mental state causally depends for its character on the states immediately before it. (Lewis 1976, 17–18)

> Let us say that, between P today and P [in the past], there are direct memory connections if P can now remember having some of the experiences that P had [then]. . . . There are several other kinds of direct psychological connection [such as] that which holds between an intention and the later act in which this intention is carried out [or] those which hold when a belief, or a desire, or any other psychological feature, persists. . . . We can now define psychological continuity . . . as the holding of overlapping chains of such direct psychological connections; and then [say] that P2 at t2 is the same person as P1 at t1 if and only if P2 at t2 is psychologically continuous with P1 at t1. (Noonan 2003, 10–11, paraphrasing Parfit 1984, 205–207)

Despite their differences in emphasis, these definitions agree on the following: the identity of a person (persona) depends on psychological continuity, i.e., a

continuous chain of “direct psychological connections,” which involve some mix of similarity (your psychology is largely the same from moment to moment), causal dependence (your present state depends on your preceding state), and **(p.232)** representational match (your present memories match your past experiences, and your present intentions match your future actions).

I will accept the Neo-Lockean account of personal identity as a starting point for thinking about the combination of personas. It poses the following four questions, which will be discussed in the next four subsections:

1. First, these are definitions of identity across time; they tell us whether someone has survived some process, not how many people are present at a particular time. Thus to make sense of part-whole relations among personas at a given time requires extending the spirit of these proposals to cover identity at a time rather than across time.
2. Second, if we can define personas in an at-a-time fashion, how does the idea of combining personas fare against the five internal problems for combinationism outlined in chapter 2?
3. Third, in particular, this approach seems to imply maximality (anything entirely contained within a persona fails to count as a distinct persona, just because it is contained within a persona), since any putative component persona will presumably be psychologically connected to the other parts, and hence will be counted as identical with the whole. So it is not clear how a persona could contain distinct component personas.
4. Fourth, even if we can make sense of component personas, as well as component substrates, it is unclear how the two are connected—whether a given system’s being divisible in one way will have anything to do with its being divisible another way.

7.1.2. Identity at a Time and across Time

The Neo-Lockean conception of persons is often motivated by presenting thought experiments of two sorts: those in which a continuous stream of psychological processes is carried on in a sequence of different brains, through each successive brain being somehow rewired so that the new psychology can be “downloaded” (see, e.g., Locke 1836, 229–230; Williams 1970), and those in which two disconnected streams of psychological processes take sequential control of the same brain (see, e.g., Locke 1836, 228–229; Olson 2003). Many find it intuitive that in the former cases a single person has “moved” from one body to another, while in the latter cases two distinct people are “sharing” a single body. A prime example of the second sort of case is that of Dr. Jekyll and Mr. Hyde (Stevenson 1886), where two diametrically opposed personalities alternate in the control of a single **(p.233)** body, one baffled and horrified to discover evidence of the atrocities committed by the other.

These across-time examples can be repurposed to illuminate the individuation of personas at a single time. Surely we can make sense of Jekyll and Hyde not alternating in control of their body, but rather coexisting (cf. Olson 2003, 341–342). When we ask that body for a name, we sometimes get an answer like “Dr. Jek—no, Hyde!—no, let me speak—it’s Hyde, yes, definitely—agh!—oh all right, go on then—Dr. Jekyll, pleased to meet you.” When we give it an opportunity for troublemaking, we observe a few moments of spasms and contortions, as though some sort of struggle is going on, followed by either refusal or embrace of the opportunity, perhaps with one arm trying to interfere with what the rest of the body is doing. At other times we might find the body behaving calmly, striking a middle course between respectability and vice and speaking in the following odd way: “Hello. This is Dr. Jekyll and Mr. Hyde, both pleased to meet you and at your disposal.” If pressed, the body explains that “they” have come to a cooperative arrangement, though temptations sometimes prompt a return to squabbling and contests over control of limbs.

Here it seems intuitive to say that we are dealing with two different people (at least if we thought so when Jekyll and Hyde appeared sequentially rather than together). But there might be no distinction between Jekyll and Hyde at the level of substrates: the entire brain might be active in generating both streams of thought. It might even be that each personality sleeps periodically, leaving the brain under the control of the other, and in these states our scans reveal widespread brain activation that differs only subtly from what is detected when the other personality is awake. We might reasonably conclude that both personalities arise from different patterns of activity in the same set of neurons. So if Jekyll and Hyde are two people, it must be because they are two personas.

So what makes Jekyll and Hyde distinct, in this case? It makes no sense to appeal to the above-quoted definitions of identity across time, because those presuppose a definite number of people present at each time who can be compared with one another. But the relations that those definitions appeal to are clearly relevant to what makes Jekyll and Hyde seem distinct. For instance, instead of appealing to the kind of representational matching that obtains between a memory and a past experience, we might appeal to matching between simultaneous beliefs, desires, and decisions. All of Hyde’s beliefs are consistent, and all of Jekyll’s are, but they might disagree sharply on some points; Hyde desires things to which Jekyll is highly averse, and vice versa; the things Jekyll is attempting to do make **(p.234)** sense in light of Jekyll’s beliefs and desires, but not Hyde’s, and vice versa. Rather than reading these as observations, we could treat them as partly definitional: we assign a belief to Jekyll only if it coheres with the other mental states we assign to Jekyll.

Similarly, whereas the quoted definitions appeal to the causal dependence of one stage on another, we can instead appeal to the causal interdependence of different present mental episodes, in the sense of “causal interdependence”

defined in section 2.2 of chapter 2. Readers may by this point have noticed that the sorts of relations that individuate Jekyll and Hyde seem to be just those relations that were discussed in chapter 2 as versions of "the unity of consciousness." Indeed, it is often suggested that subjects are individuated by unified consciousness (see, e.g., Bayne and Chalmers 2003, 55–57; Dainton 2008; Bayne 2010, 289ff.), though this claim will be as ambiguous as the term "unified consciousness": it can mean global consistency, causal interdependence, functional unity, representational unity, or phenomenal unity. Let us suppose that Jekyll and Hyde are as disunified as any two ordinary people, with respect to all these different relations: despite arising in the same brain, their experiences and other mental states are no more representationally unified, conjointly accessible, etc. than mine are with yours. This seems a good basis for distinguishing them from one another, and this distinction would be between personas, not substrates.

So we can supplement the above-quoted definitions of identity of subjects over time with the following:

> Two personas p_1 and p_2 existing at one time are identical if and only if the experiences and other mental states of p_1 are sufficiently unified in various ways with those of p_2.

This definition is limited in that it presupposes that we can already identify two personas and their experiences; a more thorough statement of the same idea would be something like this:

> Two experiences e_1 and e_2 existing at one time belong to the same persona if and only if they are sufficiently unified in various ways.

This definition makes clearer that, according to psychological combinationism, we are to start with experiences and their relations, and on that basis "construct" personas.

(p.235) 7.1.3. Psychological Combinationism Defended against the Five Arguments

Psychological combinationism says that personas, constituted by sets of sufficiently unified experiences, can be related as part and whole. In chapter 2 we saw five arguments that aimed to rule out combinationism of any form: arguments about subject-summing, unity, privacy, boundaries, and incompatible contexts. How strong are these arguments applied to psychological combinationism?

First, I think that the subject-summing argument has relatively little force. Personas are by definition an ontologically nonfundamental sort of thing, something constituted by more basic facts. More specifically, facts about the existence, identity, and properties of personas are meant to follow from facts about the existence, properties, and interrelations of experiences and other

mental states. And facts about experiences and mental states are surely facts about the intrinsic features of subjects, if anything is, so a full description of one or more subjects would include a catalogue of its experiences and their interrelations, which are the right kind of facts to entail the existence of another subject (cf. Mendelovici 2018, 16).

What about the privacy argument? The experiences which constitute a persona are just those which are ascribed to it, and so any persona constituted by the experiences of another would also be the subject of those experiences, and would thus share experiences with that other. If strong privacy (no nonidentical subjects can share experiences) is true, then it immediately follows that the two subjects are identical, in which case strong independence (which rules out explanatory relations between nonidentical subjects) is vindicated. But there is no good reason to accept strong privacy, since we can accept weak privacy (no non-overlapping subjects can share experiences) instead. In fact, if subjects are personas, then weak privacy is actually true by definition. Personas are constituted by experiences, so for them to overlap simply means for them to share experiences. Hence of course they can only share experiences with other personas they overlap with.[1]

What about the unity argument? Here there is in a sense little to say. Psychological combinationism says nothing distinctive about conscious unity; it simply explains how, given experiences and the unity relations among them, personas can be defined and can combine. Thus for a full account of how conscious unity is explained and grounded, it will need to be supplemented with some other theory.

(p.236) For instance, rather than having a distinctive account of phenomenal unity, it might be supplemented with the primitivist account offered by panpsychist combinationism, or with the reductionist account offered by pure functionalist combinationism. If we took a rivalrous view of the metaphysical and psychological conceptions of subject, then this would involve psychological combinationists taking over some elements of these alternative combinationist theories, while rejecting their accounts of subjecthood; if we took a more conciliatory view of the two conceptions, we might instead see this as psychological combinationism adding an extra layer of depth to those theories, without denying the truth of any of their claims.

Finally, what about the paired boundary argument and incompatible contexts argument? Here is where psychological combinationism faces a distinctive sort of objection: the criterion of identity for personas given in the previous section more or less straightforwardly entails the principle of boundedness (premise **D3** from chapter 2). I mentioned this above as the threat of “maximality” and

address it in the next subsection. But first let us walk through how this entailment runs. Above I suggested the following criterion:

> Two experiences e_1 and e_2 existing at one time belong to the same persona if and only if they are sufficiently unified in various ways.

This features the phrase "belong to the same persona," which becomes ambiguous if we allow for experiences to be shared by more than one persona. It could mean "belong to one of the same personas" (i.e., there is at least one persona which has both) or "belong to all the same personas" (i.e., any persona which has one has the other). The second disambiguation would give us:

> Two experiences e_1 and e_2 existing at one time belong to all the same personas if and only if they are sufficiently unified in various ways.

This can be rearranged into:

> If (and only if) two experiences e_1 and e_2 (existing at one time) are sufficiently unified, they belong to all the same personas (i.e., any persona p that undergoes one also undergoes the other).

This entails:

> If two experiences e_1 and e_2 are sufficiently unified, then any persona p that undergoes one also undergoes the other.

(p.237) Which is pretty much equivalent to the following:

> **Premise D3 (Boundedness):** For any experience e_1 belonging to a subject s, if another experience e_2 is unified with e_1, then e_2 must also be had by s.

Consider an example to bring out how the problem works: suppose we have a composite persona (p_1), undergoing experiences e_1 to e_6, and we try to say that it contains two component personas p_2 and p_3, undergoing e_1 to e_3 and e_4 to e_6, respectively. But from the fact that p_1 has all of e_1 to e_6, and from our definition, it follows that all six must be sufficiently unified (otherwise they would not all belong to a single persona). But now consider the relationship between p_2 and p_3: all of p_2's experiences are sufficiently unified with all of p_3's, and so p_2 and p_3 must be the very same persona. Consider also how p_2 (or p_3) relates to p_1: again, all of p_2's experiences are sufficiently unified with all of p_1's, and so it follows that they are the same persona. In effect, by defining personas according to unity, we have ensured that any component persona will "dissolve" into the whole.

7.1.4. Addressing the Boundary Argument

The boundary argument against psychological combinationism can be addressed by showing ways that two experiences can be sufficiently unified (to establish

the composite persona) and yet not (to keep the component personas distinct). One way to do this is to appeal to the *division-relativity* of various forms of unity, in particular functional unity, in a way that is illustrated by one of the examples discussed in chapter 6: the Nation-Brain. The radio-related experiences (pressing buttons, seeing lights) of two distant citizens are access-unified at the nation-scale but not at the human-scale: they can be jointly accessed to guide the overall functioning of the nation but not to guide the overall functioning of either citizen. Conversely each citizen's radio-related experiences are access-unified with their other, non-radio-related experiences at the human-scale but not at the nation-scale. This means that the psychological combinationist could give an analysis of the Nation-Brain very similar to the functionalist combinationist's: there is one persona for the whole nation, and within it one persona for each citizen.[2]

(p.238) Systems like the Nation-Brain, however, are not very relevant to the question of inner conflict. They rely on a difference in level so extreme that the details of the conscious goings-on at each level are more or less entirely inaccessible to subjects at the other level: the Nation-Brain has no idea it is implemented by citizens (unless we present evidence of this to its robotic avatar—perhaps by letting it tour the nation, i.e., tour its own "brain"), and the citizens have no idea they are implementing a mind (unless we show them evidence about its overall functioning—perhaps by introducing them to the avatar). Can there be component personas that operate on roughly the same scale as the whole they form? I believe there can.

There is another way to evade the boundary argument, allowing for component and composite personas even at the same scale: to appeal to the multidimensional vagueness of the above definition—the looseness of "in various ways" and "sufficiently unified." This allows for subsets of experiences, within a connected set that constitutes a persona, to be more closely connected with each other and thus constitute a component persona. That is, component personas make sense when they correspond to more integrated "clusters" within the larger conscious field.

The messiness in the constitution of personas by streams of psychologically related experiences is not unusual; it is similar to the messiness in the constitution of institutions by streams of socially related human activities, or the messiness in the constitution of organisms by streams of biologically related chemical events. And in all three cases, this multifariousness of relevant criteria and thresholds allows us to make sense of composition among the entities thus constituted.

For instance, citizens of all provinces in a federal state are socially connected in the ways that are required for them to count as members of a single political community. But citizens of each particular state are connected in those ways to a

greater degree, given at least some weightings of various factors, and their connections are sustained by distinctive causal mechanisms (e.g., state bureaucracies), and this allows us to say that they are members of a political community that does not also include members of other states. Thus there is a composite political community (the federal nation) composed of component political communities. We can also have composite organisms composed of component organisms, when the events **(p.239)** in a single component part are biologically connected to a greater degree (given at least some weightings) than they are to events in other parts. One example of such a setup would be colonial animals like the jellyfish of the order Siphonophora; another, arguably, would be multicellular organisms in general, each of whose cells is in some respects an organism in its own right. Indeed, even some organelles within our cells, like mitochondria, show organism-like features.

This multidimensional vagueness of "sufficiently unified" was already present in the notion of "psychological continuity" employed in the Neo-Lockean theory of individuation across time. Clearly, someone can lose some memories, goals, beliefs, etc. while retaining others. But how many can we lose in a short space of time before becoming a different person? As Parfit (1984, 231–233) argues, there must be borderline cases, in between those who retain the same persona despite abandoning a few plans and losing a few memories, and those who lose their persona despite retaining a few plans and keeping a few memories, where someone changes enough to no longer be clearly the same, but not enough to be clearly different.

Not only is there vagueness in the exact threshold for change of personas, but there is also vagueness about the right weighting. One person might lose all their memories of their life and circumstances while retaining allegiance to the goals, values, and projects that had previously animated them; another might keep their memories but cease to see any value in their previous goals. Which has seen the replacement of one persona with another, and which a mere change in a single persona? (Does it depend on what sort of person they were before?) This does not seem like a factual question, but rather a question of how we choose to use concepts—which sorts of continuity we decide to weight more heavily in organizing our thoughts about "people."

To see how this messiness extends to individuation at a time, consider an adjusted version of our Jekyll and Hyde case, where the two have no secrets from each other, even in their most private thoughts. When Hyde indulges in a fantasy of cruel mayhem, Jekyll must experience every grisly detail, and if Jekyll realizes that a vulnerable person is undefended in the next room, Hyde immediately and gleefully jumps into action on this belief. Judged by the standards of access-unity, their experiences are highly unified; judged by the standards of global consistency, they are highly disunified. Should we, then, count them as one or as two?[3] Or suppose that their phenomenology involves a

single field of perceptual and bodily **(p.240)** sensations, with distinct sets of urges, feelings, and plans focused on particular parts of them. When we prick the body's finger, a single experience of localized bodily pain is the focus both of Hyde's outrage and desire for vengeance and of Jekyll's perturbed contemplation of possible reasons we might have pricked him. The two might even be aware of the pain as having this dual character, since it co-presents the other's experiences to both subjects: Jekyll contemplates the causes of a pain that he experiences as prompting outrage and anger (from which he studiously distances himself), while Hyde craves revenge for a pain that he feels as calmly contemplated (while dismissing that contemplation as impotent and spineless). Here we seem to have two overlapping sets of mental states, such that it will be hard to say whether, say, Hyde's total set (including both the pain and the anger) is sufficiently unified with Jekyll's contemplation (some elements are, others are not).

It seems to me that the most perspicuous description of cases like this would be to say that there is a single person by low standards (which require only relatively low degrees of relatively few forms of unity), but two people by higher standards (which require higher degrees of more forms of unity). All the experiences associated with this body form a single loose cluster, within which two (possibly overlapping) subclusters can be distinguished. These subclusters are parts of the looser cluster, and Jekyll and Hyde, the personas they constitute, would then qualify as parts of Jekyll-Hyde, the persona it constitutes, by the standards set out in chapter 1, section 1.3. They are simultaneously existing things, in the same ontological category, such that the two parts are existentially independent (except insofar as they overlap), but the whole is existentially dependent on them together. If the Hyde experiences were not occurring (perhaps due to some sort of selective neural inhibition due to a carefully calibrated device invented by Jekyll's scientist friends), the Jekyll experiences could continue, but if neither the Hyde experiences nor the Jekyll experiences were occurring, there would be no Jekyll-Hyde experiences.

For different purposes it might make sense to focus on a different set of standards: when assigning moral responsibility for Hyde's crimes, the inconsistency of their desires might be more important, but when studying perception we might focus on the integration of their sensory experiences, and when holding a conversation their access-unity might matter most. There are innumerable ways that different relations might hold to different degrees among sets of experiences, and for different setups and different purposes there would be innumerably many different versions of the concept "subject" (or "person") that would be most useful to track. No doubt if some new, unfamiliar setup became common, there would be corresponding evolution in the terms we used to think about each other, just as **(p.241)** changes in the structure of society prompt changes in the relationship concepts that we employ.

Perhaps it would be most practically useful in everyday life, dealing with mental systems that are fairly sharply differentiated from each other, with no overlap and all the unity relations running together, to individuate personas in a maximal way, with no allowance for part-whole relations among them. But there is no principled reason not to recognize the coherence of a compositional interpretation of the clustering together of experiences in tighter and looser ways. In particular, there is no principled reason not to use this framework (as I do in section 7.4) to make literal sense of everyday talk of component personas.

7.1.5. How Far Do Persona Divisions and Substrate Divisions Line Up?

Consider the third question posed at the beginning of this section: What connection is there, if any, between the compositional structure of personas and that of substrates of experience? It is certainly possible to have one without the other. A Cartesian immaterial soul might be strictly simple, admitting of no division into component substrates, and yet be prey to the same sorts of division and dissociation as we are at the level of personas. On the other hand, a collection of objects might all support consciousness both individually and collectively but be so perfectly synchronized with one another, and underlie such a flawless and untroubled psyche, that no division into component personas could be made.

However, we humans, and other conscious evolved animals, are clearly divisible both with respect to the substrates of our experience (organs, ganglia, cells, molecules, etc.) and with respect to our personas (complex personalities, conflicting desires, identities based on different social roles, etc.). How do these two forms of division relate? We can expect them to correlate somewhat, because substrates are relevant to causal interdependence: if two experiences are grounded in the same substrate (e.g., the same brain area), they can more easily affect each other (and less easily avoid doing so), as well as tending to be affected by the same things (anything that influences that brain area). So on at least some dimensions they are likely to be more unified than they are with experiences based in a different substrate. If we weighted the particular sorts of causal interdependence associated with the sharing of substrates very highly, and made the standards for “sufficient unity” high enough, we would have a sense of “persona” such that all the experiences based on that particular substrate counted as belonging to a single distinct persona. We might then speak of the persona corresponding to, say, my left hemisphere system, constituted by all the experiences which are grounded in **(p.242)** my left hemisphere and brainstem, individuated by the particular forms of causal interdependence that hold in virtue of sharing this substrate: this persona would be a component of my total persona.

However, these personas that correspond to component substrates are gerrymandered beings, dependent on specifying the standards for identity of personas in a way that ignores all the other forms of unity that hold among my

experiences. Moreover, not all experiences can be tied directly to one part of the substrate: they might arise only from the interactions between many, and so would not belong to any of the personas corresponding to the division of the substrate into substrate parts. A final limitation of any division of personas corresponding to that of substrates is that it is likely to be undetectable to the subject themselves, and to anyone lacking a complete and sophisticated neuroscientific understanding of their individual brain. To take a simple example, it appears that the colors of different objects are processed and identified by a single brain area, while other features of the same objects are processed and identified elsewhere (see Zihl et al. 1983; Heeger et al. 1999; Théoret et al. 2002; Anzai et al. 2007). So a division of personas corresponding to substrates would involve a persona constituted by all my color experiences, unbound to shape and motion experiences, but would not involve a persona constituted by all my visual experiences of a particular object. This surprising fact is clearly not something I could tell just by reflecting on my experiences from the inside.

If we make divisions according to what seems important and introspectively detectable (like the division between my identity as a family member and my identity as a professional, for instance), we can have no idea, prior to neuroscientific investigation, whether these component personas correspond to any component substrates. We could of course attempt to define substrates corresponding to those personas, looking for the set of neurons whose activity is necessary and sufficient for the occurrence of the experiences constituting, say, my sense of identity as a family member. But here again, the result will be unhelpfully gerrymandered from the perspective of brain anatomy—it will be "one neuron here, two over here, another one there, etc." (And two distinct personas may have the very same substrate.) Because of the brain's massively interconnected character, top-down and bottom-up divisions will often diverge widely in unpredictable ways.

All this talk of picking whichever specification we want of "sufficient unity" and "various ways," thereby individuating innumerably many subtly different sets of personas, may seem oddly casual. If the psychological conception of subjects is right, then *I* am a certain persona, and my existence, boundaries, and identity across time seem to be objective facts, independent of anyone's decision about which terms or concepts to use. What I experience right now, surely, is not a **(p. 243)** matter of convention, something that could be equally well decided in many different ways. I feel the force of this intuition, but it seems to me that it is best satisfied outside the framework of the psychological conception (cf. Williams 1970, 176–177; Parfit 1984, 214, 277–280; Thomson 1997, 225), because it seems akin to the intuition that "I," the thing that has experiences, am not myself some kind of collection or system of experiences. If these intuitions are right, then the metaphysical conception of subjects is likely true, and subjects are not personas but substrates. The previous four chapters

showed how substrates can combine; here I have tried to show how personas can as well.

7.1.6. Analogies of Self and Society

Repeatedly in this chapter I will draw analogies between human beings and social groups (army battalions, bands of assassins, and political coalitions, among others). Such analogies are common and natural when thinking in this area, from Plato to the present, and for this reason they deserve some warning comments.

First, are social groups composite subjects, according to psychological combinationism? That is, are they composite personas, constituted by sets of sufficiently unified experiences, for some specification of "sufficiently"? Well, for any social group there are a set of experiences (those of its members), and those experiences interact in a variety of ways, often in ways that share information, connect representations, and partially re-create the functional roles of states like intentions, beliefs, and desires. These relations are somewhat like, though also somewhat unlike, the unity relations among an individual's experiences; in chapter 6's section 6.2 I discussed in more detail the sorts of unity-like relations found in social groups and the ways in which these relations nevertheless fall well short of what is found within a human individual. If we were willing to stretch our notion of a subject to breaking point, we might focus on the forms of unity that do hold among some or all the experiences of the members of some social group, take those forms of unity as our standard for "sufficiently unified," and thus take those experiences to constitute a composite persona, a group subject. This subject would be distinct from the group itself (the substrate of its experiences) in the same way that a human subject, according to the psychological conception, is distinct from their brain and body.

It is probably only by overextending our normal ways of individuating subjects that we could count actual social groups as composite subjects. It is better, then, simply to say this: according to psychological combinationism, social groups often exhibit a sort of clustering and organization of experiences which is similar in kind to that exhibited by a human mind. Individual human personas and the structured **(p.244)** mass of experiences that make up the conscious goings-on in a social group differ only in degree of integration, not in their core nature.

There is, however, a very important difference between social groups and human individuals, namely that social groups are divisible into parts (human beings) who *both* are separate substrates of experience and also have separate, sharply distinguishable personas. This contrasts with the cross-cutting of the divisions of a human brain into substrate parts and persona parts. There is consequently something misleading about any representation of inner conflict which pictures the conflicting parts as little people in the head, arguing or fighting or cooperating. Such representations implicitly present the component personas as

having separate bodies, i.e., separate substrates, and thus as being distinct prior to the relationships they enter into with each other. But in fact component personas are constructed by relationships among experiences and other mental states: they arise out of the interactions among brain processes rather than simply entering into those interactions. Of course something a little like this is also true of individual human beings' personas: they are formed through the social interactions that the growing human being enters into rather than preexisting those interactions. In this sense individuals are themselves social constructs. But something does preexist social life, namely the sharp difference in physical information conduits for between-brains and within-brain interactions, which makes it nearly inevitable that the experiences which arise during social life and depend on the same brain will be much more unified with one another than they are with experiences in other brains.

If we are to think of component personas in political terms, it is best not to analogize them to individual members of a society, but to political movements, clubs, or parties—subgroups within society which arise out of the patterns of interaction among human individuals. An individual's reconciliation of competing desires or identities, for instance, is not like a "social contract" we might imagine a group of castaways entering into, recognizing a need for cooperation in order to preserve their preexisting lives, because competing desires or identities are already the products of mental development and pattern formation. It is much more like an agreement between rival political parties or gangs, operating within and grounded in the social life of the same population.

7.2. Composite Subjectivity and Self-Consciousness

How do you tell which person you are? How do you distinguish your actions from events that just happen, your thoughts from other people's thoughts you merely detect, your body from someone else's body? This is a large question, and I will **(p.245)** not be able to fully answer it, but one thing that is fairly clear is that we do not, in the first instance, accomplish this discernment by first knowing things about various people, and then reasoning our way to identifying one of them as ourselves. Sometimes we do that; for instance, when watching a blurry video of several people involved in a drunken escapade, we might first identify the physical or behavioral traits each person displays, and then, based on knowing that we ourselves have certain traits, identify which of the drunks must be us. But this contrasts sharply with our everyday experience of ourselves, in which questions about which person is us generally do not even arise; in everyday life we distinguish self and other swiftly, automatically, and without thinking, independently of any prior knowledge about what sort of person we are. Each of my actions *feels mine*, for example, and the actions of others don't.

The automaticity of self-identification is what generates a puzzle for combinationists. What happens when there is a composite subject made of several component subjects, and they all distinguish self and other in roughly

the way that we do? Does each part identify itself, and distinguish itself from the other parts—which would leave the whole, which shares its experiences with those parts, with a patchwork sense of self and no experience of itself as a whole subject? Or does the whole identify itself and distinguish itself from things outside it—which would leave the parts, which share their experiences with that whole, misled about their own identity, falsely thinking of themselves as the whole? Or do both try to distinguish self and other simultaneously, leading to a contradictory and unstable sense of self?

This puzzle is particularly pressing if the combinationist wonders whether we ourselves might be composite subjects, made of simpler subjects. We seem to have generally fairly consistent self-consciousness as a whole human being: How is this established? Moreover, we generally do not have a sense of ourselves as a composite thing, a sense of our component subjects as distinct from one another: How do our parts fail to identify themselves as distinct from one another?[4]

7.2.1. An Inapplicable Solution Used by Social Groups

We can clarify the problem here by considering a social group organized to act as a collective agent,[5] because such social groups typically form a consistent sense of self in a very *different* way from anything we human individuals could be doing. **(p.246)** The individual parts first identify themselves, and then secondarily "identify with" the group while remaining aware that they are not identical with it. Seeing how this solution works for social groups, and why it cannot be extended to us, will throw into relief the combinationist's challenge in explaining how a human individual might be a composite subject.

Suppose, for instance, that an army battalion is trying to ambush the enemy; most of the group have hidden themselves in positions where they can neither see nor be seen, while a small group scouts out the enemy and relays their position to the rest. The ambush must be sprung only when the enemy "is nearby," but the representation of something as "nearby" is implicitly self-specifying, and so means different things for the scouts and for the ambushers. This poses a risk of miscommunication: if the signal "The enemy is nearby" is sent by the scouts when the enemy is near *to them*, but interpreted by the ambushers as meaning the enemy is near *to them*, the ambush will be sprung too soon and the scouting will have been useless.[6] For the signal to play its proper role, being detected and used to guide behavior ("perceived" and "believed") by the group as a whole, it matters which particular subject is implicitly represented. And if each soldier simply relies on their automatic, spontaneous sense of who they are, the resultant confusion will be disastrous.

Of course it is clear how the army battalion avoids this kind of confusion. The scouts can evaluate the enemy using the position of the ambushers as their point of reference, mentally "putting themselves" in that position, and send the "is

nearby” signal only when they perceive that the enemy is near the ambushers, not the scouts themselves. The battalion ensures that representations of “self” are consistent in which “self” they are about, by relying on each member’s explicit knowledge that although they are a member of the group, they are also an individual agent, distinct from the other members. But this solution is available only because the battalion recognizes the problem—it recognizes that it is made of members, each of which recognizes they are themselves, distinct from each other and from the battalion as a whole. They can then reflect on the content of their own experiences and how that content needs to be translated to be usable by the group as a whole. Their drive toward consistency is founded on a prior recognition of inconsistency.

But this cannot be a fully general account of how a composite subject can have a consistent sense of self, for two reasons. First, it invites the question: **(p.247)** How did the members establish the individual self-consciousness that let them recognize their distinctness from the group? Second, it seems clear that individual humans do not in fact work like this; our parts did not first recognize their distinctness from each other, and the problems it poses for coordination, and devise some policy to ensure consistency. If they did, their awareness of being multiple subjects would be inherited by the whole. The intuitive attractiveness of anti-combination reflects that if there are many communicating subjects who compose me, they have no idea that they are many. And this is something that cries out for explanation: people usually notice when they are in constant close interaction with others, so why do the conscious parts of us remain oblivious to each other?

7.2.2. The Patterning Principle and the Detection of Harmony

So combinationists must answer two related questions: How can self-consciousness in a composite subject be made consistent *prior to* any self-consciousness in the parts, and how can our parts be capable of self-consciousness *without* becoming aware of their own individuality? I propose the following answer: a given subject’s spontaneously occurring self-consciousness refers, by default, not to that subject itself but to the largest composite it belongs to whose parts are all, in a sense to be defined, “harmoniously connected.”

The composite subject which is this whole system will self-represent in the strict sense, but all the component subjects within it will refer to something larger than themselves as “me,” even when they seem to be self-representing. It follows that those component subjects, in virtue of being harmoniously connected to each other, will perceive each other as extensions of themselves. All the inputs they receive from each other they interpret as telling them either of their own voluntary actions or of events going on “in them.”

In this regard they are in something like the position opposite from a schizophrenic suffering from “thought insertion,” who perceives their own internally generated thoughts and experiences as produced by outside forces. Where the schizophrenic perceives what is actually “internal” as being “external,” the conscious parts of a normal human perceive events that are actually “external” as being “internal.” Each thus feels itself to be alone and responsible for all the mental activity in the whole system, which then inherits this unanimous judgment of solitude. (Schechter [2018, 156–180] defends an account of this sort for the split-brain phenomenon: there are two thinkers [“Lefty” and “Righty”], but they each identify themselves with the whole person, self-attributing the actions produced by the other.)

(p.248) What is harmonious connection, and why would it determine self-consciousness? Rather than defining “harmonious connection” directly, I will progressively flesh out the idea by considering how we usually form spontaneous ideas of self and other. For our parts can employ only the cognitive mechanisms that we do, and so whatever it is that lets them categorize things as internal and external, as “self” and “other,” must also be what lets us do that. If combinationists claim that our parts miscategorize each other as “self,” they should also think that we could make a similar miscategorization in the right circumstances:

> **Patterning principle:** Our spontaneous impressions of whether an event is external or internal, and of whether it is our voluntary action or not, are determined by the patterns of correspondence and divergence we detect between it and other things.

Since both internal and external events can stand in the same patterns, the patterning principle implies that we might self-ascribe actions, states, or events which in fact occur externally and independently of us, if we detect the right pattern in them. But how plausible is the patterning principle?

I think most of us are inclined, if we reflect on what determines our impressions of internality, externality, and causal responsibility, to accept some role for patterning, but only a limited one. At some point, we tend to think, we fall back on a direct metaphysical insight into certain events beings ours, either in the sense of being our actions or of being “in” our own minds. Thus I think we are normally inclined, on initial reflection, toward a “hybrid” view, with some role for detecting patterns and some role for direct insight. It is this direct insight that poses problems for combinationists: shouldn’t each of our parts know automatically that some of the thought processes that guide this human body are *theirs*, while others are not? So the strategy I will pursue on behalf of combinationism is to argue that patterning considerations can in principle entirely explain the relevant class of judgments, making direct metaphysical insight superfluous.

The most plausible role for patterning considerations is determining our impressions of the causal relations among external events. What makes it seem to us that one event we perceive is caused by another? Surely the answer has to be ultimately in terms of some sort of covariation, things either changing at the same time (or in quick succession) or remaining constant together while other things change. If I see the cup hit the floor and at the same moment I hear a noise, it will seem that the collision caused the noise. If I open the oven door and immediately start to feel heat, which ceases when I close the door, it will **(p. 249)** seem that the door being open caused the heat. If I see a candle being lit and someone across the room suddenly starts screaming, and stops only when the candle is extinguished, it will seem as though the candle was somehow hurting them. Whether we express this in the language of Bayes's Theorem or in that of Humean laws of association, the basic idea seems clear enough: we respond to regularity among the changes we perceive in the world. We might say that we observe "harmony" between events when they are correlated with each other in the relevant ways, with this term serving as a placeholder for whatever the statistical relations are that we respond to.[7]

But even given some idea of which external events are causing which, how do we identify some of these events as *our own* actions? There has been a lot of empirical work on this question, but for the most part it is accepted that we rely on considerations of patterning; the dispute, for instance, between the "comparator model" (see Helmholtz 1866; Blakemore et al. 2002; Frith 2012; Carruthers 2012) and the "multifactorial model" (see Synofzik et al. 2008; Moore and Haggard 2008; David et al. 2008) is a dispute over the particular weighting and mix of factors used, over whether there is a single privileged comparison or not. What is agreed on is that some brain system has to compute, based on signals from perception and from the internal processes that produce action, which events are "done by me." This idea is borne out by the possibility of "tricking" subjects into self-ascribing responsibility for externally caused events by manipulating their perceptions, e.g., by using mirrors to show what looks like their own hand making movements similar to those they intend (Nielsen 1963; Ramachandran and Rogers-Ramachandran 1996; Lynn et al. 2010; Ebert and Wegner 2010; cf. Wegner 2002), or by direct brain stimulation (Desmurget et al. 2009; cf. Fried et al. 1991). We might put the point by saying that we identify external events as our own voluntary actions when and only when we perceive them as harmonizing with our internal decisions and volitions: if we see our arm move just after we've consciously decided to move our arm, the harmony between these two events is what gives us our strong impression that the arm rose *because* we raised it.

But this just pushes the question back a step, to the question of why we self-ascribe the decision itself; clearly we cannot say that we regard it as our own decision because it harmonizes with a prior decision to make that decision!

(p.250) 7.2.3. The Patterning Principle and Internal Events

How can the patterning principle explain our self-ascription of internal events? Some internal events, like imaginations, may fit neatly into the same model as external events. Some experiments suggest that faint sensory stimuli may be miscategorized as imaginary when they match what subjects were independently attempting to imagine, i.e., when they harmonize with a prior intention (Perky 1910; cf. Segal 1972). But many of the events we experience as being "in our minds" are not preceded by any distinct decision to produce them—our decisions, thoughts, impulses, and so on seem to be "ours" *on their own*, with no need for us to compare them with any other events. How can the notion of harmony even apply here? I think there are actually at least three ways for the patterning principle to cover these internal events: harmony with background psychology, lack of resistance to the will, and dedicated comparator mechanisms.

First, we might self-ascribe events not just when they harmonize with particular other events, but also when they harmonize with our background psychology, reflecting the fact that our mental events are typically "caused by a combination of our background beliefs, desires, and interests" (Campbell 1999, 617). We might focus in particular on the mass of dispositional intentions, desires, and goals that could be called "the background will" (as contrasted with "the occurrent will," our present conscious feelings of desire and intention). The background will is the underlying structure of what we want or intend to do in different situations, given different contingencies, when confronted with different stimuli. It is probably impossible to ever fully articulate this structure—we cannot write out a list of what we would want or do in every possible situation—but there does seem to be a set of persistent, determinate facts about us in virtue of which we are inclined to will some things, and in virtue of which we perceive things as unwanted or desirable, frustrating or welcome. I may have never thought before about whether I would like to receive a free pair of glittery gold sneakers or to have all my fingernails double in length, but immediately upon such things happening I would have quite definite reactions, positive or negative. It seems clear that even before they happen, it is an objective fact about me that I "would like" or "would not like" such things—such facts about me are my "background will."

(Note that we are not considering the obviously regressive idea that a subject judges some thoughts to be theirs because of its fitting well with the background psychology that they have already *judged* to be theirs; the background psychology need not be self-ascribed, indeed need not be the object of any kind of thought or awareness, in order to govern and guide what self-ascriptions are in fact made. This evaluation of harmony could be, and probably is, largely inaccessible to reflection: **(p.251)** sometimes the only way we can find out what

it is we really want is to expose ourselves to actual or hypothetical cases and see how we react.)

Nevertheless this proposal can be only part of the story, since we can regard a thought as "ours" despite its content being wildly out of character, just as we can think that someone *else* is thinking exactly what we would think in the circumstances (cf. Gerrans 2001, 233ff.; Schechter 2018, 156–180). Being "in character" for us seems neither necessary nor sufficient for being perceived as "ours." Thus the defender of the patterning principle should supplement the appeal to background psychology with other factors.

A second option is to appeal to a sort of *instability*: the thoughts and feelings which we ascribe to ourselves flow into one another, affecting and being affected by other things we self-ascribe, and by our own decisions and wishes, while external things remain fixed as our attention, plans, or ideas flow over them. Contrast a perception of a red square with a mental image of a red square (even one that arises in the mind unbidden): the latter will shift, recede, disappear, or transform according as I attend to it, ignore it, suppress it, connect it with something else, consider something related or something unrelated, and so on. And those changes which thus affect it are themselves similarly mutable: while we do not choose each step, we can intensify or inhibit things voluntarily if we try to. By contrast, the perceived square either stays constant during all those fluctuations, or else changes in a manner uncorrelated with them.[8] We might say that while neither is positively voluntary, the perception *resists* my will, whereas the idea does not. (This idea has a long history: see Descartes 1985, 2:26–27, 55; Locke 1836, 484; Berkeley 2008, 41.)[9]

A final idea is that the "efferent copy" mechanism employed in the monitoring of motor actions could also be present with thoughts, feelings, and other mental events. Feinberg (1978) and Campbell (1999) both suggest this, advocating regarding thought "as a motor process." The idea is that the brain processes that produce actions or thoughts do not just produce those actions or thoughts, but also produce "efferent copies," signals reporting that such-and-such an action or **(p.252)** thought has been produced. These copies are processed by a "comparator," some brain system that also takes in feedback from the actual execution of these actions and thoughts, and from other internal and external events, and monitors the correspondence or lack of correspondence among them. Since the comparator could operate below the level of conscious awareness (cf. Campbell 1999, 617–618), it would easily account for cases where something seems internal to us despite not cohering with other consciously accessible elements of our minds. But it might seem extravagant for every single conscious event to be part of such a monitoring process, with every whim, twinge, flicker of doubt, or snatch of memory generating a copy to be submitted to a comparator mechanism. If so, we might posit a comparator-type monitoring

mechanism for some but not all mental events, relying on the first two proposals to explain our impressions of the others.

7.2.4. Harmony and Harmonious Connection

I conclude that the sort of "harmony" which leads us to regard events as occurring "in our own mind" can be something less than conformity to a definite prior decision: it can be any combination of conformity to background psychology (especially the background will), sensitivity to a willful stream, or the verdict of a dedicated comparator mechanism, with the particular balance of factors being a topic for empirical research. I will use the term "volitional harmony" for the specific sort of harmony which our minds look for when categorizing things as internal or external (as opposed to when discerning causal relations among external events). The term is still basically a placeholder; we do not know exactly what balance of factors is involved, nor what mechanisms implement and determine volitional harmony in the actual human brain. But it is a placeholder whose place is circumscribed by our limited but nontrivial scientific understanding of how human beings distinguish self and other.

If the patterning principle is true, and the relevant "patterns" are something like what I have called "volitional harmony," it follows that whether some thought or action seems spontaneously to be "mine" depends not on whether it really is, but on its relationship to my will (both occurrent and dispositional) and to other things which seem to me to be "mine." If two subjects, who discern self and other using the same basic mechanisms we do, are interacting harmoniously, so that what each does harmonizes with the psychology of the other, then they will each self-ascribe everything the other does that they are aware of. Let us say that subjects are "harmoniously connected" if they are set up so as to interact in this way all the time: then two harmoniously connected subjects will go their whole lives without ever realizing they are not alone. The self-representations they both **(p.253)** form will be based, not specifically on what is true of them, but on what is true of the system that comprises them together with any other subjects they are harmoniously connected to.

I think it is plausible that the parts of a human being are harmoniously connected. The electrical behavior of a brain part, and the physiological structure that underlies that behavior, is very sensitive to that of surrounding parts and has developed in constant interaction with them for years. Moreover, they are alike in their biological requirements (pH, temperature, salinity, etc.), their ways of coding information, the timescale on which they act, and so on. It is hard to imagine better conditions for harmonious connection.

7.3. Composite Subjectivity and Agency

There is a special challenge to combinationism that challenges whether a composite subject who *acts*, in the sense of being genuinely responsible for the

things they do, can be composed of parts which also act. Horgan (2007, 190), for instance, points out that “experiencing one’s behavior as produced by oneself [i.e., as one’s own action] is fundamentally different from experiencing it as caused by internal states of oneself,” and if this is true of one’s own mental *states*, it seems even more true of one’s own mental *parts*. So the combinationist needs to explain how action by our mental parts can ever be enough to really count as action *by us*. (This objection is posed specifically against constitutive panpsychism by Mørch 2014, 198–201.)

This challenge should be distinguished from a couple of nearby issues that are not specific to combinationism, but which also deal with whether our everyday experience of agency is veridical or some form of illusion. The most famous and long-standing issue is whether we act *freely*, and whether our doing so is compatible with our decisions being the effects of prior causes. In the twentieth century, philosophers have also worried about the “causal exclusion” problem mentioned briefly in chapter 1, of whether in the law-governed world revealed by physics, there is room for our decisions, and other mental states, to be causes of behavior at all, free or otherwise: if our behavior is fully explained by the operation of fundamental physical laws governing the microscopic parts of our brain and body, then what work is left for our mental states to do? I will not try to solve these much-discussed problems here;[10] rather, I will try to defuse the specific worry that parts and wholes are always in competition for agency. To do this I will consider **(p.254)** two sorts of relation which are in some ways analogous to the relation between composite and component subjects: the relation between a subject and their own mental states, and the relation between multiple human beings acting at once. In both cases it seems that agential exclusion is sometimes a problem but sometimes not: by considering what makes the difference in these more familiar cases, we may get an idea of what is needed for conscious parts and wholes not to be in agential competition with one another.

7.3.1. Agential Competition between a Person and Their States

Consider some ways that a movement of my body might fail to be my own action despite being caused by states of me, inspired by Davidson’s (1973, 79–80) famous case. A mountain climber is holding his friend on a rope during a climb, and he releases his hold so that his friend falls to his death. But how exactly did this effect come about? Perhaps a gust of wind buffets him so hard that he cannot keep his grip; then releasing his friend is not an action, but something that happens. Or perhaps he decides that, to lighten his load and improve his chances, he will drop his companion. In both cases there are features of him that are causally relevant (e.g., how strong his finger muscles are), but only in the latter case would we call it an *action*, something caused by him as an agent. Davidson famously makes the point that this difference is not simply about whether the cause was a mental state of his: he might let go inadvertently as a result of being startled or afraid, and this would seem to be causation by his

mental state but still not action by him. It is not even enough for the cause to be a mental state that rationalized the action: the thought “I could improve my own chances by dropping my friend” might occur to him, and unnerve him so much that he inadvertently drops his friend—again this seems not to be an action attributable to the climber himself. To qualify as his own action, the climber’s mental state must cause the movement of his fingers “in the right way.” Spelling out this “right way” precisely is not at all easy, but fortunately the combinationist does not need to do so; they simply need to show that there is at least some plausible way of distinguishing cases where the causal role of a subject’s mental state does, and does not, exclude the person as a whole from agency, such that conscious parts could also sometimes exclude the whole from agency, and sometimes not.

A first pass might be that in the agential case the action is a result not just of one stray mental state but of the agent’s whole set of mental states, of their “whole mind” and thus of “them as a whole.” But of course it’s rarely, if ever, true that **(p.255)** every single mental state someone is in makes a difference to their action (most just won’t be relevant), and certainly not every single mental state makes a difference in our selfish mountain climber who deliberately drops his friend.

One key difference between the case where the mountain climber seems to act and the cases where he does not is that in the former he makes a *decision* to drop his friend. By ‘decision’ I mean that there is a process which is both potentially open to many considerations (e.g., motivations to drop or not drop the friend, ideas about alternative options available) and also “unified” in the sense that it generally issues in a consistent plan of action (even if the climber has many conflicting desires to do different, incompatible things, he cannot decide to do them all—he must pick). When actions arise from a mechanism like this, which allows many different motivations to “have their say” but enforces singleness of action, it seems intuitively right to say that the action reflects not just the one mental state that directly motivated it but also all those which “had their say.” Of each of the climber’s mental states we can at least say that *if* it had been relevant to whether or not to drop the friend, it was in a position to make a difference to the action (which we cannot say about the mental states of other people). In this sense it seems that we can say, roughly, that the deliberate action reflected *all* the climbers’ mental states, but when they are startled by an unnerving thought, their behavior reflects only one “stray” mental state.

Could we say then that something counts as someone’s action only when it results from a decision they make? No, because so many of our actions are not preceded by a conscious decision, and indeed happen too swiftly and fluidly to be deliberated about. But the “openness” characteristic of decision-making, the sense that any other mental state which was relevant could make a difference to whether the action was performed, *is* still present in such cases. If we found that

our fluid everyday actions (our walking, sitting, picking things up, glancing around, and so on) kept happening even when we formed new beliefs or desires which directly spoke against performing them, we would feel vividly as though we were suddenly no longer in control of them, were not acting but just seeing our bodies move. So it might not be too far off the mark to say that something is our action if it is produced by a process that, like conscious decision-making, allows for all of our mental states to have input insofar as they are relevant to it.

One attractive feature of this suggestion is that it ties agency to the defining structure of a persona. According to psychological combinationism, what defines a persona is some set of unity relations among mental states—coherence, interdependence, mutual access, and so on. The details of this structure are variable and complex, but it is clearly bound up with what allows many different mental states to feed into a decision-making process that issues in a single united action. By **(p.256)** their mutual submission to this decision-making process, different motivations are made interdependent and coherent with one another, and this mutual submission plausibly requires that they already be mutually accessible. If this is so, then the structure that defines a persona (sufficient unity of various sorts) plays a crucial causal role, and it seems to be just in cases like this that we would intuitively count the agent themselves as having acted.

This is not a full solution to the challenge of defining agency. For one thing, I have not defined "relevance"; for another, what distinguishes the interdependence involved in a decision from that involved in, say, someone's mood? (Perhaps the climber would not have been startled enough to drop his friend if he had been more cheerful and relaxed, and many different mental states could have made a difference to how relaxed he was feeling.) My aim is just to show that something's being caused by a state of me is sometimes compatible with its being my action, and sometimes not, and to gesture in the direction of the sort of factors that make the difference. This will help to illuminate how something's being caused by a *part* of me can also sometimes be compatible with its being my action, and sometimes not, and what makes the difference.

7.3.2. Agential Competition between Discrete Agents

There is obviously much more to say about the role of psychological structure in agency, but this is not the place to say it. What I have said above is not very surprising and is not specific to combinationism. What is specific to combinationism is that the question changes from "Is there any real agency?" to "Is it *my* agency or *my parts'* agency?" Surely something cannot be my action if it is *someone else's* action?

However, reflection on various forms of social cooperation shows that distinct agents are not always in competition for agency. For example, suppose A hires B to murder C, knowing that B is a very reliable assassin. A is culpable for C's resulting murder, even though their effecting this murder went through B: B is a conduit of A's agency just as a gun or knife would be. But B is also culpable: making themselves a conduit of A's agency does not erase their own. Of course there are circumstances in which the eventual act of murder might seem less fully attributable to one or the other of them, such as if B was coerced into carrying out A's orders, or if A did not really expect B to follow those orders. But at least in some cases it makes sense to regard them both as having agentially caused that outcome, and to that extent as not being in causal or agential competition.

The above example involves something like what in chapter 6 I called "authority relations": one person (B) takes on the expressed intention of another (A) and **(p.257)** conforms their own intention to it. But a similar kind of agential noncompetition is plausibly present also in joint action and in actions based on joint commitment. When a team of assassins work together to kill a target, operating based either on their shared intention "that we kill this target" (a joint intention) or on having each committed publicly to the goal of their doing so (a joint commitment), it may be basically random which of them ends up striking the killing blow—and whichever of them does so may well have depended, both for their intention and for their success, on the others. In a case like this it seems artificial to insist on calling one the agent and the others mere accomplices (even if making such a distinction may sometimes be of social use).[11]

In each of these cases, where distinct agents appear not to be in agential competition, there seems to be a kind of "alignment" among the wills of the distinct agents. In authority relations this is asymmetrically established: the obeyer aligns their will to that of the commander. In joint action, the process is multilateral: each member of the group aligns their will to that of the others, until they all share a certain intention. In joint commitment, the alignment is indirect: rather than participants aligning their will to what the other participants intend, they each align theirs to the same public idea, like a constitution, a law, or just a verbal statement. Although there are significant differences in how people's wills come to be aligned in different cases, this common thread is crucial: different agents come to share agential responsibility for particular actions by aligning their wills with each other's.

This answer dovetails with the above suggestion that someone is responsible for the effect their states bring about, if they bring it about through a mechanism that is open to influence from all their other states—if they bring it about, we might say, in a unified way. We can now suggest that mechanisms which display this openness also serve to align the wills of different subjects whose states

enter into them. If part of me wants money and part of me wants to relax, the process by which I reach a decision about how to balance these goals is also a process which brings the wills of these two parts of me into alignment. The actions produced by that mechanism are then both attributable to me as a whole (and not just to states of me), and also attributable to both those parts of me.[12]

(p.258) 7.3.3. Alignment of Wills between Whole and Part

In the previous section I defined volitional *harmony* as whatever structural relation of congruence among the wills of distinct subjects served to determine whether they would perceive each other's actions as "mine" or as "someone else's." Two subjects which consistently interact in volitionally harmonious ways are "harmoniously connected," and it is the harmonious connection between the parts of each human brain that stops them from recognizing each other as distinct. What is the relationship between volitional harmony and alignment of wills?

The alignment of wills in various forms of agential cooperation is clearly not harmonious connection, since the participants are fully aware of their distinctness from each other. In part this is because they have had so many other occasions, acting individually, to learn their own identities; in part this is because even when two people share an intention, and communicate regularly on how to implement it, much of the detail of each one's action will still be independent of the other's intentions. (I may tell you to "go over there quickly," but I can hardly specify the particular strides you should take.) Perhaps if these two deficiencies were made up—if some people acted together always, and determined every relevant detail of their actions jointly—the result would qualify as harmonious connection, and they would never need to form individual self-concepts.[13] Alternatively, perhaps harmonious connection requires something more than the alignment of wills involved in things like joint action; I cannot say for sure because both notions are still placeholders, awaiting further empirical and conceptual investigation. But harmonious connection certainly seems *sufficient* for alignment of wills: alignment of wills involves creating, briefly and perhaps in an attenuated sense, something like the volitional harmony that constantly pervades a normal human brain.[14] If this kind of alignment of wills is the key thing that lets distinct agents share responsibility for the same effect, then the harmoniously connected parts of a human brain should eminently qualify for such shared responsibility. They are "working **(p.259)** together" in such a way that they can share responsibility for the actions they jointly produce.[15]

This is not yet an answer to our question: Even if the parts can share responsibility *with each other* for some or all of their joint effects, what about the whole they form? Though it may sound strange, I think the same basic answer applies: a whole is not in agential competition with its parts as long as its will and their wills are aligned. When one part's will diverges from that of the

whole, there is a meaningful question of whether something is the action of one or of the other. When they do not diverge, the question can be answered only "both."

This alignment of wills between whole and part is not quite a matter of authoritative command, or joint action, or any other relation that obtains between discrete agents, because the two wills are constitutively related: the whole's will is grounded in the existence of, and relations among, its parts, and by extension in their wills insofar as those wills are relevant to how they relate. But this does not make alignment of wills automatic: one part's will might diverge considerably from that of all the others, and thus at least partly diverge from the whole's will. In such cases, the whole would experience a strange impression of acting and yet not acting; this is the experience of inner conflict, which I discuss in the next section.

7.4. Composite Subjectivity and Inner Conflict

What should we make of the experience of "inner conflict"? I mean here the broad class of cases in which we would find it natural to say things like "I wasn't myself," "I couldn't control myself," or "I'm at war with myself," or in which we would describe someone as struggling against some mental state of theirs, like addiction, temptation, compulsion, instinct, or conditioning. Cases like this, of varying degrees of severity, have prompted metaphors and theories involving a mind with parts which can oppose one another and struggle for control either against each other or against the person themselves. Such talk is often not interpreted literally, and surely is often not intended literally, in part because it is very hard to make sense of it literally—what are these things that must be fought, and who is it who must fight them?

This is where combinationism can help. I will not try to argue that any particular instance of "inner conflict" talk should be taken literally; rather I will simply ask: What *could* it mean, if it were?

(p.260) 7.4.1. What Are the Parties to an Inner Conflict?

Let us first ask: What, in the most general terms, do people find themselves in inner conflict with—what aspects of their mental life tend to generate this kind of experience? I think the two key factors are (i) incompatible desires and (ii) a "failure of reconciliation" among those desires.

By "incompatible desires" I mean simply desires which, given the agent's perceived situation, cannot both be satisfied in any foreseeable way, but can each be satisfied by some course of action that frustrates the other. They "pull in different directions," motivating incompatible courses of action. But incompatible desires by themselves are not enough for inner conflict; our desires are routinely incompatible in various ways, and while we often regret not being able to satisfy them all, we usually find it easy enough to prioritize, pick a preferred course of action whose benefits outweigh its costs, and carry it out.

Inner conflict involves what I call a failure of reconciliation. But what does this mean?

Successful “reconciliation” need not involve the losing desire disappearing, or being revised so as to no longer be incompatible with the winning one (though sometimes it does). Often we continue to feel both desires even while forming and unproblematically executing a plan of action that satisfies only one. What is important is severing any direct link between desire and action: the losing desire is reconciled to the decision when it is denied any power to produce actions in the direction of its satisfaction. When reconciliation fails, the losing desire controls or constrains action in defiance of the decision that frustrates it, either by causing actions which satisfy that desire, or by preventing actions which decisively frustrate it. For example, a heroin addict who has decided to quit cold turkey may find both that their desire for heroin can make them actively consume heroin “in spite of themselves,” and also that this desire can prevent them from throwing their heroin away.

Failure of reconciliation seems to be made more likely both by the strength of the desires that must be frustrated and by the relative causal independence of the conflicting desires. Desires which arise from the same sets of mental mechanisms, which “stand and fall together,” seem easier to reconcile with one another and less often give rise to inner conflict. By contrast, inner conflict more often arises when two conflicting desires have very different causal bases; for example, classic cases of being overcome by temptation usually involve one of the following three things: a desire based on something bringing pleasure (e.g., desire for food or rest) conflicting with a desire not associated with pleasure (e.g., desire to conform to a training regimen); a desire which is strengthened by stimuli in the immediate environment (e.g., desire for some money right in front of one) conflicting with a **(p.261)** desire which requires actively “remembering” something distant or abstract (e.g., desire to avoid pain the next day); or a desire based in simple, developmentally and/or evolutionarily old systems (e.g., danger detection) conflicting with a desire based in later-developing, more distinctively human systems (e.g., adherence to social rules). And plausibly, addiction is so prone to generate inner conflict precisely because it involves a chemical mechanism for forming and maintaining desires independently of the rest of the subject’s psychology.

So let us say that inner conflict involves two or more mental things of some sort, distinguished by being associated with incompatible desires which are sufficiently strong or causally independent for reconciliation to fail.

7.4.2. Inner Conflict as Involving Component Subjects

Let us now suppose, as a combinationist might, that these two opposing mental things are conscious subjects, each a part of the whole person, and the sense in which they are associated with these desires is by “having” them in the same

sense that the whole person does. What would follow from such a supposition, given the claims of sections 7.2 and 7.3, that our impressions of self and other track volitional harmony, and that agential competition is avoided by alignment of wills?

It seems clear that a failure of reconciliation involves a failure to align wills, and thus a failure of volitional harmony. The two putative parts have different desires, and neither accepts restriction by the other in their capacity to produce action. They are like two people who not only want incompatible things but insist on pursuing those different things even when it means subverting or opposing the other's efforts. This failure may be only partial, mitigated by alignment of wills on many other matters—rather like squabbling members of a political coalition who nevertheless remain able to act jointly on some issues. Subjects whose wills are not aligned clearly cannot share responsibility for the actions they produce. If my addictive desire for heroin drives me to consume it in spite of the strenuous efforts of "that part of me which wants to quit," the latter entity by supposition is not the agent of the consuming. (Conversely, the addictive part is not the agent of my throwing away my heroin.)

The more significant question is whether the whole subject shares responsibility for such actions, which turns on whether the whole subject's will coheres with that of the particular part responsible, which in turn depends on what counts as the whole subject's "will." And in a case like this, where the whole harbors conflicting desires and cannot reconcile them under any single decision-making mechanism, it is not clear that there will be a single determinate thing answering the description "the whole subject's will." Given this, it will likely be somewhat indeterminate **(p.262)** whether the wills of the whole and of a given part are aligned enough to share responsibility, and thus somewhat indeterminate whether the person themselves has acted when that part of them produces an action.

Consider another version of Davidson's mountain climber, who has weighed up their various motives for and against dropping their friend, and found the reasons against far more compelling. Suppose they form a settled intention not to drop their friend, but nevertheless at the last moment find the temptation to save themselves by lightening their load too strong to resist, and drop their friend in a "moment of weakness." The sharp divergence between one element of their psyche and the decision reached by the person as a whole prevents us from seeing this act as equally fully theirs and their parts': it is partly theirs (unlike the case where the same thought or desire simply unnerves them enough to loosen their grip) because it operated through their psychological mechanisms for decision-making and rational action, but not fully theirs because it disrupts and usurps those mechanisms without submitting to the proper protocols.

Of course this result is a matter of degree. If the two conflicting parts are roughly equal in various ways (neither has more control of the decision-making mechanisms, and neither shares more of the whole's desires), it will be harder to say nonarbitrarily that the whole's will either fits or conflicts with either part's will. If one of the parts is considerably "larger" in the sense of comprising far more of the subject's desires—in particular, if all their desires except one can be reconciled to the verdicts of a single decision-making mechanism—then we should think of the whole subject's will as approximating the will of that part, so that whatever that part does is done by the whole subject. The smaller, more "isolated" part, by contrast, is in conflict with the will of the whole (because it is in conflict with the will of the other, larger, part) and so the whole can more reasonably disavow the actions it produces.[16]

Another implication of supposing the mental parts involved in inner conflict to be subjects is that they will recognize each other as distinct, because their conflict **(p.263)** involves a lack of volitional harmony. By the arguments of section 7.2, the effects these component subjects have on one another will appear to each as alien, not as their own actions. (Of course this loss may be only partial, and in a healthy person's life it occurs against the background of a long history of unified self-perception, in which all the person's parts have seen each other as extensions of themselves.) Then the whole, inheriting the judgments and perceptions of each part, will get the following impression (where parts 1 and 2 act on desires A and B):

- I am the one who acts on desire A (inherited from part 1).
- I am not the one who frustrates desire A by acting on desire B (inherited from part 1).
- I am the one who acts on desire B (inherited from part 2).
- I am not the one who frustrates desire B by acting on desire A (inherited from part 2).

For example, suppose A and B are the desire to indulge an addiction and the desire to break an addiction, respectively; the composite subject would then simultaneously feel themselves to be the one trying to indulge the addiction, the one frustrating that effort, the one trying to break the addiction, and the one frustrating that effort.

This inconsistent impression of one's own identity seems to me to accurately capture the phenomenology of inner conflict: there seem to be two distinct things, and yet I seem to be both of them. Of course this contradictory impression may be more or less sharp; as well as fully resolute inner conflict, with both sides giving their all, there are many varieties of half-hearted allowing-oneself-to-succumb, and in the latter case there will be intermediate degrees of volitional harmony, yielding odd impressions as of something's being somewhere

between self and other, in ways whose details deserve more empirical and phenomenological study than can be given here.

7.4.3. Inner Conflict among Substrates and among Personas

But if the mental parts involved in inner conflict are subjects, what sort of subjects are they—brain lobes, neural networks, subpersonalities, or what? In light of the reasonable disagreement over the essence of subjecthood, we should try construing them both as substrates and as personas. It is certainly possible for two substrates of experience to be related in such a way that they form a single subject which experiences inner conflict in the way that we do. Plato (1997, 246a–254e) famously imagines the soul of an internally conflicted person as a team composed **(p.264)** of two horses and a man, trying to work together but only sometimes succeeding; if such a team did exist, and its members stood in such intimate relations that they generally perceived each other as extensions of themselves, then we would have a case where the structure of human inner conflict is realized by a collection of discrete substrates (namely, three organisms). But in fact it seems unlikely that the actual brain is organized like this, with separate lobes responsible for different goals and desires. The brain's division seems much more to be functional, with different areas dedicated to different sorts of information processing that can subserve many different desires (e.g., both the desire for money and the desire for self-respect can influence and make use of visual perception, auditory perception, imagination, strategic thought, etc.). Given this, it seems that if we identify the substrate for the desires associated with one side of an inner conflict, it may well turn out to also be the substrate for the desires associated with the other, or at least to overlap significantly with it. If our aim is to illuminate the conflict between these two subjects, these considerations count against construing them as substrates.

Can we, instead, make sense of the subjects involved in inner conflict as personas? Yes, and the very features that characterize inner conflict (incompatible desires and failure of reconciliation) provide a guide on how to do so. Their incompatible desires, and consequent incompatible drives to action, constitute an important failure of global consistency between the contents of their mental states; the failure of reconciliation indicates a lack of causal interdependence. If we weight these forms of unity highly, we will find two sharply distinguished clusters of experiences within the overall consciousness of the conflicted person. By a high-threshold definition of personas, these clusters will constitute two distinct personas, contained within the whole persona constituted by the whole cluster which meets a lower-threshold definition of personas. And this will be true whatever the division of the underlying substrates—even if these different personas arise out of activity of the same brain areas.

So it does make sense to interpret inner conflict as involving a literal conflict between distinct subjects. Of course there are some major caveats to this literal construal: the distinct parts are subjects only according to the psychological, not the metaphysical, conception; they are psychological subjects only given particular ways of weighting forms of unity, and particular thresholds for these forms; and, most interestingly, they may be sharply distinguished only "on one border," so to speak. This is because the forms and degrees of unity that define them may not be transitive: experience A might be largely consistent with, and interdependent with, experience B, which is likewise consistent and interdependent with experience C, even though A and C are too contradictory or independent to meet the **(p.265)** same standards. The two incompatible desires involved in inner conflict might therefore both qualify as sufficiently unified with many of the same other desires and experiences, despite not being sufficiently unified with one another. We would then have difficulty individuating the personas involved; there could be thought to be one, or to be two, with the choice between those descriptions being to some extent arbitrary (cf. the problem posed by "ring species" for the biological definition of a species, e.g., Moritz et al. 1992).

7.4.4. Inner Conflict and Identity over Time

Further complexities arise when we consider identity across time: How long do the component personas involved in inner conflict survive? If the same pattern of conflict recurs frequently (e.g., someone is driven to self-destructive acts by uncontrollable anger every time they feel insulted), the Neo-Lockean account of personal identity suggests that the rival parts on each occasion are identical to those on the other occasions. The defiant part on one occasion is both very similar to and causally dependent on the defiant part on the previous occasion; that is, the two are psychologically continuous in the same way that the whole persona is psychologically continuous across time. This is why it can make sense to think of someone as engaged in an ongoing, recurrent struggle with a persisting part of themselves (e.g., "their anger").

But what about times when there is no inner conflict, either because the stimulus is absent or because the person has been able to achieve reconciliation among all their desires? Does the defiant persona (e.g., the person's "anger") still exist—that is, is it identical to (by being psychologically continuous with) any persona which exists at this time? On the one hand, it may seem not, because on those more peaceful occasions there is no subcluster of experiences that stands out from the rest on account of being more tightly unified—not because there has been any loosening of the unity that had previously tended to obtain between the experiences associated with the defiant part (e.g., the person's perceptions of offense, impulses to become angry, feelings of simmering grudgeful resentment against particular people), but because of a strengthening of the unity between them and other experiences. That is, the component persona seems to have disappeared by "dissolving into" the whole, rather like a

black-and-white drawing which is “erased” by filling in the white spaces with black ink.

On the other hand, we seem to miss something if we just say that the angry subpersona has ceased to exist. That would imply that if the person ever achieves full integration, so as to no longer be troubled by inner conflict, their component personas have been permanently destroyed—turning what seems like a happy **(p.266)** reconciliation into a sort of intrapsychic fratricide. We want to say that integrating all aspects of one’s personality is, at least sometimes, a laudable and rational aim, not that it is rational for conflicting component personas to obstruct integration as a means of self-preservation. One thing we could try saying is that whenever inner conflict is not occurring, the component persona is identical to the whole person. After all, there is no shortage of psychological continuity between the two, even if the inner conflict involves a limited form of discontinuity. But this turns the process of overcoming inner conflict (or just temporarily ameliorating it) into a “fusion case” of the sort which has proven endlessly paradoxical for theories of personal identity. The basic problem is that if the part has become the whole, then they are the same thing. But the whole was the whole all along, and if it’s the same thing as the part, then the part was the whole all along. Yet we started by saying that it was not. This sort of logical problem arises whenever two things go from being distinct to being identical; I will not here attempt to solve it (though in chapter 8 I suggest some ways that combinationism may affect the plausibility of different solutions), only to note the present case as one instance of this general form.

A final possibility is to keep the component personas in existence, but also keep them distinct from the whole, by saying that even when no inner conflict is occurring—even when all unity relations hold just as strongly between the experiences constituting that component persona and those not—the component persona still exists just *because* it existed earlier and was distinguishable from its surroundings then. That is, we might permit ourselves to individuate entities by first looking for patterns that stand out at a particular time, and then tracing the development of that pattern over time, continuing to see it as existing even when it is better connected with its surroundings and thus no longer stands out. After all, personas are not metaphysical bedrock; they are high-level patterns, and it is perfectly appropriate for us to construct our criteria and definitions for their individuation in a flexible way, to enable us to capture whichever high-level patterns we have reason to care about.

7.4.5. Inner Conflict and Dissociation

Finally, what about dissociative identity disorder (DID), in which a single body manifests multiple “alters,” rather like Jekyll and Hyde in our earlier thought experiment? DID is usually treated as a disorder afflicting a single individual, whose therapeutic goal should be integration of all their alters so as to no longer have DID. Yet on the face of it, psychological accounts of personal identity, like

the Neo-Lockean accounts discussed earlier, imply that the alters are distinct people, and **(p.267)** thus entitled to be recognized and respected as such. This tension between popular philosophical accounts of personal identity and the practical approach generally taken to DID has been recognized (e.g., Bayne 2002), and I believe that one of the merits of combinationism is that it can simultaneously do justice to both sides. DID is indeed a condition affecting a single individual,[17] while at the same time involving a literal plurality of component individuals. In fact, combinationism can make sense of DID as just a radical extension of ordinary inner conflict: two or more sets of brain processes consistently give rise to clusters of experiences that are well-unified internally, but strikingly disunified with those arising from the other processes. Because the personas constituted by these clusters at different times are psychologically continuous, they should be recognized as real, distinct, enduring subjects—at least on the psychological conception of what subjects are. If integration is achieved, the same three options as above are available for describing what has happened: that the alters no longer exist, that they are now all identical to the whole subject and thus to each other, and that they still exist but without standing out from the rest of the subject's mental life as distinct entities.

The biggest difficulty in giving this combinationist account of DID is explaining why the disunity involved is not just volitional but also extends to thought, memory, and knowledge. Some hypotheses link DID with traumatic experiences in childhood (see, e.g., Ellason et al. 1996), which would make sense if volitional disunity is the crucial factor. Intensely unpleasant experiences are ones whose subject intensely desires not to be having them, and thus for another subject to volitionally harmonize with them will require the latter to come to also intensely desire that this ongoing state not obtain—which subjects will understandably resist doing, since it amounts to taking on some of the other's intense unhappiness. But how this leads into a situation where the different alters are unaware of each other, and lack memory access to each other's experiences, is much less clear. Indeed the whole topic of how cognition relates to the will is a perplexing one: in general we maintain a clear distinction between what we want to be true and what we think is true, but our occasional failures (manifested by "wishful thinking" on the one hand, and "adaptive preferences" on the other) show that this distinction is not guaranteed. Moreover, DID need not always originate in trauma and may have multiple different sorts of cause. To say anything positive here would be to speculate about a matter best studied empirically.

(p.268) 7.5. Conclusions

A division into component personas works differently from a division of the neural substrates of experience, in that it looks at the pattern of relations among experiences even if they bear little resemblance to the structure of the underlying mechanisms that enable them. Because it is defined by the high-level information processing that enables introspection, agency, and self-

consciousness, it is more readily visible to us introspectively, and to the view of others who observe our behavior. The occasions when we spontaneously find it natural and useful to think of people as having person-like parts will tend to be of this sort, involving component personas, not component substrates.

Nevertheless the combining of personas faces the same five internal problems as the combining of substrates, most especially the boundary argument, which I addressed in section 7.1. It also, however, throws into relief some special bridging problems concerning self-consciousness and agency, defining characteristics of the special sort of subjects we call “persons.”

The problem around self-consciousness was this: How could any composite whose parts were all self-conscious achieve self-consciousness as a single whole? And would it not automatically thereby be aware of its own compositeness, in a way which we humans seem not to be? Yet if some composite subjects (like us, presumably) do not have individually self-consciousness parts, why don’t those parts reflect on and identify themselves as distinct?

The problem around agency was this: When my body moves, is it me or some part of me that performs this action? Intuitively we would like to say that I am the agent, even if parts of me participate as well, but how is this compatibility to be ensured? Doesn’t the causation of my actions by some smaller agent within me imply a lack of control on my part?

In trying to solve these problems, I gave a central role to the volitional relations among component subjects, the way that the will of one relates to the will of the other. In so doing I could not help but frame individual reflection and decision-making as something like a social process, a meeting and conversation among many beings. I am thus led toward the idea that “the higher mental functions have their origin in and, therefore, share important features with, interpersonal activity” (Fernyhough 1996, 48; cf. Gregory 2017). The process by which each of us matures and becomes a rational person is, according to psychological combinationism, essentially a social process conducted so successfully as to no longer appear to be social.

However, knowing what lessons to draw from the deep analogy between individuals and groups depends on a topic I have not even touched here: the **(p. 269)** value of *autonomy*, of making up “your own” mind and living according to “your own” will. It is a truism that to live together in society, people must accept some limitation to their autonomy in this sense. Questions about how far autonomy must or should be restricted, and how such restrictions are justified, are central to political philosophy, and combinationism only makes them deeper by pushing them down into the psyche itself: for many component subjects to live together as a well-integrated composite subject, they must forswear their own individual control over action, and even their own capacity to know their

identity. Conversely, if individual parts of the self have *independent* wills, letting them know themselves and act individually, the whole's autonomy is compromised: they are no longer "master of themselves," but act inconsistently based on caprice or circumstance, as one desire or another wins out and directs their actions. Indeed combinationism goes beyond expanding the range of cases where this kind of conflict over autonomy occurs: it unsettles the very idea of "your own mind," since a composite subject's mind is not *just* theirs but also "belongs" to their parts. To evaluate the values involved in these conflicts, and how they should be resolved, is a major undertaking in moral and political philosophy. I have not tried to do any of that ethical work, but merely to lay out the metaphysical and psychological framework within which we can best appreciate what is at stake.

Notes:

(1) Note that, if subjects are substrates, it is fine for distinct subjects to be experientially identical, because they can still be distinguished by their nonexperiential properties—human bodies are, for example, much heavier than brains. By contrast, psychological combinationism rules out complete sharing of experiences, on the basis that if two personas have all the same experiences there will be nothing to distinguish them.

(2) The two forms of combinationism do, however, have different implications for the identity of the Nation-Brain over time: if the citizens were issued a set of radical revisions to their button-pressing instructions, so as to remain in contact but generate very different behavior and cognition at the nation-scale, psychological functionalism would imply that the original Nation-Brain was now gone, replaced by a new subject grounded in the same substrate (like a new mind in an old brain). Functionalist combinationism, by contrast, would say that the Nation-Brain persists, though with a very different personality. On the other hand, if the radio transmissions were temporarily stopped so that citizens could be progressively replaced with something else—tiny robots, microchips, or even actual neurons—so that afterward the same organizational structure could be started up in a new (and probably smaller) system, functionalist combinationism would imply that the Nation-Brain was gone, even though its persona was now being run on a new substrate. By contrast, psychological combinationism implies that the Nation-Brain would have survived, leaving behind its component personas as it moved to a new substrate but remaining the same subject because it remains the same persona.

(3) The split-brain patient seems to exhibit the reverse condition: two streams of consciousness which are in large part causally independent but which display the same consistent personality, values, and goals. Cf. Schechter (2009), who argues that Tye (2003) is torn between one-subject and two-subjects accounts of the split brain for this reason.

(4) The problems discussed in this section were first suggested to me in conversation by Benj Hellie.

(5) Even if we hold back from accepting chapter 6's argument that social groups share conscious experiences with their members, we can make perfectly good sense of such groups producing and circulating representations which function like thoughts, beliefs, and perceptions, and these often involve either implicit or explicit self-representation.

(6) Similar problems arise with other implicit self-representations: the scouts should not report how, e.g., dangerous or vulnerable the enemy is by reference to the scouts' own capabilities, but by reference to the larger group's.

(7) There may or may not be additional, innate or learned "models" we are particularly prone to recognize and perceive as causal. For instance, we might automatically regard the transmission of rectilinear motion as "how things work," and be more sensitive to such patterns than to, say, S-shaped motion (cf. Cheng 1997). But these models cannot be the whole story, since we can identify and recognize causal relations even among unfamiliar or bizarre items.

(8) Again we might worry: doesn't perceiving things as responsive to "our will" require having already self-ascribed our will? But as with the first suggestion, this misunderstands the proposal. The patterning principle says that our spontaneous impressions of some X being "in us" or "outside us" are based on detecting harmony between it and some Ys, but does not require that the Ys be *themselves* the objects of any judgment or impression. We might be unconscious of the Ys, yet still have our conscious awareness of X affected by its harmony with them.

(9) Aren't there inner mental states which are stubbornly resistant to our efforts to ignore or induce them and which display stability over time? What about obsessive or haunting ideas and feelings, which refuse to go away whatever we do? But these are not, it seems to me, genuinely indifferent to our will. Rather, they recur and return constantly, in spite of being partially inhibited, or at least modulated, by our efforts to dispel them.

(10) Panpsychist combinationists will presumably think that the causal exclusion problem is solved by the Russellian thesis that physical laws are actually realized by the basic experiential properties; functionalist combinationists may prefer to solve the problem in the opposite way, by saying that all mental properties are functional properties realized by underlying physical brain properties. My concern in this book has not been to defend either option, but to show how a combinationist theory can secure sufficient ontological intimacy between composite subjects and component subjects for either of these views to work, chiefly by having experiences shared between wholes and parts.

(11) Singling out a single agent might be useful to ensure that each member has an incentive to refrain from executing the murder, even when they have already contributed considerably to the group's pursuit; this is similar to the obvious social interest in distinguishing murder from attempted murder (namely, to give people an incentive not to kill someone, even when they have already tried many times).

(12) I think this model is very intuitive when applied to social decision-making. A decision is usually said to have been made by some group of people, and the resultant actions attributed to them, only if they all "had their say" in its being made, if they all did or *could have* had some appropriate sort of input into the process. But it is also usually attributed to them only if the eventual outcome of that process is carried forward, or at least tolerated, by them all, as opposed to being fought by some, enforced by others, and ignored entirely by yet others.

(13) I discuss the possibility of this kind of enormously prolonged and intensified joint action in Roelofs (2017a), rejecting five different arguments that it would either no longer contain multiple agents, or that they would necessarily be impaired or defective when considered as rational agents.

(14) Schechter (2012a) offers an account of the split-brain phenomenon that fits this model: normally our two hemisphere subjects would have their will aligned (or, as she puts it, their agency unified) by direct transfers of information across the corpus callosum, but in the split-brain patient this alignment is achieved only by the pressures imposed by sharing a body: this co-embodiment secures unified agency for the whole patient, but on a different basis than the unified agency of a normal human whole.

(15) This does not mean they necessarily are all the agents of every single action produced by any of them; after all, some might have no causal role to play, or even receive any information about, certain actions produced by the others. How exactly to assign actions to interacting agents is a complex question: my goal is simply to show that there need not be a general principle of competition or exclusion among them.

(16) The claim that someone is not (determinately) the agent of actions produced by a desire of theirs which is in conflict with their broader psyche may seem to let too many people too easily "off the hook." But it is advanced based on the presumption that the two parts really are resolutely and determinedly opposing each other, i.e., that the person has tried with every ounce of effort to restrain their wayward desire. In many cases people put up a "token effort," or willingly allow an apparently wayward desire to control them because they prefer to act inconsistently, and be in denial about it, than to really give up the prospect of satisfaction. Here there is not really a conflict, fundamentally: the two parts of their mind have the same will, namely to allow there to be the appearance of

inner conflict. I will not attempt to fully describe or analyze the subtleties of self-deception in such cases, beyond pointing out that they do not threaten the validity of the preceding analysis of cases where the conflict is entirely genuine; rather, they presuppose it, since if there could never be genuine conflict there would be no sense in “staging” it.

(17) I remain agnostic on whether DID is necessarily a disorder when it does not cause dysfunction or distress; its status as disorder is denied by those who describe themselves as simply “being multiple” (e.g., Amorpha 2010; Astraea 2017).