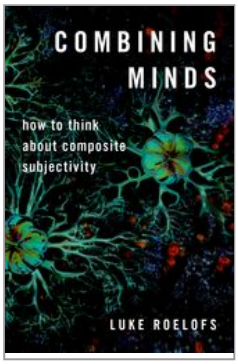

Combining Minds: How to Think about Composite Subjectivity
Luke Roelofs

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

What It Is Like for Two to Become One

Luke Roelofs

DOI:10.1093/oso/9780190859053.003.0008

Abstract and Keywords

This chapter considers in depth a particularly perplexing thought experiment that has puzzled theorists of personal identity, namely person-fusion. Philosophers have wondered: if two people were to go through a process of brain-interfacing that gradually changed them into a single, integrated, person combining the traits of both, should that be regarded as the destruction or the survival of the original subjects of experience? The chapter discusses this question, and tries to think through what this process might be like for the people involved: how they might experience a gradual shift from recognizably interpersonal ways of relating to the other person, to relations more like those that hold within a single person's consciousness.

Keywords: person-fusion, subject of experience, personal identity, brain interfaces, thought experiment

CHAPTER 7 OUTLINED psychological combinationism, on which subjects are constituted by sets of unified experiences and component subjects are constituted by clusters of experiences which are either *more* unified than the larger mass that contains them or unified relative to a different scale. Because there are so many ways that experiences can be called "unified," there are many valid ways to divide a subject into component subjects—many different patterns within the stream of their consciousness that are worth recognizing. Of particular relevance for humans are unity relations that involve the will: when component subjects' wills are "aligned," as those of cooperating humans are, the behavior that results is a genuine action by all of them, and by the whole they form. And when their wills are also "harmonized," covarying reliably and in

detail, they will each experience the other as extensions of themselves. When sharply opposed wills arise within a human mind and cannot be brought into harmony or alignment, the human being's unified sense of self and of agency begins to break down.

In this chapter I seek to illustrate and enrich psychological combinationism, along with the other theories of combination I have been developing, by applying them to an extended thought experiment involving the fusion of two persons **(p. 271)** into one—or, in combinationist terms, two persons becoming conscious parts of a composite person. Over the course of this thought experiment, subjects like us become subjects like our parts, and it is one of the advantages of combinationism that it can make sense of this as a gradual change of status rather than an abrupt transition.

8.1. Introducing Mind-Fusion

The best way to get a sense of what it is like to be a component subject in a unified mind is to imagine becoming one, and so most of this chapter is devoted to a thought experiment involving the fusion of two minds into one. My primary aim is to make vivid the abstract principles defended in previous chapters; my secondary aim is to argue that combinationism expands our options for understanding survival and identity in cases of fission and fusion.

8.1.1. Mars Needs Fused Humans

Suppose that some technologically advanced Martians decide, for reasons best known to themselves, that they would like to be able to combine many humans into a sort of "hive mind" creature, with a single consciousness controlling many bodies and drawing upon the cognitive resources of many brains. Suppose that a team of their best scientists gets funding for a series of experiments aimed at creating such a "multihuman."

They do this by implanting specially designed pairs of electrodes into the brains of two human participants, such that when a certain circuit in one brain fires, it activates an electrode, which sends a radio signal to the other, which immediately generates a surge of current in the other brain. This allows activity in one brain to produce or influence activity in the other brain, and hence allows the thoughts and experiences of one subject to produce or influence thoughts and experiences in the other. If the number and sophistication of these electrodes were increased enough, the causal connections between the two brains could eventually be as fast, as complex, and as reliable as the connections between the two hemispheres of either brain individually. At this point the only thing stopping us from saying there is a single, highly integrated nervous organ, with parts in two different skulls, is that the electrical signals are traveling via radio, not along an axon.[1]

(p.272) Of course this description simplifies what will be an incredibly fraught and complicated process. The Martians are ahead of us technologically by a few hundred years, but not by thousands, so while they can create and implant these electrodes, they have limited foresight about, and control of, what happens next. To prevent messy failures that only deplete their supply of humans, they run the procedure carefully and gradually, over a period of months or years, so that increases in the bandwidth of the implants are interspersed with periods of exploration, assimilation, and adjustment by the human participants, who can spend time working out what signals they can now send and receive and how to respond to these developments.

Moreover, let us suppose that the experimenters design the implants to mimic nerve cells as closely as possible. Obviously there are limits, since nerve cells do not emit or receive radio signals, but they may incorporate things like neurotransmitters, axons, ion channels, and so on. In particular, the manner in which the implants "multiply," spreading to connect more and more circuits across the two brains, could be made responsive to the way in which the participants use it. Just as neuronal connections become stronger or weaker based on the history of their activity, so can the links between implants, and between each implant and its brain. (As Churchland [1981, 88] says, "We simply trick the normal processes of dendritic arborization into growing their own myriad connections with the active microsurface of the transducer.") This way the experimenters need not constantly be performing repeat surgeries: the first operation puts in a biomechanical device which thereafter grows and expands into the brain.

8.1.2. Four Observable Outcomes

So far I have described the various high-tech interventions which our Martian scientists are making into the brains of their human subjects. The goal of these interventions is to create a single being, with a unified mind and coordinated, intelligent behavior across both bodies, which combines the personality, memories, and values of both humans. But their success in this endeavor is not guaranteed. While I will focus on the "good cases" in which they succeed, other results are possible, and the outcome is determined not simply by the setup and techniques used, but also by the way the humans handle the process. They have to undergo a transformative, potentially traumatic experience, and their temperament and attitudes to each other will make the difference at each stage between experiencing it as communion or as invasion.

We may distinguish four "ideal types" of outcome, allowing that actual outcomes may be intermediate between them or entirely unexpected. The intended outcome **(p.273)** can be called "merging"; this is when there is a unified mind (i.e., a single maximal persona) that controls both bodies and is recognizably continuous with both original personas. The combinationist can still say that this is actually one of *three* minds, since the originals can survive as parts of it. But

they display no more independent thought or sense of individuality than the conscious parts of a human brain.

The second-best outcome for the Martians is to have a single persona controlling both bodies but displaying recognizable continuity with only one of the original two personas. In such a case, we must surmise that the other persona has been suppressed, assimilated, or somehow subsumed into the resultant being without being at all manifest in that being's behavior. Call this outcome "domination." For instance, one of the two participants might be aggressive, defensive, and unwilling to allow another access to its thoughts, while the other is submissive and deferential and values acceptance over autonomy. The development of the experiment might then involve the former constantly seeking to interfere more and more with the latter's mental processes, while resisting any countervailing interference. By the end, one human has in effect "colonized" and "assimilated" the other's brain into itself, and thereby taken control of their body.

Third, we might end up with two recognizably separate personas, controlling different bodies or alternating in control of both bodies, despite the organic connectedness of their brains. This would be somewhat similar to a case of dissociative identity disorder: two psychologically distinct but internally integrated personalities control (simultaneously or sequentially) a single organic structure. These two would probably be recognizable as the original people, who had built up psychological barriers to replace the physiological ones that had previously separated them; a combinationist might still think there are, strictly, three minds here, but the composite mind they formed would at best be like that of a familiar social group like a club rather than an ordinary human mind. Call this outcome "dissociation."

This might arise if both participants were very concerned to maintain the privacy of their own mental processes but had little desire to explore or enter into those of the other. At each stage of the experiment, they might respond to the new way of influencing each other's minds by setting up, independently or cooperatively, policies and habits to minimize its effect.

Finally, the process might be too traumatic and too invasive for either subject to survive. They might both end up so deeply psychotic and fragmented that the resultant being is not recognizable as either, and perhaps not even as a single individual. This might mean that neither body's behavior was coherently interpretable at all, or even that at a certain point both bodies collapse into catatonia or epilepsy, **(p.274)** having somehow killed each other from within, and never wake up. One or both might even become traumatized to the point of violent paranoia, seeking out the other's physical body and stabbing or strangling it in order to silence the voices in their head. (If there is a composite

mind here, it is most similar to unstructured aggregates like “all the snakes in Ontario.”) Call this outcome “dissolution.”

So to get the philosophically most interesting “merging” result, we might need to run the experiment several times. But we may suppose that the Martian experimenters have as much concern for human life as human experimenters have for the life of rodents. My reason for sketching these four possible outcomes is that it will be illuminating to refer to them at various points, noting how the way the participants handle a particular aspect of the process might make one or the other outcome more likely. While I will focus on the responses which most conduce to the merging outcome, these will be best appreciated by contrasting them with those which conduce to domination, dissociation, or dissolution.

8.1.3. What Has Happened to the Original Participants?

As well as uncertainty about which of these four outcomes will be observed, there is room for uncertainty about what any of those outcomes would mean for the fate of the original participants. In this respect the procedure is a bit like the famous thought experiment discussed in chapter 2, where two people’s bodies are given the other’s memories, personality, beliefs, and so on; even once we have established every observable fact, and even once we know exactly what has happened in scientific terms, the question can still be asked whether two people have “switched bodies” or instead “switched minds.” Similar things apply to *Star Trek*–style teletransportation thought experiments, where a device scans, records, and disassembles the person who enters, then transmits the recorded information to a device elsewhere which constructs an exact replica of their body out of different particles. Even a full scientific understanding of how the device works is not enough by itself to answer the question whether the device is transporting one individual from place to place, or killing them and making a new but exactly similar person elsewhere. Such questions turn not on the scientific question of what would in fact happen, but on the question of what defines a person.

Particular puzzles are raised by the “merging” outcome of our imagined procedure because of the difference in number of people at the beginning and the end. If we start with two people but end with one, which (if either) of the original people has survived as the resultant person? (A similar question could be asked if we ran the reverse version of the procedure, “splitting” a single person into two people: Which of them, if either, does the original person survive as?) This is the puzzle of **(p.275)** fusion and fission, which has been discussed extensively but for which no fully satisfying solution seems to be available (see, e.g., Wiggins 1967, 50; Parfit 1971; Perry 1972; Lewis 1976; Nozick 2003). Other science-fictional thought experiments pose the exact same puzzle; the two most often discussed are the “teletransporter malfunction,” where a teletransporter, instead of just scanning, disassembling, and then re-creating a person, re-

creates them twice over, creating two perfect copies of a single person (Parfit 1984, 199ff.; cf. “Second Chances,” S6E24 of *Star Trek: The Next Generation*, and “Tuvix,” S2E24 of *Star Trek: Voyager*), and the “double-brain transplant,” where someone’s brain is safely bisected into two halves each sufficient to support consciousness, and each is then implanted into and given control over a different body (Wiggins 1967; Shoemaker 1984; Parfit 1971, 1984). And in chapter 7 I noted, in passing, that if the alters of a DID patient are distinct subjects, then the dissociation that produces them is a form of “fission,” and any therapeutic reintegration they may achieve is a form of “fusion.” Let us consider some of the options available for thinking about what happens to the original participants of our Martian experiment, to see both why the puzzle is puzzling and what combinationism might contribute.

On the one hand, it seems as though both of the original participants are gone: they no longer exist, having vanished into the new whole. This would clearly be the right thing to say if the observed result of the procedure was “dissolution.” But if the observed result is “merging,” with both bodies walking and talking and seemingly alive, governed by a single personality, it seems wrong to simply regard this as the destruction of both participants, as though they had simply died. They are learning new ways to communicate, to learn from each other, to work together, and this seems like a basically constructive, not destructive, sort of process. While the participants clearly do change, this change is gradual, organic, and often beneficial in terms of overall capacities. In particular, the resultant person still has the memories of both participants’ lives, and in that sense will seem to itself to have existed prior to the procedure.

But if the original participants survive, where are they? It might seem attractive to identify them each with the single person who controls both bodies at the end, but this rapidly leads to contradiction: if participant 1 is the same person as the resultant person, and so is participant 2, then participant 1 is the same person as participant 2. But clearly they are distinct people—otherwise we would not even have a case of fusion at all. We could avoid this contradiction by saying that participant 1 survives as the resultant person, but participant 2 does not (or vice versa), and this would seem the right thing to say in a “domination” case, where the resultant person shows psychological continuity with only one of the **(p. 276)** participants. But in a “merging” case, where each is equally continuous with both, there is no nonarbitrary way to say which participant has survived and which has not.

There is a way to say that both participants equally have survived: to say that they are jointly identical with the resultant person rather than individually identical with it. They are not *each* that resultant person; rather they are it, together. In short, each original person has become a part of a person. In a fission case, where a single person splits in two, the corresponding analysis

would be that two parts of them have survived, while they have become a pair of persons. Let us call this “the compositional approach.”

The compositional approach is not often explicitly defended (though see Moyer 2008). Parfit (1971, 7) rejects it for “greatly distort[ing] the concept of a person. . . . It is hard to think of two people as, together, being a third person.” The thesis of this book is that although this is true (it is indeed hard to think of two people as, together, being a third person), it is a feature of our everyday thinking that needs to be changed, not embraced. In short, the compositional approach “distorts” the concept of a person because it violates the anti-combination intuition.

8.1.4. Combinationism and the Fusion of Persons

The fact that the compositional approach to fusion and fission conflicts with anti-combination does not, of course, mean that combinationists need to accept it. Indeed, combinationists need not endorse a single approach for all sorts of fission and fusion. For instance, teletransporter malfunctions, which make two or more identical copies but do not in any literal sense “split” the original person, might need to be treated differently from things like double-brain transplants or amoeba-like physical “splitting.” Part of my aim in distinguishing the metaphysical and psychological conceptions of subjects, and remaining neutral between them, was to emphasize that combinationism does not by itself imply anything about what subjects are or what it means for one of them to survive over some period of time.

I do think, however, that it is worthwhile to explore how the compositional approach might be spelled out, to show the sort of options that combinationism makes available. So in the next two sections I will lay out a compositional way of understanding what has happened in my imagined Martian experiment when the observed result is “merging.” I will treat the process not strictly as “two becoming one” but as “three becoming three”: at the beginning there is a pair of subjects and the two subjects that compose it, and by the end the former has become an intelligent composite subject, and each of the latter has become a component subject **(p.277)** within it. It is like the process of welding two pieces of metal together: nothing has really been created or destroyed, but whereas before the two pieces stood out as units, and the whole that they formed did not, afterward their behavior is sufficiently integrated (e.g., any movement of one will mean a corresponding movement of the other) that the whole now stands out more than either part. The transition is thus not a discontinuous jump from two to one, but a gradual shift in which of the levels exhibits greater integration, and thus stands out as more salient.

8.2. Fusion from the Perspective of the Parts

In this section I discuss my thought experiment of fusion with a focus on the two human individuals involved. My aim is to show that what happens to them can

be understood as an extreme form of various familiar relational phenomena: they are each interacting with another person, who they at first perceive as "other" but come eventually to perceive as an extension of themselves.

8.2.1. Let's Talk, Brain to Brain

At first, the two participants will be related somewhat like two conversational partners, or two people with pagers. Each can, by thinking a certain way, produce a certain kind of experience in the other. Depending on where the implants are first put, this might be one seeing lights when the other thinks hard about math, or one hearing words whenever the other feels sad, or something else. Each participant may struggle, at first, to distinguish "normally occurring" experiences from those produced by the other. But suppose we give them a supportive environment, where they can talk normally with each other (so as to ask "Do you feel anything when I do *this*?"), and have the time and inclination to practice controlling and interpreting the implants. This will probably let them devise a mutually understood "language" and come to perceive implant-generated experiences as signals betokening another mind, just as we perceive words, hand gestures, or facial expressions.

In learning this language, the participants would be employing the correlation or lack thereof between their own volitions, the reported volitions of the other, the experiences they undergo, and the reported experiences of the other, thereby conforming to the patterning principle from chapter 7. Assuming that each was honest and open, they could distinguish experiences arising spontaneously, experiences voluntarily produced by them, and experiences produced voluntarily or unintentionally by the other. But they could do this only because these various **(p.278)** experiences were generated by "unsynchronized" systems, which varied independently rather than being volitionally harmonized at all times.

With time, they will also be able to reliably discern, from incoming signals, what effects they are producing in the other—paralleling the ability to, for instance, read in someone's face how they feel about one's utterances. Even before this they can guess, infer, or wonder about their effects on the other—including how the other is judging them based on "hearing their thoughts." Each may thus become "self-conscious" about whether the other finds the electronic relaying of their thoughts impressive, amusing, disgusting, etc. They may then make efforts to reduce or control the amount of information they send out, either by avoiding the thoughts which send out signals, by using feedback information to find ways of thinking those thoughts without being detected, or by learning which thoughts the other finds hardest to identify. Each might, that is, try to develop a "poker brain" in the same way that we can develop a "poker face."

Participants could also try to control the other's knowledge of their mind by asking the other to deliberately ignore the signals they receive, to direct their attention away from them. If they trust each other, they may simply request this,

much as we might request that someone look away while we change clothes or type a password. If these ways of politely ignoring the other, and the ways of establishing a “poker brain,” became habitual, they might eventually lead to the “dissociation” outcome, with the individuals surviving as two separate minds supported by one substrate. But if they do not trust each other, then in order to ensure their privacy, each participant must ascertain whether the other is attending to the signals they are receiving via the implant. But then each must, to protect their own privacy, invade the other’s privacy. This prepares the way for a conflict which might end in “domination,” if one wins conclusively, in “dissociation,” if both manage to repel the other, or in “dissolution,” if each psychologically cripples the other.

Fortunately, the participants may display emotional responses besides defensiveness. In the right circumstances, humans strongly desire both to know others and to be known by them. So if the right circumstances can be contrived, the participants may relish the connection which their implants give them and spend much of their time engaged in silent but energetic conversation. No doubt they will also want a degree of privacy, so expedients like the “poker brain” will be employed to some extent, but not so as to become automatic and inflexible. What will secure this happy result? Primarily it will be the temperaments of and relationships between the participants selected; perhaps for best results they should be a pair who have already established a strong and stable friendship or romantic partnership, who feel comfortable exposing their own mental lives and are enthusiastic about getting to see the other’s. Psychological health and stability will **(p.279)** also be important, to handle productively the tensions and arguments which will inevitably arise when two people, even people who love each other, are installed permanently inside each other’s skulls.

8.2.2. Is This Telepathy?

Suppose that the experimenters have, by luck, wise choice, or trial and error, selected two stable and mature human beings who are willing to be merged with each other but refuse to either conquer or be conquered. One interesting question is whether the communication they become capable of counts as “telepathy”—or rather, since that term could mean many things, is it a fundamentally different form of communication from the ways that people communicate normally?

If “telepathy” just meant communication not by any of the specific mechanisms that humans normally use—in particular, those that operate through the sense organs—then of course the participants are engaged in telepathy. But if telepathy means direct, *unmediated* awareness of another’s experiences, then they are not; the access is mediated by the electrical connections (radio waves or nerve signals) which link the two subjects. This means that the access is fallible, for something may interfere with the signals traveling along those connections. In this sense, what the two participants have is no more telepathic

than is an everyday spoken conversation: one subject's thoughts are encoded in some form of energy signal that is picked up and decoded by the other. When the energy signal is sound waves between mouth and ear, or light waves between face and eye, the communication is comparatively slow, unreliable, and low in information: when it is nerve impulses from one part of the brain to another, it is comparatively fast, reliable, and rich in information. Over the course of the Martian experiment, the radio communication between the participants goes from being more like the former to being more like the latter, but the difference is only one of degree.

Is the electrical communication perhaps somehow "less filtered"? It might be, especially at first. When we talk, we can exert a lot of control over the signal that gets to the other person, letting us craft it to be misleading or uninformative. But when the implants are first put in, there is none of that: each receives a signal that varies based simply on what the other is *actually* thinking or feeling, not on what they wish to claim they are thinking and feeling. But deliberate control, and with it insincerity and deception, can arise simply from each coming to understand how the connection works. Later on, as they become so used to each other that they slowly cease to think of themselves as distinct, this capacity for insincerity and deception might fade away again. So in a sense we might say that the participants have telepathic access at first (when they are probably unable to make sense of it) **(p.280)** but then lose it as they become more aware of the implants' workings, and may regain it once they become so attuned that filtering the signals they send to the other comes to seem perverse.

In another sense, telepathy might mean literally sharing experiences. Are there particular mental events which both subjects undergo? Of course there may be experiences that guide the behavior of both participants, but this differs only in degree from the way that my decision might guide the behavior of many other people, via my verbal instructions to them. I maintain that two discrete subjects cannot literally undergo a single experience: whatever is going on in one brain will be an experience of that subject, and not the other. And if an experience is realized in a dispersed manner, spread across both brains, then it belongs to neither individually; rather, that experience belongs to the composite of both, and each has some part of it, some component experience.

So I conclude that "telepathic" communication, in the sense of sharing particular experiences, does not (and probably cannot) occur between discrete subjects. They can communicate in a fast, reliable, and informative way, and they may in some cases do so with little room for insincerity or pretense. Yet they do not know each other's experiences *directly*, the way they know their own experiences, and they do not literally share experiences. This is a consequence of accepting (albeit in weakened form) the privacy of experience. If the two

participants are discrete, weak privacy precludes their sharing any single experience.

Because the procedure involves no telepathy, we could not use it for direct between-subjects quality comparisons. Suppose we connect one normally sighted participant and one blind participant. Or suppose we connected two participants who were, unbeknownst to us, spectrum-inverted relative to each other, one seeing green whenever the other would see red, and so on. What would happen?

Maybe one participant would report, with surprise, having a kind of experience whose quality they could not previously have imagined. But this might also happen when two experientially normal subjects are connected, if the signals each received stimulated their brain in previously unknown ways. Or maybe they would simply report experiencing familiar qualities under new circumstances or in new arrangements—because the incoming signal caused the same kind of brain activity as normal sensory stimuli. In either case, the type of quality which the receiving subject experiences is fixed by the signal's effects on their brain, and thus depends on the experiential capacities of that subject and their brain. They may find previously unknown capacities unlocked by the procedure, but this is still not the same as their directly sharing the experiences of the other person. In this respect, the participants are not fundamentally different from any two of us who talk.

(p.281) 8.2.3. Unifying the Participants' Experiences

What about conscious unity? Here panpsychist combinationism disagrees somewhat with other sorts of combinationism, but certain things can be agreed on. In most senses of "unity," the participants' experiences become gradually more and more unified. Before the procedure, the participants' experiences are no more representationally unified, or access-unified, or causally interdependent, or globally consistent, than any other two people's. By the end of the procedure, they are just as unified, in these respects, as the experiences of any ordinary person. This transition is not some sudden magical transformation; there is no crucial moment when unity "switches on" all at once. It is a gradual increase in the amount of information that is shared and integrated between the participants and the way that this establishes feedback that enriches the experiences of each. From each one's perspective, this means that their phenomenology of co-presentation—their awareness of the mental life "behind" the signals they receive from the other—becomes richer and more informative. Each participant comes to be constantly aware, not of the other's experiences as external events, but of how things seem to the other. That is, each participant comes to see the world through the other's eyes and have an awareness of this other perspective constantly in the background of their own perspective.

As the participants learn to tell what kind of experiences in the other mind produce what experiences in their own, they will come to perceive the latter experiences as the revealed aspect of—the expression of—a mental process that is not fully given but whose content they increasingly can discern. By extension, the absence of certain experiences in their own minds, and the particular combinations of experiences they do have, will also come to seem meaningful, like a fragment of a whole other stream of consciousness. Eventually, when each experience of either participant conveys something not just of *their* other experiences, but of the experiences of the other participant, we will be able to call the whole set unified because each experience will be felt as merely one fragment of a conscious whole that includes both minds.

Through this process, we will also find pairs of experiences in the two participants interacting to yield more complex experiences. For instance, when one of the participants perceives something, the signals received by the other may activate a memory of a similar thing perceived in the past, and the signals of this memory received by the first may then contextualize and color their perception of this new thing just as their own memories would. This growing disposition to think together will likely be accompanied by a growing difficulty in thinking **(p.282)** separately: when relevant thoughts from the other subject spring to both minds so quickly and readily, it will be hard for either not to be influenced by the other's ongoing thoughts.

The closest we can get to this kind of psychic intimacy is probably "internalization": when someone has exercised a formative influence on us, we can come to involuntarily and unreflectively see the world as they would see it, with this seeing becoming a constant background to our own seeing. What stops that from counting as conscious unity is that it is in virtue of a past interaction, not a present one, which is precisely what the procedure we are imagining changes. Each participant's constant background awareness of the other's perspective is a product of an ongoing interaction—though this need not mean that it snaps into existence simply because the implants are turned on. If the flood of new stimuli is unwelcome, it will prompt defensive measures or cause a traumatic breakdown in one or both psyches. But even if it is welcomed, there will still have to be a process of attunement and mutual learning, to help the two minds communicate more effectively rather than to help one re-create the other in their absence.

But what about phenomenal unity? Intuitively, causal interactions, joint access, and so on, are a very different kind of thing when accompanied by phenomenal unity and when not. When two representationally connected experiences are phenomenally unified, there is a composite experience with the richer representational content; when they are not, the complex content may be represented but not consciously. When two jointly accessible experiences are phenomenally unified, there is a composite experience which is access-

conscious; when they are not, information is accessible but not in virtue of any single experience being access-conscious. When do the participants' experiences go from phenomenally disunified to phenomenally unified?

By hypothesis, there is no discontinuous break in the gradual enriching of the participants' interactions, which makes it very awkward to insist that at some precise point the phenomenal transition must suddenly occur; indeed, I do not think there is any such precise point. But different forms of combinationism have different ways to avoid insisting on a precise point of transition. Combinationists who think that consciousness is a nonfundamental property, explained by some underlying physical or functional structure, can say the same about phenomenal unity: it is not a fundamental property, but rather explained by some underlying pattern of causal, functional, or informational relations. This allows them to say that phenomenal unity is vague: there is a midpoint at which it is neither determinately true nor determinately false that the participants' experiences are phenomenally unified, because their relations are neither definitely rich enough nor definitely not.

(p.283) Panpsychist combinationism cannot take this line: consciousness is a fundamental property, and so cannot be analyzed as meeting some ill-specified threshold on some underlying scale. A state is either conscious or not, and so a pair of experiences is either conscious jointly or not; thus phenomenal unity is not vague. Panpsychist combinationism instead says that there is no precise transition point, because there is not really any transition. Phenomenal unity is pervasive in the universe, and so the participants' experiences will be phenomenally unified even before the procedure begins. They were simply unable to introspectively register this unity because no set of introspective mechanisms could register the experiences in both heads. Eventually this changes, in ways that will be discussed more fully in the next section, and the whole becomes able to register the phenomenal unity among all its experiences. But this is a matter of its discovering something that was there all along.

8.2.4. The Growth of Shared Responsibility

As the procedure progresses, the two participants are likely to spend less time attending specifically to one another and more time attending jointly to external things. In the early "conversations," each focuses on how to convey things to the other and on how to interpret the other's own expressions. Their goals are to learn about the other, or to cause the other to believe certain things (true or false) about them. This contrasts with shared attention, where they are not attending to each other, but are rather aware of each other in the background of something they are both attending to, and which they are attending to partly because they know the other is attending to it.

To illuminate this shift, we might compare two people first meeting, engaged in “getting to know” one another, with two old friends jointly considering a shared problem. In the latter case, each may ask the other for their view on a particular part of the question, and will maintain a constant awareness of what the other knows, may not know, can do, refuses to do, and so on, but only in the same way that they maintain a constant awareness of their own capabilities and knowledge, without having to focus attention on them.

The growth of shared attention corresponds to one aspect of their growing conscious unity: their experiences tend more and more to transfer attention, so that if one focuses on something, the other’s attention will also be drawn to that thing (even if that thing is known to them only via signals from the other). This will contribute to a growing difficulty in assigning responsibility to one participant or the other. Each time one begins to focus on an action, the other becomes aware of this, and has a few related ideas, which the first immediately becomes aware of, **(p.284)** and so on. Eventually every action performed by either body will be the product of multiple rounds of feedback between both minds, making it all but impossible to isolate the exact contribution of each. In the terms of chapter 7, they are now consistently “aligning their wills,” acting on joint rather than just individual intentions, so that neither can escape responsibility for the actions of either.[2]

A similar difficulty in assigning responsibility may occur with rapid and habitual actions, because at some point it will likely be possible for one participant to initiate actions in the other’s body, perhaps acting via a momentary urge it can prompt in the other’s consciousness or perhaps bypassing the latter. Even if the second participant is able to inhibit the action just as they can inhibit their own, at this stage they will likely have become fairly comfortable with each other and not prone to automatically inhibiting every impulse they receive from the other. For instance, if one participant is graceful and perceptive and the other clumsy and oblivious, the first may start adjusting the second’s posture to avoid the trips and spills they were previously prone to. The second, seeing no reason to block out this helpful intervention, allows it to become habitual, so that every action performed with the second body is a blend of contributions from both minds.

But even while it becomes harder to assign responsibility to just one participant, it also becomes less practically important to do so. One reason we normally care which of two people did something is because we need to know what to expect from those people in the future. But if neither participant will act separately in the future, this prospective need becomes less urgent. It is also normally important to assign responsibility so as to allocate rewards and punishments, but with the two participants so closely linked in body and mind, and likely

strongly caring about and empathizing with each other, we cannot cause any pain, joy, or inconvenience to one without affecting both.

There will also be long-term effects on each participant's personality. Just as people often change to reflect and accommodate their families, the participants will be prone to absorb values, beliefs, and habits from each other—while any sharply opposing or incompatible traits are likely to be removed, either violently (in a case tending toward domination) or by persuasion and negotiation (in a case tending toward merging). Any persistence of sharp conflict increases the odds of dissociation or dissolution.

Drawing on my defense of the patterning principle in chapter 7, we might say that their background psychologies and background wills are more aligned with each other. If I am right that human beings perceive events as their own voluntary actions to the extent that those events harmonize with their will, then they may **(p.285)** start experiencing each other's actions as their own. For instance, when I decide to picture a yellow flamingo, and the image of a yellow flamingo arises, the close correspondence between what I intended and what happens makes me feel that I imagined the flamingo voluntarily. Now if one participant decides to picture a yellow flamingo, which is noticed by the other, who is generally better at visual imagination, their habitual response might be to form an image of a yellow flamingo and transmit it to the first participant. The first participant, experiencing an intention followed by a matching image, will likely feel as though they have voluntarily imagined that flamingo. The distinction between doing something oneself and doing it with the other's help becomes increasingly irrelevant as these options become indistinguishable in speed and reliability. Arguments over responsibility may still arise, when something goes wrong and each tries to blame the other. But these are unlikely to end with any clear answer; rather, the participants need to be willing to let go of the question. Only if they can think of their actions as "ours," rather than trying to carve out "mine" and "yours" in each case, will the procedure lead to successful merging.

8.2.5. Can the Participants Still Know Their Own Minds?

But couldn't each participant still consciously decide to "take charge" of their own mind and make a decision that is their own, not shared? Supposing the other is supportive and does not deliberately interfere, isn't individual responsibility still possible, by carefully "screening out" incoming signals?

There are three difficulties facing such efforts at individual responsibility. First, as noted above, it will grow harder and harder to discern which thoughts come from outside and which from inside. Second, many internally generated thoughts will now reflect the past influence of the other, and may even require consultation with the other in order to properly understand them (e.g., one participant may have been persuaded of a certain belief by the other, but

remember the conclusion better than they remember the arguments). As a result, evaluating the reasonableness or virtue of one participant's thoughts or actions will require evaluating those of the other that fed into it.

The third and most interesting difficulty is that to decide something "by oneself" requires distinguishing not only between what is one's own and what is another's, but also between what is one's own and what is a random passing whim or chance thought. Occasionally I might get the urge to slap an annoying person, but it does not follow that if I gave into that urge every time it arose I would be acting more autonomously, for it may be that a concern for civility and mutual respect is a far more important part of "who I really am" than this occasional urge. **(p.286)** This is essentially the same point discussed in chapter 7's section 7.3.1, that behavior caused by my mental states are not actions of mine, unless they go through some process which, like deliberate decision-making, is open to all my mental states and so can reach outcomes that reflect the balance of all of them.

This means that in order to find out what we really want, among a certain amount of random statistical "noise," we need to engage in "soul-searching." By this I mean a kind of calm taking stock of one's thoughts and feelings, a conscious effort to be open to all our desires and to discern which are stable and which are fleeting. This effort at "self-consultation" contrasts with the sort of blinkered focusing where we just pursue to completion what we have already embarked upon, pushing aside all new thoughts and feelings that arise, attending only to the implications of a particular line of thought or the means to a particular end. While this makes us more likely to successfully complete our task, it impedes efforts to establish our true wishes, and increases the risk of devoting ourselves to something we do not really want.

So to decide "for oneself" properly requires "soul-searching," which requires openness to one's "whole mind," a lowering of thresholds for admission to attention. But in our thought experiment, this kind of openness will actually invite thoughts stimulated (even if unintentionally) by the other. The attempt to block out or ignore thoughts coming from the other requires the opposite of soul-searching: vigilantly keeping watch over one's thoughts and driving away any that do not fit certain criteria. Thus the distinction "talking to the other person versus doing it myself" comes increasingly to line up with the distinction "soul-searching versus blinkered focusing." We can find parallels to this in real life: talking with someone else is often the best way for us to work out what we really want, as long as they are supportive and open-minded. In a sense "making up my own mind" is an illusion: our autonomy is the product of our relatedness. The participants in the thought experiment just take this to an extreme.

8.2.6. Surviving as Part of a Person

Many people would describe the merging outcome by saying that the two participants are no more; they no longer exist, having been “absorbed” or “dissolved” into the whole. Yet I would like to say that they survive as parts of a person, and combinationism is interesting in part for allowing us to say this sensibly. This does not mean entirely rejecting the idea that the participants are in some sense “gone”; indeed, this idea is correct in at least two superficial senses.

The first superficial sense is that the two participants have ceased to be salient things, or things which it is useful to think in terms of. It is no longer sensible for **(p.287)** someone trying to understand the situation to organize their thoughts around “participant 1” and “participant 2.” The second superficial sense explains this fact; each participant has lost a significant degree of “independence,” in that their body’s actions are no longer primarily controlled *only* by their own mental processes, and their mental processes are no longer primarily controlled *only* by their own mental processes a moment before. But despite these facts about the two participants, a defender of the compositional approach will maintain that they still, strictly speaking, exist. The success of fusion is not marked by a change in which things exist, but by a shift in the salient divisions, whereby the distinction between the two individuals becomes increasingly irrelevant. Each one’s story is not a story of something being destroyed, but of something growing and forming new connections to an external thing—indeed, an external thing which, if the experiment succeeded, probably managed to elicit intense feelings of friendship, love, and acceptance. It would be perverse to think of this as a form of self-destruction.

Of course, what the participants’ survival consists in will depend on what subjects are, and I have tried to make combinationism compatible with both the metaphysical and the psychological conceptions, as well as with mixed views which recognize both conceptions as identifying important senses of “subject.” On the metaphysical conception, the original people are the two substrates—crucially, the two brains. Since these two brains clearly still exist, the contentious claim is that they are still subjects. On the psychological conception, the original people are two personas, and the contentious claim is that they persist as two component personas in the overarching persona of the two-bodied being that results. If we had observed the outcome I earlier labeled “dissociation,” it will be very clear that this is the case: the two component personas will verbally assure us of their own distinctness and existence. But in the “merging” outcome, things are less obvious; as noted above, the behavior and thought of the two-bodied being shows no salient or obvious division into two personalities but is seamlessly integrated.

But section 7.1 of chapter 7 argued that personas can exist even when they do not stand out sharply from their background, and the mental life of the pair is still likely to show two subclusters, slightly more unified internally than they are with each other.[3] Moreover, experiences arising within one particular brain are likely to show greater causal interdependence with one another than they do with experiences arising in the other brain, because of the simple effect of physical proximity, shared dependence on particular mechanisms, and being the "first port-of-call" for the same streams of incoming sensory information. So even if **(p.288)** there is no preservation of any discernible boundary between personality quirks, memories, habits of speech, or dispositions inherited from the one participant and those inherited from the other, there is still a sense in which the original personas can be delimited from each other as parts within the whole.

The basic reason to accept the continued existence of the original participants as component subjects is that the type of thing happening in each brain at the end is basically the same as what was happening at the beginning. Neurons fire, stimuli produce responses, information is processed and filtered and integrated to enable a coherent and meaningful worldview. These activities have not been disrupted or stopped, just connected with other such activities so that they work together. Why think that these activities no longer support a conscious perspective of their own?

One might say that they no longer support a perspective of their own because they contribute to supporting a larger conscious perspective. But this follows only given the rejection of combinationism, which claims precisely that supporting a perspective of their own can be a way of contributing to a larger perspective. Or one might say that their psychological integrity has been destroyed, and they no longer display the same psychology as before. But any theory of personal identity must account for the way that our psychologies change over time, and one of my major aims in this section has been to show that what happens to these participants is just an extreme version of things that happen to all of us when we become intensely related to someone: shared attention, long-term shifts in personality, or finding that we best understand our own desires by discussing them with another.

8.3. Fusion from the Perspective of the Whole

In this section I shift to consider the whole composed of the two human participants. Its perspective is, at first, much less familiar to us and harder to make sense of. Nevertheless, I argue, we have conceptual tools available that can give us some idea of what it is like to be a pair of people midway through this process, and perhaps even what it is like, if anything, to be a pair of people prior to this process—i.e., what it is like to be any of the many pairs of people which actually exist.

8.3.1. Being a Pair of People

Different combinationists will take different views on whether pairs of people are composite subjects, but all should agree that they are not *intelligent* subjects. The many experiences going on in a pair do not integrate enough information overall **(p.289)** to generate structured consciousness; they do not work together to produce intelligent functioning at the pair-level. This is why it seems bizarre to attribute mental states to a pair, once we are clear that this ascription is not to be merely distributive; obviously we can say "There's a happy pair," meaning that each member individually is happy, but the pair cannot really be happy itself. Because intelligent functioning is division-relative, intelligent functioning in each of a thing's parts does not guarantee intelligent functioning in the whole. But it is also not incompatible with it, and human sociality shows a way to connect the two. As discussed in chapter 6, joint intentionality in its various forms effectively allows the many mental states of some group of people to jointly play a certain functional role for the group as a whole. And our two participants are ideally placed to intend, think, and act jointly: they have unfettered access to the other's mental states, time to build up cooperative habits, and (hopefully) a deep bond of trust. Given this, we can expect them, more and more frequently, to each think and behave in ways that are conditional upon the other's "doing their part." When they work together like this, it begins to make more sense to describe the pair in properly psychological ways—as "wanting X," for example, when both participants want X, know the other wants X, and are prepared to work together to get X.

As joint intentionality becomes easier and more automatic, it will approach the ease and automaticity of individual action. At a certain point, the pair will start to qualify as functioning intelligently, and its intelligent functioning will be a natural outgrowth of the intelligent functioning in its parts.[4] In tandem, the increasing integration of information that enables this intelligent functioning will be connecting the participants' experiences in more and more specific ways, until the whole starts to qualify as having structured consciousness. At this point, it will be an intelligent subject—or rather, since "intelligent functioning" and "structured consciousness" are vague, it will first reach a point where it is indeterminate whether it qualifies as an intelligent subject (rather like, perhaps, an insect or a snail, though differing considerably in containing two definitely intelligent subjects as parts), and then later reach a point where it is determinately an intelligent subject.

On pure functionalist versions of combinationism, being an intelligent subject is the only way to be a subject, so the pair becoming an intelligent subject is also its becoming conscious. But this becoming conscious does not involve the appearance **(p.290)** of any new experiences: the experiences themselves were going on in the pair all along, but only when the whole functioned as a subject did it qualify as "having" them. By contrast, panpsychist combinationism allows for subjects which are not intelligent—subjects like fundamental particles, which

have experiences but without any cognitive capacity to access, think about, or act on those experiences. And panpsychist combinationism says that in this "stripped-down" sense of subjecthood, all wholes inherit experiences from their parts. On such a view, the pair of people was conscious all along, but without any of the functional and behavioral complexity we normally associate with consciousness. The pair had the same experiences as its parts, but they, and not it, also have their overall behavior intelligently guided by those experiences. It was conscious only in the minimal sense that elementary particles, according to panpsychism, are conscious. Moreover, according to panpsychist combinationism, its experiences were phenomenally unified all along, but prior to the Martian experiment the two clusters associated with the two brains integrated so little information between them that the composite experience they formed was a mere blur, a blend of all their different experiences in which none could be distinguished from the others. As the two brains integrate more and more information, elements of their two conscious fields become phenomenally bound, and structured consciousness slowly coalesces from what was previously a homogeneous blend.

8.3.2. How the Pair Knows Itself

Even if the pair is conscious from the beginning, it cannot at first introspect on its consciousness as a whole. In the early stages of the procedure each person has good access to their own experiences, but very limited access to the other's. Each will thus form introspective impressions of only part of the composite's phenomenal field, or at least, introspective impressions in which the rest of the field features hazily and peripherally. This could still be called a sort of "introspection by the composite": the composite is introspecting on half its experiences at a time, using a different brain each time. But this falls short of what we might think of as "proper" introspection and attention, which involve the whole phenomenal field. These, I will argue, become possible for the composite only when it can act simultaneously with both brains and connect its two acts properly.

For proper introspection, the composite needs to not only have each brain introspect but have each inquire with the other about the other's part of the field, and incorporate what they are told into their own view. It may never get a complete survey of the phenomenal field in a single brain, but this does not mean it never gets a complete survey: the complete survey is distributed across the two **(p.291)** brains. This allows the composite to think about its own thoughts, but it does so by means of each participant separately identifying and then putting together "what *I* am experiencing" and "what *they* are experiencing." These two impressions will be distinguished as introspective and testimonial, respectively, and so the composite can think of itself as such only on the basis of first thinking of each part as a distinct part.

But as the procedure progresses, the consequent “we-thoughts” will follow the antecedent “I-thoughts” more and more automatically, as any attempt at introspective stocktaking by one participant automatically prompts the other not only to also take stock, but to listen to its partner and share its own results. With increasing automaticity, it will become less salient to each that it is hearing things from someone else. By contrast, the impressions in each brain of what “we” are experiencing will grow more salient, not to mention easier to focus on, for segregating out only the experiences from one’s own brain will come to take more and more effort. Eventually, for a participant to introspectively review all and only what they individually are experiencing, without attending also to what the other is experiencing, will come to seem both pointless and nearly impossible, compared to reviewing everything in both minds. The participants come to be conscious of themselves only as a whole, not as parts. And thereby the whole becomes conscious of itself as such.

8.3.3. New Forms of Phenomenal Blending

While in one sense the composite has gained new capacities throughout the procedure, there is also a respect in which it has steadily lost capacities. In particular, it has lost various capacities to do one thing *without* a certain other thing happening. The implants ensure that certain mental or bodily actions taken with one body will have automatic consequences in the other body—such as the other participant knowing what was done, or feeling some emotion, or supplying some useful or disruptive feedback. It is often worthwhile and overwhelmingly tempting to prioritize speed and efficiency over carefully scrutinizing every step in a mental process, and so when there is no strong reason to keep things separate, a useful pattern of action may become so habitual that its components cannot occur separately.

This loss of capacities is a form of *confusion*, as defined in chapter 4. The composite mind becomes unable to perform one mental act without simultaneously performing another, just as each participant’s mind may be unable to attend to the microexperiences of its microscopic parts without simultaneously attending to many others. At first this confusion will be weak and shallow: shallow because the effect in the second mind might still potentially be inhibited or avoided, and weak **(p.292)** because even once both mental actions have occurred, the separateness of the two brains’ attentional systems will allow for distinct attention. But over time, as the links become stronger and faster, and as the two brains’ capacities for attention become more and more coupled together, there may be instances where it becomes in practice impossible even to attend separately. Obviously this will not happen for all mental events; there are benefits to preserving some separation of function, if only so as to know at each moment which body’s eyes are generating which visual experiences. But pairs of events which there is no reason to keep distinguishing may become strongly, robustly, and symmetrically—i.e., radically—confused.

Since confusion is subject-relative, this does not guarantee that the parts suffer the same confusion. They may suffer confusion between a given thought or experience of theirs and the *feedback* that it automatically prompts from the other brain, but this is still confusion among events in their own brain, contrasting with confusion among events in the two brains.

This raises the interesting possibility of the whole experiencing phenomenal qualities which its parts cannot, which are blended out of, and thus nothing over and above, the qualities experienced by its parts. To see this, imagine that one subject has very different color experiences than the other, outwardly responding the same to stimuli but experiencing different qualities—seeing gred, grue, and grurple where the other sees red, blue, and purple. For these to blend they must be both phenomenally unified and radically confused; this is unlikely to happen with perception (due to the separateness of eyes), so we should focus on imagination. Consider the whole visually imagining the color blurple, a blend of gred and blue, by imagining gred in one brain, and blue in the other, without any ability to separate the two. What is this like for the parts, e.g., the subject who can see and imagine blue but not gred? Since their blue-experience is representationally unified with the other's gred-experience, it co-presents that other experience in some fairly informative way. Yet since the link is not telepathic, they do not experience and cannot even imagine gred, and by extension do not experience and cannot imagine blurple. Yet at the same time, this joint imagination is entirely satisfactory within the psychological economy of the merged minds, generating no sense of dissatisfaction or frustration in either part.

Thus the subject imagining blue has an experience which co-presents gred informatively, without actually instantiating the character of gred, and yet generates no further curiosity or sense of lacking access to the quality gred. This is a hard description to make sense of, but not impossible. I think the best way to make sense of how the co-presentation can be highly informative without actually conveying the character of gred is to think of it as a sort of acquaintance that enables recognition but not recall; the subject will recognize next time the **(p.293)** other subject is imagining gred, but cannot themselves visualize or otherwise capture the quality—a little bit like when we say "I don't know how to describe it, but I would recognize it if I saw it again." And the best way to make sense of its producing no sense of frustration or curiosity is to appeal to the patterning principle: the subject imagining blue thinks they can imagine gred, and indeed blurple, because whenever they try to (by resolving to imagine "that quality"), this prompts the other to imagine gred, thereby composing blurple, and all the feedback that any subject (composite or component) gets is "Success: quality visualized." In fact much of that feedback is coming to each component subject from the other, but the degree of volitional harmony they have built up makes this fact impossible to detect from the inside.

8.4. Conclusions

When I think about the procedure our two imaginary humans have undergone, I find combinationism more attractive. If minds cannot be parts of minds, then there must be some moment when two become one, and I find it hard to accept any such abrupt transition in a gradual process. If minds can be parts of minds, but component minds do not explain the mentality of their composite, then when we have explained the behavior of the composite in terms of the two original minds closely and automatically cooperating (as I think we can), it would be superfluous to posit a further distinct mind belonging to the whole. Yet if the whole gives as much indication of being an intelligent conscious subject as any of us do, then it seems arbitrary to deny it that status. I find myself forced to look for some account on which the genuine consciousness of the parts and the genuine consciousness of the whole are not only compatible but are two sides of the same coin. Over the course of this work I have attempted to formulate three versions of such an account and to articulate both why they are attractive and why they might be rejected.

Notes:

(1) Churchland (1981, 87–88) briefly discusses this sort of procedure in a futuristic human society; Rovane (1998, 141) discusses something similar but conceived of as a voluntary undertaking “something like a marriage arrangement” (cf. Parfit 1971, 19; Roelofs 2017a).

(2) In Rovane’s (1998, 141) terms, this is part of their establishing “overall rational unity.”

(3) For an example of an integrated persona which has recognizable component personas from earlier individuals who fused, see “Tuvix” in S2E24 of *Star Trek: Voyager*.

(4) Arguably, full-strength ascription of psychological properties to the pair requires it to be a “group agent” in the robust sense described by List and Pettit (2011), which requires not just a succession of temporary joint actions but also a stable and consistent set of collective attitudes. This seems likely to be what happens in our imagined procedure, as each participant becomes more and more able to persuade and be persuaded by the other, and more and more willing to adhere to past joint decisions.