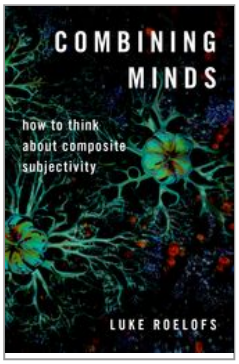

Combining Minds: How to Think about Composite Subjectivity
Luke Roelofs

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

A Universe of Composite Subjectivity

Luke Roelofs

DOI:10.1093/oso/9780190859053.003.0001

Abstract and Keywords

This chapter introduces the topic of the book—composite subjectivity—and explains why it matters. This involves clarifying how the key term “combination” is used and how key ideas like “composition” and “consciousness” are understood, as well as reviewing the various reasons why philosophers have tended to deny or neglect the possibility of composite subjectivity, and the implications they have drawn from doing so. The chapter explains the significance of mental combination for panpsychism’s combination problem, for collective consciousness, and for a variety of other issues in the philosophy of mind, and sketches out the book’s plan of attack.

Keywords: composition, consciousness, panpsychism, combination problem, collective intentionality, metaphysics

NOTHING IN NATURE is an absolute unit. Human bodies, and everything around them, are built up out of smaller things, themselves built up out of smaller things, and they are parts of larger wholes, which are themselves parts of larger wholes. Moreover, they overlap with other things marked out by other ways of drawing boundaries: the world can be carved up in many different ways, but it is the same world however we carve it. This compositional structure is central to our ability to make sense of the physical world: both scientifically and in everyday life, we understand things by breaking them down into simpler components and putting them together into new and more complex wholes.

Are our minds likewise wholes made of mental parts, themselves made of smaller and smaller minds, which can be divided in many alternative ways or

recombined into more complex mental wholes? If they are, that would be a deep and interesting continuity between mind and matter; if they are not, that would be a deep and interesting discontinuity. But there is something puzzling about the thought of minds composed of, contained in, and overlapping with other minds: How should I relate to these other minds that I contain, or are contained within, or overlap with? Are they "self" or "other"? Do they know that I exist? Does it really **(p.4)** make sense for my point of view to be nothing over and above theirs? In this book I defend and explain this idea: that each of us has a composite subjectivity, formed of many individually conscious parts.

To bring out what is at issue, consider an everyday physical object, like a rock. Theories in physics never mention rocks, but this is no shortcoming: rocks fit neatly into the world described by physics because they are made up of parts (electrons, quarks, etc.) that physics does mention, and these parts combine with each other in intelligible ways to yield a rock. In particular, note three things. First, many key properties of the rock are "division-invariant" (a "division" is simply a certain way of organizing things into larger or smaller units): its mass or volume, however, remains constant whether you treat it as one thing, or as two halves, or as trillions of atoms. Carve it up however you like, when you add up the features of the parts, you will get the same result.[1] Of course, many of the rock's other properties are not like that: the rock's texture or solidity or symmetry is not automatically shared by its halves, let alone its component atoms. But—and this is the second key point—these properties are typically "intelligibly division-relative": when we analyze them we can see that their nature explains why the same set of division-invariant properties would support them on one way of carving but not another. Suppose the rock is striped when considered as a whole rock, but not when considered as a set of atoms (atoms are not striped, after all). But it is not as though stripedness mysteriously pops into existence at the rock-scale: analyzing what "being striped" means reveals that having thin adjacent regions of different colors is all it takes to be striped, and it is obvious why the whole can satisfy that description even though slices of it do not. Third, the relations that hold within the rock (its structure of "internal relations") can equally well be viewed as external relations between distinct objects, namely its parts. If its different-colored slices are chemically bonded together, we can equally well describe this as several objects, each of which is bonded to another, external object, or as a single object, which is internally tightly bound.

The idea of combining minds is challenging because these three things seem to be absent when we start to think about the mental world, and in particular about consciousness, about what it feels like to be me at a given moment. I may be a part of a club, or be made of many neurons, but consciousness seems, at first glance, to be present only at the human-being level, not at the level of the club (even though it has conscious parts) or the level of the neurons (even though they form **(p.5)** a conscious whole).[2] So consciousness does not seem to be

division-invariant, like mass, but it is also far from clear whether consciousness is intelligibly division-relative, like being striped. When I consider a club which has conscious people as parts, I do not know how to redescribe the consciousness of those people as a property of the club, the way I can redescribe the rock's parts being differently colored as the rock being striped. Likewise it is hard to see how my consciousness could be redescribed as some fact about my many neurons. And, most strikingly, the relations that hold within my mind seem very different from those that hold between my mind and the minds of others. My desires can motivate my decisions, my perceptions can inform my beliefs, my thoughts can conjure up imaginings, all apparently *directly*. By contrast, any interaction between me and other people is strikingly *indirect*: my desires and perceptions and thoughts seem radically cut off from theirs and can affect them only by first producing some outward movement of my body that they can perceive. In these ways the conscious mind seems like an anomaly, an absolute unit in a world of endless division and composition.

It is because composite subjectivity seems so challenging that I want to defend it. Historically, many philosophers have gone the opposite way and deemed it impossible. In fact, this supposed impossibility has sometimes been used to argue that consciousness requires an indivisible, immaterial soul, separated from the material world and from all other souls by unambiguous boundaries. At other times the challenge of composite subjectivity influences philosophical debates more subtly, by obscuring possibilities which would rely on it. But I think that denying composite subjectivity is a mistake, an illusion of indivisibility that misattributes the imperfect unity that our brains manage to achieve to a fundamental inner unity at the heart of the mind. It is also human-centric: we are too used to thinking about the minds of things with centralized brains, highly integrated internally and separated from all other brains by thick layers of bone. But there are creatures on earth whose nervous systems are not packaged into such neat units, and for all we know even more such creatures elsewhere. Developing brain technologies offer the prospect that we might be able to forge connections between human brains just as strong as the connections within a human brain, or establish divisions within a human brain just as sharp as those between different brains. If we want **(p.6)** to understand all of these possibilities, and free our imaginations from the narrow bounds of present-day-human-like minds, we need to think about composite subjectivity.

I analyze the idea that compositive subjectivity is impossible in terms of a principle I call "anti-combination," which in turn employs the notion of a "mere combination" and is opposed to a doctrine I will call "combinationism," defined thus:

> **Anti-combination:** The experiential properties of a conscious subject cannot be mere combinations of the experiential properties of other subjects which compose it.

> **Combination$_{df}$**: One property-token P is a "combination" of other property-tokens p_1, p_2, p_3. . . , if and only if (i) P is fully grounded in the ps, and the real relations[3] obtaining among them, and (ii) P is fully explained by the ps, and the real relations obtaining among them.[4]
>
> **Combinationism:** The experiential properties of a conscious subject are sometimes mere combinations of the experiential properties of other subjects which compose it.

A more intuitive way to put the definition of "combination" is to say that a feature of me is a combination of features of my parts if those other features, and the way they are connected, both suffice to ground my having the feature in question and could be appealed to in an explanation of why I have it. Obviously we still need to unpack terms like "grounded," "explanation," "parts," and "experiential properties"; in section 1.3 I do so, but for now I will leave the idea rough. Anti-combination claims that when something is conscious in a certain way, this is always a further fact "over and above" the fact that its parts are conscious in certain ways and related in certain ways: combinationism denies this. Combinationism allows for composite subjectivity: experiential properties of x that are mere combinations of the experiential properties of the parts of x. Combinationism also implies the possibility of partial overlap of consciousness: three subjects such that **(p.7)** A is a part of both B and C, and its experiential properties contribute to grounding and explaining both of theirs. The opposite of a mere combination (a property of a whole which is neither grounded in nor explained by any properties of its parts or their interrelations), I will call an "emergent" property.[5] It is often thought that there are no genuinely emergent properties in nature, and there seem to be none in the physical realm.

I think anti-combination is deeply intuitive to many people and has a powerful grip on our thinking about consciousness. I also think it is false; each of us in fact has a composite subjectivity, built up out of the consciousness of the parts of our brains. But it is hard to make sense of combinationism, of the idea that "what it is like to be me [is just] what it is like to be each of those [parts] (somehow experienced all at the same time)" (Goff 2006, 59). How can my perspective be nothing over and above what it is like for many other beings to stand in some relation? What does it mean to talk about "what it is like to be all of them" and to identify that with the apparently unitary consciousness I enjoy?

My aim in this book is to make it plausible that anti-combination is false; equivalently, it is to defend combinationism and the possibility of composite subjectivity. This will require theorizing mental combination so that it shares the three features, mentioned above, that make physical combination so satisfying: mental properties should be either division-invariant or intelligibly division-relative, and mental relations should be simultaneously within-subject and between-subjects. We should be able to see how, for each component subject in a composite subject, the other parts are external things and its relations to them

are relations of self to other, even though for the whole, none of the parts is external, and their relations are just the internal structure of its own mind.

This chapter sets the stage by defining and contextualizing this project: Why does composite subjectivity matter? Who has endorsed anti-combination? And what do these various terms (“part,” “consciousness,” etc.) mean, exactly?

1.1. The Anti-Combination Intuition

The reason for this book’s existence is the breadth and strength of the anti-combination intuition. The idea that consciousness, or mental properties more generally, cannot combine is often asserted as obvious, as a premise that can be assumed without argument (e.g., Van Inwagen 1990, 118; Lowe 1996, 39; Merricks **(p.8)** 2001, 111; cf. Bechtel 1994, 16). Other writers assume it implicitly in the way they frame or defend their views. And there is certainly something compelling in the idea that a conscious point of view has a special, indivisible wholeness. This widespread and powerful conviction makes composite subjectivity seem impossible.

In this section I review some of the ways that I believe the anti-combination intuition has influenced philosophical thought on a range of topics; in the next section I examine in more detail a topic which has recently occasioned especially focused attention to questions of mental combination, namely panpsychism and the “combination problem” thought to face it.

1.1.1. The “Achilles” Argument and Anti-Nesting Principles

First, anti-combination has a long and prominent history in the form of what is sometimes called the “Achilles” argument, so christened by Kant (1997, A351) for being the strongest argument of its kind. The argument maintains that because “representations that are divided among different beings . . . never constitute a whole thought” (A353), conscious minds must be simple and without parts, and therefore cannot be material. Versions of this Achilles argument appear in authors stretching from Plotinus (1956, 255–258, 342–356) to Brentano (1987, 290–301).[6] And for these authors, since physical things are always divisible, the indivisibility of conscious minds proves that they lie outside the physical world.

In recent philosophy of mind, most explicit discussion of anti-combination has dealt with the combination problem for panpsychism (treated in section 1.2), but that does not mean that other theories of mind are not influenced by it. For one thing, advocates of both machine-state functionalism (Putnam 2003, 215–216) and integrated information theory (Tononi 2012, 59, 67–68) have stipulated “anti-nesting” principles, by which no system can qualify as conscious if it is contained within, or contains as parts, other conscious systems: consciousness can be in the part or in the whole but not both. (This claim is of course not equivalent to the claim that conscious minds must be simple, i.e., without any

parts at all, but both show the same sense of a special difficulty in mental combination.)

1.1.2. Consciousness in Social Groups

The anti-combination intuition exerts a strong but sometimes unrecognized influence when philosophers consider the status of large groups like nations or crowds. Although many philosophers vigorously debate the possibility and nature of **(p.9)** collective *intentionality*—collective beliefs, collective goals, collective intentional actions, etc. (e.g., Velleman 1997; Gilbert 2000; Bratman 1997; List and Pettit 2011; Tollefsen 2014)—the proponents of this kind of collective mentality will generally resist the idea of collective *consciousness*. While a group might plan, or investigate, or deploy strategies to achieve goals, it is assumed that there could not be something it is like to *be* a group: whatever we may do together, surely there is nothing it is like to be us. Even those who go furthest in defending genuinely collective mental states stop short of collective consciousness: Gilbert (2002) and Huebner (2011) both argue in support of collective emotions, but do so by trying to break the link between emotion and consciousness, arguing that genuine emotions may be devoid of phenomenology. Thus they seek to prevent "the implausibility of collective consciousness . . . impugn[ing] the possibility of collective emotions" (Huebner 2011, 102).

The intuition that groups are not conscious is so strong, in fact, that it has been used to argue against particular theories on behalf of others. For instance, Block (1992) offers thought experiments in which we are supposed to find consciousness in certain beings composed of other conscious beings implausible. Block regards consciousness in such entities as "an absurdity" (79), and therefore rejects theories of consciousness that imply it.[7]

Historically, proponents of the Achilles argument also often support their claim that minds cannot have parts by drawing an analogy to groups of people, as in this quotation from Brentano (1987, 293):

> If, when we see and hear, the seeing were the property of one thing and the hearing the property of another, then how could there be a comparison between colours and sounds? (It would be just as impossible as it is for two people, one of whom sees the colour and the other of whom hears the sound.) (cf. Plotinus 1956, 346; De Courcillon and Timoleon 1684; for an example directly targeting panpsychism, see James 1890, 160)

I believe that it is primarily anti-combination that explains this intuitive resistance to the idea of collective consciousness, and in particular the discrepancy in attitudes toward collective intentionality and collective consciousness. People generally want to avoid irreducible, emergent group minds which float above individual minds; such beings seem mysterious, unparsimonious, and even ethically **(p.10)** threatening (Searle 1990, 404; List

and Pettit 2011, 9). We can avoid emergent minds while allowing for collective intentionality, as long as that intentionality can be explained through and grounded in that of individual members. But if consciousness in a whole cannot be explained by or grounded in consciousness in its parts, then this anti-emergence attitude rules out collective consciousness. Combinationism dissolves this conflict: social groups can be nothing over and above their members, while still being literally conscious. That is not to say that combinationism entails that collective consciousness is ubiquitous or even actual; it merely shows that the possibility is not absurd.

1.1.3. Human Beings with Impaired Unity, and Other Unusual Cases

The anti-combination intuition also constrains the way we think about cases where the unity of a human person is in some way impaired.[8] Consider two complementary cases: the split-brain phenomenon and dissociative identity disorder. Both cases seem somewhat intermediate between what we would normally count as one mind and what we would normally count as many minds, but accepting anti-combination forces us to regard those as the only available options. In the split-brain syndrome, the nervous fibers connecting the two cerebral hemispheres are severed, and although the patient appears normal in everyday situations, very strange results can be produced when stimuli are segregated so as to be processed by only one hemisphere. Simplifying greatly, it seems that, when responding with the motor organs controlled by that hemisphere, the patient shows full awareness of the stimulus, but with the organs controlled by the other hemisphere, they claim not to have seen it. It seems like there are “two people,” one for each hemisphere, each with half the body’s full set of sensory and motor organs, and each responding rationally to its own sensations but unaware of the other’s. Dissociative identity disorder also involves what seem like “partial persons,” but in a completely different way: the “alternate personalities,” or “alters,” are not tied to any particular sensory or motor organs, but instead take sequential or simultaneous control of the whole body. What differentiates them is the memories they can report, the personalities they display, and what they claim about themselves—or, to put it another way, what the patient reports, displays, and claims when exhibiting a particular alter.

(p.11) In both of these cases, there is compelling reason to think both in terms of a plurality of individual people and in terms of a single individual showing unusually dissociated behavior. It is very natural to think that the right analysis is somewhere in between—that there are two of *something* person-like, and also one overall unit that they together form. But as long as we hold onto the anti-combination intuition, it will be almost impossible to conceive of any such intermediate option. If there are two people, then there is no single person who subsumes both; if there is a single person, then the constituents or components of that person cannot really be people. This contrasts with how we think about physical things; after all, if we ask how many *brains* the split-brain patient has,

any uncertainty as to whether to say “one” or “two” seems merely semantic, because we can always shift away from a language of countable brains and speak in terms of neural parts and their relations. We can say that all the normal parts of a brain are present, but they are no longer interacting as they were before.

Human beings with impaired unity provide one set of cases where nature asks us questions that anti-combination makes it hard to answer. But there are other cases too, such as conjoined twins who share not only parts of the body but also parts of the brain (cf. Langland-Hassan 2015, N.d.), and animals whose nervous systems are less centralized than ours. The most salient example of the latter is the octopus, two-thirds of whose neurons reside in its eight arms, each of which appears to have significant autonomy and responsibility for advanced sensorimotor processing (see, e.g., Sumbre et al. 2001; Gutfreund et al. 2006; Gutnick et al. 2011; Godfrey-Smith 2016; Diamante 2017). Other creatures with decentralized nervous systems, like sea stars and jellyfish, are less clearly conscious at all, but if they are their consciousness presumably will lack the tight integration into a single unit that ours displays (cf. Sterne 1891). Similar things hold for eusocial animals like ants and bees, whose colonies are sometimes regarded as candidates for mentality in their own right: if there is consciousness here, it likely inheres both in individual ants and in the colony, in which case combinationism provides the most natural way to make sense of it. Hypothetical alien hive-minds and futuristic speculation about artificial intelligence provide more scope for envisaging minds which can overlap, contain, and compose one another.

1.1.4. Ordinary Psychological Division

As well as unusual human beings, we all experience what sometimes feels like the beginnings of “internal division,” when mental conflicts become so intense as to prompt descriptions like “I am at war with myself” or “I am enslaved by my passions.” There is a long history of philosophical psychologies that posit some **(p.12)** sort of division of the mind into parts to explain this kind of experience, from Plato’s (2000, 111ff.) division into reason, spirit, and appetite to Freud’s (1949) division into ego, superego, and id. Other writers, though they do not speak of particular parts, do discuss the importance of avoiding various sorts of “inner division” (Frankfurt 1987; Korsgaard 2009). Cognitive scientists regularly analyze everyday cognitive processes as involving the collaboration of brain systems whose behavior is simpler than ours but still mind-like; examples include the “homunculi” posited in various models of (human and artificial) intelligence (see Selfridge 1959; Dennett 1978b, 1991; Lycan 1995); the informationally encapsulated “modules” suggested to autonomously handle certain domain-specific tasks, from language understanding to face recognition (see, e.g., Fodor 1983; Cosmides and Tooby 1992; cf. Prinz 2006); and the quick-and-dirty “system 1” and slow-and-careful “system 2” distinguished by dual-systems theory (see, e.g., Evans 2003; Frankish 2010; Kahneman 2011). Indeed,

one influential conception of how psychological explanation works *in general*, the "neomechanist" approach advocated by Bechtel and Craver (see, e.g., Bechtel 1994; Craver 2015; cf. Rosenberg 1994), identifies as central the decomposition of systems into components, each of which performs parts of the task whose performance by the whole is to be explained. And after considering the split-brain phenomenon, it is hard not to wonder what stops the individual hemispheres of my brain from being conscious in their own right, even in normal cases when they are closely communicating with each other (cf. Nagel 1971, 409; Blackmon 2016).

Of course, it is not clear how much of this talk should be taken as asserting a literal compositeness in human minds, and even if a given statement does assert compositeness, it is not clear what type of entity the parts are meant to be. The Freudian id, for instance, seems to be a mental being of some sort, but should we think of it as a conscious subject with its own phenomenology? And if I fail to achieve some sort of ideal "unity of self," what or who are the things which exist in place of the single self which was aimed for? Indeed, one might worry that treating components of the human mind as mind-like is just a form of anthropomorphism, a "Homuncular fallacy" that inappropriately extends the concepts applicable at one level to another level (Bechtel 1994, 19–20; cf. Bennett and Hacker 2003). But it is only anthropomorphism if those components are not in fact mind-like, and we cannot decide this in advance of examining how they actually work.

Certainly, I do not mean to insist on a literal reading of all such talk, nor to suggest that it is always true, when interpreted literally—just to undermine the confidence often felt that they *cannot* be literally true, because no part of a subject could literally be conscious itself. Combinationism does not say that all parts of subjects, or any particular parts, are themselves subjects: only that we should be open to **(p.13)** the possibility. The implications of combinationism for consciousness in parts of the mind are thus rather like the implications, for consciousness in nonhuman animals, of realizing that human beings are not endowed with any special, supernatural soul; we must evaluate each entity for consciousness on its own merits, not prejudge it based on how it relates to us whole human beings.

1.1.5. Overlapping Parts of a Human Being

Finally, we can even set aside all unusual or problematic human conditions and the details of human brain anatomy, and just consider some unremarkable everyday assumptions. Surely a human head is intrinsically capable of supporting consciousness; surely human heads exist; but surely we are not ourselves heads but rather have heads as a part of us. Philosophers have pointed out that it seems to follow that both we and our heads are conscious, so that for every human being there are in fact two different conscious beings—or more, if we also consider such entities as "the top half of a human being," "the brain of a

human being," or "all of a human being minus 1 foot" (Merricks 2001, 95; Unger 2006; cf. Burke 1994, 2003; Gilmore 2017). Some philosophers have regarded this as a paradoxical result (a problem of "too-many-minds"),[9] but typically they accept that a parallel result is unproblematic for physical properties: my head and I are two different entities with mass, two different entities which possess eyes, two different entities which fill this region of about a cubic foot of space (Merricks 2001, 106; Unger 2006, 378–379). These results are unproblematic because we accept that one being with mass (for instance) can be part of another, and that then they will both have mass, but the mass of one will include the mass of the other. What makes the parallel result with consciousness seem absurd is the assumption that even when one conscious being is part of another, the one's consciousness cannot *contain* that of the other—i.e., the assumption of anti-combination.

So a successful defense of combinationism would have implications for a variety of debates. We should reject the Achilles argument for substance dualism and should refrain from adding anti-nesting principles to our theories of mind. We should accept the possibility in principle of conscious social groups (and not reject theories just for having that implication), though this leaves open which particular groups might qualify. We should also regard the division, multiplication, and overlap of people as no more theoretically problematic than that of physical objects—and in particular should take seriously models of the split-brain, or of **(p.14)** dissociative identity disorder, or of everyday life, on which there is both a single whole person and also one or more component persons.

Chapter 3 contains a direct answer to "too-many-minds"–style problems about heads and brains, with subsequent chapters considering more complicated cases. In chapters 5 and 6 I discuss the split-brain phenomenon, the consciousness of cerebral hemispheres, homuncular accounts of cognition, and other interesting test cases, as well as sketching how combinationists should think about social groups, both real ones and thought-experimental constructs. And in chapter 7 I discuss how far a theory of experiential combination can support literal or near-literal readings of everyday forms of both "inner division" and dissociative identity disorder.

1.2. Motivating Panpsychism

Although combinationism is relevant to many issues in the metaphysics of mind, it is perhaps *most* relevant to panpsychist theories of consciousness, and in particular "constitutive Russellian panpsychism" (hereafter CRP). Indeed, I have taken the term "combination," as meaning explanatory, nonemergent composition, from the "combination problem" supposed to face CRP (Seager 1995, 280). I am sympathetic to CRP, but my aim in this book is not to argue directly for it: rather, by defending the possibility of composite subjectivity, I hope to solve the combination problem, and thus remove the most serious

objection to CRP. So it will be useful to say a bit to acquaint the reader with the view and some reasons for holding it.

Panpsychism (from Greek *pan-* and *psuche*, meaning “all-” and “mind”) is the view that consciousness is omnipresent among the fundamental things of the universe: all matter is conscious. If the fundamental things are particles, then every particle of matter has a point of view, an iota of consciousness, which is unimaginably simple but nevertheless differs from our own consciousness only by degree. If the fundamental things are something else—waves, fields, strings, spacetime, or the universe as a whole—then panpsychism says correspondingly that those have some rudimentary glimmer of consciousness. It does not automatically follow that “everything” is conscious, for many things (shoes, ships, sealing wax, etc.) are nonfundamental, built up out of the basic constituents. Whether those composite things are conscious will depend on exactly how consciousness combines.

1.2.1. The Explanatory Argument for Panpsychism

There are four major arguments in support of panpsychism, which I will call the “explanatory” argument, the “causal exclusion” argument, the “intrinsic **(p.15)** natures” argument, and the “continuity” argument. In the following pages I will outline these, as well as distinguishing between constitutive and emergentist panpsychism, with the explanatory, causal exclusion, and continuity arguments favoring the constitutive option, and between Russellian and dualist varieties, with the causal exclusion and intrinsic natures arguments favoring the Russellian option.

The first argument begins with dissatisfaction about “physicalism” as an explanation of consciousness (Nagel 1986; Seager 1995; Chalmers 1995; Strawson 2006), combined with a desire to hold onto “naturalism.” By “naturalism” I mean the view that the world contains a single basic type of stuff, whose behavior is governed by a single set of simple, general laws, and that these laws are those revealed by science. The most common version of naturalism among contemporary philosophers is physicalism, the view that the world is entirely made up of matter, and matter is exhaustively described by physics. But some philosophers reject physicalism, even while accepting naturalism, holding that matter is not *exhaustively* described by physics—there are fundamental aspects of matter that physics is blind to. In particular (they tend to say), there are certain things each of us can know about matter, such as that one particular portion of matter (the one between our ears) sometimes feels and thinks and experiences, which go beyond both what physics itself says and what can be deduced from any physical description, no matter how detailed.[10] Because facts about my consciousness are left out by any purely physical descriptions, these “naturalistic anti-physicalists” infer that consciousness must be itself a fundamental feature of reality, no more derivable from physical properties than mass is derivable from charge.[11] Yet consciousness does seem

intimately tied to the workings of the physical brain, suggesting that they are linked by *a posteriori* laws of nature (like those saying how much charge electrons have), even if not by *a priori* conceptual necessities (like those saying that a square has four sides). The resultant picture is one on which fundamental "psychophysical" laws bridge the gap, connecting physical and experiential properties just as different physical properties are related by fundamental physical laws.

(p.16) So far this is not necessarily panpsychist; the fundamental psychophysical laws might be "emergence laws," associating consciousness only with certain physical composites (e.g., just with human brains) and not with all matter. The next move is to point out that fundamental laws tend to be simple and general, allowing for a great variety of forms to be built up gradually from a small set of widespread basic elements. They do not attach a fundamental element to a precisely specified sort of rare and complex structure. Moreover, it is arguably this simplicity and generality that makes physics an explanatorily satisfying framework—by contrast, more narrowly applicable "emergence laws" seem unsatisfyingly ad hoc. Consequently, we should expect psychophysical laws to put consciousness more or less everywhere: once naturalistic anti-physicalism is accepted, panpsychism is the natural conclusion.

1.2.2. The Intrinsic Natures Argument for Panpsychism

Next, the intrinsic natures argument tries to show that any alternative to panpsychism is both unparsimonious and obscure (Seager 2006; Strawson 2006; Coleman 2009; Goff 2017a, 139ff.). It works from three premises: that the properties expressed by the language of physics are in some sense "merely structural"; that conscious properties, alone among all the properties we understand, are more than structural; and that all structural properties must be instantiated in something with a nature that goes beyond structure. There is some dispute over how best to define "merely structural" (cf. Stoljar 2013; Mørch 2014),[12] but the rough idea can be brought out by observing that, apart from spatiotemporal terms, all the fundamental properties ascribed by physics—force, energy, mass, charge, etc.—are defined by their place in equations linking them to other such properties. What is it to have charge? It is to have whatever property it is which produces forces when at some distance from another thing with charge. But what is force? It is whatever accelerates mass. What is mass? It is whatever resists acceleration when exposed to force and exerts attractive forces on other things with mass. And so on. All of this still seems to leave unanswered the question of *what* these interdefined properties actually are; physics tells us about the structure that physical properties fit into, but not what those properties actually are. By contrast, it seems as if when we experience some feeling, like sadness or pleasure or the sensation of heat, we *do* know what it is to have that property—perhaps we know this in an inarticulate, **(p.17)**

hard-to-express way, and perhaps not in all respects, but we are not caught in a network of "whatever it is that . . ." as we are with physical properties.

The intrinsic natures argument then proceeds as follows. Even if physics does not tell us what physical properties really are in themselves—even if it does not tell us their "intrinsic natures"—they must still be something. The structure physics reveals must be implemented in something. So what could this something be—what are the intrinsic natures of physical properties? Experiential properties are one candidate: maybe what we call "mass" is a certain kind of experience, whose characteristic effects and distribution are captured by the equations of physics. Or perhaps the intrinsic natures of physical properties are something completely inconceivable to us, something of which we neither have nor can form any positive conception—call this the "noumenalist" option, borrowing Kant's term for the forever-unknown way things are in themselves (Goff 2017a, 170). The panpsychist claims that these are the only two options, and that we should prefer the first option, which implies panpsychism. Why should we prefer it? Partly, perhaps, because it is unreasonably defeatist to adopt a noumenalist position when a more positive view is available. But also, partly, because noumenalism is unparsimonious. *Some* parts of the material world have conscious experience as their intrinsic nature—namely, human brain processes. So the noumenalist position, which says that some matter has one nature and other matter has a different nature, seems to be needlessly multiplying the number of basic types in the world. Hence Occam's razor cuts against it.

1.2.3. The Causal Exclusion Argument for Panpsychism

Third, the causal exclusion argument for panpsychism (closely analogous to a similarly named argument for physicalism) says that because consciousness seems to have power to affect the physical world, and because events in the physical world seem to be entirely driven by the causal powers of physical things, consciousness must be some part of the physical world (Rosenberg 2004; Chalmers 2015). If consciousness made a causal difference but was separate from physical reality, then its actions would appear as violations of laws of physics—violations which we do not seem to have yet detected and which it seems theoretically unattractive to have to predict. This argument is compatible with some forms of physicalism, but is also compatible with some forms of panpsychism; what it is not compatible with are dualist theories on which consciousness "floats free," independent of the physical world and its causal laws.

(p.18) #### 1.2.4. The Continuity Argument for Panpsychism

Finally, the continuity argument (James 1890, 147–148; Clifford 1874/86, 266ff; Goff 2013; Mørch 2014, 153–154; Buchanan and Roelofs 2018) motivates panpsychism by pointing out that scientific progress seems to show that there is nothing supernatural about humanity, nothing in our development as individuals

or as a species that marks a genuinely sharp break with the rest of nature. But for panpsychism to be false, there would have to be such a sharp break—a moment when the most “advanced” nonconscious thing was succeeded by the most rudimentary conscious thing, a moment when “the lights turn on.” Our evolution was gradual, our fetal development is gradual, and even when there does appear to be a “sudden leap” from a nonconscious stage to a conscious stage, the process always turns out to be resolvable into a gradual sequence of steps when looked at on smaller timescales. As Chalmers (1996, 297) says, panpsychism “avoid[s] the need for consciousness to ‘wink in’ at a certain level of complexity . . . [because] any specific point seems arbitrary, so a theory that avoids having to make this decision gains a certain simplicity.”

1.2.5. Varieties of Panpsychism: Constitutive and Emergentist

Not all panpsychists are combinationists: even if all the basic physical entities are conscious, the consciousness of complex physical entities might not be a mere combination of their parts. There might then be a form of panpsychism on which the experiential properties of human beings are emergent, not mere combinations, generated according to “emergence laws” that directly connect complex consciousness with certain sorts of physical composite.[13] I will use the labels “constitutive” and “emergentist” for combinationist and noncombinationist versions of panpsychism.

This dependence on the challenging idea of composite subjectivity is sometimes used as an argument within the panpsychist camp, against constitutive versions and in support of emergentist versions. But just as often, it is used in attacks on panpsychism as a whole, by critics who assume that the only or best forms of panpsychism must be constitutive. The reason for this assumption is clear when we consider the explanatory argument for panpsychism: for this argument to support **(p.19)** panpsychism against physicalist and emergentist alternatives, panpsychism needs to offer an explanatory advantage. But if microexperience does not ground or explain macroexperience, what advantage is there? To put it another way, if human consciousness “emerges” from microexperience, in the strong sense that the latter does not ground and explain the former, why not just say that human consciousness emerges from nonconscious microphysics, in an equally mysterious way?

Another reason to prefer constitutive over emergentist panpsychism is to avoid causal exclusion: emergentist panpsychism risks having microexperience leave macroexperience with no causal work to do. If the relationship between human minds and microminds is closely analogous to that between macrophysical objects and microphysical objects, then it is seems very plausible that the former, like the latter, allows for shared causal efficacy: what the parts do, the whole also counts as doing, but the result is not somehow “caused twice over” (not “overdetermined”). But if macroexperience is emergent relative to

microexperience, then it is much harder to see how both can be efficacious without their effects being overdetermined.

1.2.6. Varieties of Panpsychism: Russellian and Dualistic

As well as distinguishing constitutive and emergentist versions of panpsychism, we can distinguish Russellian and what I will call “dualistic” versions. Russellian panpsychism uses the conceptual framework developed for the intrinsic natures argument (which is traced to Russell [1927] and Eddington [1929]; cf. Schopenhauer 1969, 97–105) to understand the relationship between physical and experiential properties: the former are just more abstract descriptions of the causal and structural roles played by the latter. This unites the two sets of properties more closely than on non-Russellian versions of panpsychism, on which each physical ultimate simply has two, metaphysically independent sets of properties (hence “dualistic” panpsychism). On the Russellian view, it is not just that each particle has mass, charge, etc. and then *also* has experiential properties, but rather each particle has experiential properties, which ground a set of causal powers whose structure is captured by the equations of physics. Obviously, anyone persuaded by the intrinsic natures argument has reason to prefer Russellian over dualistic panpsychism, since only the former responds to the demand to know the intrinsic natures of physical properties. But the causal exclusion argument also provides reason to prefer Russellianism, since metaphysically independent properties, like the physical and experiential properties posited by dualistic panpsychism, seem to be at risk of causal competition. By analogy, if a particle causes some effect just in virtue of its charge, there seems to be no room left for its mass to also cause that **(p.20)** very same effect. But if experiential and physical properties are related as intrinsic basis and structural role, then effects attributed to one can also be attributed to the other.

In summary, the arguments which support panpsychism over non-panpsychist views also seem to favor the constitutive, Russellian, version over other versions —at least if that version is workable at all. But this version is arguably the hardest to spell out, most metaphysically demanding version of panpsychism. What is attractive about it is that it ties together the different facets of the world: it connects micro and macro by being constitutive, physical and experiential by being Russellian. But this simultaneously makes it easier to refute, just by finding any incompatibility between the supposedly tied-together facets.

This is why the combination problem faces CRP especially strongly. Because it is Russellian, it is committed to the fundamental experiential properties being isomorphic with the fundamental physical properties that they are supposed to be the intrinsic natures of. This means that the basic experiential properties will be both quite few in number and also extremely widespread in the universe. And

because CRP is constitutive, it is committed to all the diverse and specific forms of consciousness that exist being generated directly out of this basis.

In the previous section I tried to show that the difficulty of conceiving consciousness as composite is not exclusive to panpsychists: some version of the "combination problem" faces a wide range of views in the metaphysics of mind. But CRP implies an especially extreme and thoroughgoing sort of experiential combination, and consequently also faces an especially extreme and thoroughgoing sort of combination problem, with especially constrained resources for solving it (Chalmers [2017, 211] compares it to "trying to juggle seven balls in the air with both hands tied behind one's back"). Yet it is also, if it can be made to work, an enormously appealing and powerful theory, for precisely the same reason it faces this problem over combination: the deep unity it postulates in nature.

1.3. Defining Combination

The definition of combinationism given above employs the following five ideas: "grounding," "explanation," "part," "whole," and "experiential." Combinationism says that a whole's experiential properties can, in some cases, be grounded and explained by those of its parts. In this section I clarify what I mean by these terms.

(p.21) 1.3.1. What Is "Grounding"?

By saying that A "is grounded in" B, or equivalently that A is "nothing over and above" B, I mean to capture the, admittedly rough and intuitive, idea that once we have B, that is by itself enough for us to have A.[14] This is, in the first instance, a relation among facts: for one object or property or event to ground another is for all the facts about one to be grounded in facts about the other.

This is compatible with a few different cases: most simply, we can say that A is grounded in B if A is B, for nothing is anything over and above itself. Thus when a physicalist claims that mental events are nothing over and above certain physical brain events, one thing they might mean is that those brain events are identical to mental events: we have two terms ("mental event" and "brain event") referring to one and the same thing.

Identity is symmetrical, so if B grounds A by being A, it will also be true that A grounds B (though see Jenkins 2011). But a second, asymmetrical way for B to ground A is for A to be simply an "abstraction from" B, a less specific or detailed version of B. Thus when someone claims that a certain lizard's being red is nothing over and above its being scarlet, they plausibly mean that being scarlet is a more specific way of being red, so that once something is scarlet nothing more is needed for it to be red.

Other cases are trickier. It seems mistaken to simply *identify* the nation of Canada with its current population (it could endure after they are all dead), or its territory (that existed long before Canada did), or with a certain system of government (it could have a revolution and still be Canada). Yet it seems clear that Canada is nothing over and above certain people implementing a certain institutional structure within a certain territory (cf. Parfit 1999, 17–18). To see why this is so, observe that it seems possible to give logically sufficient conditions for there to be a nation, mentioning only people, institutional structure, and territory, and then also to give logically sufficient conditions for two nations to be the very same nation, in terms of the continuity of those same factors (continuity meaning, roughly, that though any given factor may change, there is only limited, gradual change from one moment to the next). Given all this, and given an initial specification that some nation is Canada (e.g., the nation founded on such-and-such a date at such-and-such a place), we have a complete account of what it would be for Canada to exist at any time—namely that people must implement institutions **(p.22)** within a territory in such a way that a nation exists, and that a sufficiently continuous history of people implementing institutions within a territory must connect that nation with the nation initially named “Canada.” This is grounding, but neither identity nor abstraction.

In cases like this a key role is played by the fact that we can “metaphysically analyze” the grounded entity: thinking about what a nation is reveals a set of conditions for the existence of a nation, and what grounds the nation are the entities (people, territories, etc.) whose activities fulfill those conditions. Goff (2017a, 44ff.) calls this “grounding by analysis,” noting the important implication that only entities which admit of metaphysical analysis—only entities whose nature can be analyzed into a set of conditions that other entities could then fulfill—can be grounded in this fashion. If some entities are “primitive” and cannot be analyzed except in a circular way (e.g., “occupying space” is primitive if the only analyses we can give of it are things like “filling space,” “having length, breadth, and depth,” or “being located somewhere,” which are just other ways to say “occupying space”), then facts about them cannot be grounded by analysis (see also Dasgupta 2014, 564–580; cf. Melnyk 2003, 21, 2014).

I will adopt the following rough criterion: A is grounded in B if all it takes for A to exist is for something to be true of B (to use a popular metaphor, all God would have to do to create both is to create B). When A and B are identical, this criterion is clearly satisfied: for A to exist is just for B to exist. When A is a less specific version of B, we can say that for A to exist is just for B to be any of a range of ways. And when A is grounded by analysis, the relevant state of B may be a complex and intricate one, but could in principle be articulated by analyzing the nature of A.

1.3.2. What Is “Explanation”?

What about “explanation”? In many of the above cases of grounding, understanding the grounding thing would allow us to understand the thing grounded—would let us “see why” various facts about it obtain. The difference between explanation and grounding is that one concerns the reality of things, and the other concerns possible states of understanding by us. I will say that A is explained by B when someone who fully understands B is thereby in a position to fully understand A; to put it another way, A is “made intelligible” by B.

This is not by itself a real analysis, since notions like “understanding” and “intelligibility” are little clearer than the notion of “explanation.” And not all sorts of “explanation” are relevant here; for instance, I am interested in how components existing at a given moment explain a whole’s properties at that very time, not in across-time explanations, where we explain events in terms of separate, earlier events.

(p.23) Rather, my major point of reference is the supposed explanatory gap between physical properties and consciousness, which contrasts with cases where one set of physical properties intuitively suffices to explain another (see, e.g., Levine 1983; Chalmers 1995; Loar 1990; McGinn 1989). A complete understanding of the microphysical structure and dynamics of a human brain would, in principle, let us “see” why it must have the macroscopic physical features which it in fact has (Levine [1983, 357]: “our knowledge of physics and chemistry makes [such connections] intelligible”), yet would not allow us to “see” why it feels a certain way to be that brain (Levine: it “leaves the connection . . . completely mysterious” [357]). The microphysics explains the macrophysics, but not the phenomenology. So I will index my discussion to the explanatory gap between consciousness and physics: Could there be the kind of connection between mental wholes and mental parts which defenders of the explanatory gap think is missing between consciousness and physics, but accept between physical wholes and physical parts?

Although the relations of grounding and explanation are importantly different, it is natural to regard them as connected. If one thing fully explains another, that is usually a good sign that one grounds the other; conversely, if one thing is nothing over and above another, shouldn’t we expect that understanding the latter would let us understand the former? But this cannot always be assumed: it might be that despite some A being fully grounded in some B, there is some stubborn cognitive fact about us that prevents us from understanding the one in light of the other: there might be an explanatory gap even without a real gap, just as it has sometimes been argued that the explanatory gap between consciousness and matter does not betoken any metaphysical difference, but only the presence of two ways to think about the same properties, which we are

incapable of intelligibly connecting (e.g., Levine 1983; Loar 1990; Block and Stalnaker 1999; Diaz-Leon 2011).

Even if there are cases where grounding and explanation come apart, the compositional relations between physical properties do not appear to be such a case. If, for instance, we know the present locations of all the parts of a table, it seems clear both that there is no additional reality needed for the table to be located where it is, and that there is no additional mystery for us about why it is located there. This is why I have defined combination in terms of both explanation and grounding: my consciousness is a mere combination of the consciousnesses of my parts if and only if it is *both* grounded in them and their interrelations *and* fully explained by them and their interrelations.

(p.24) 1.3.3. What Are "Parts"?

Next, what does it mean to say that something has "parts," or that one thing is "part of" another? Philosophers generally agree that parthood is a transitive relation (parts of a part of me are parts of me), and that two distinct things cannot both be part of the other, and that if I have one part, I must have at least one other part (what is "left over" from removing the first part). Beyond this the notion seems to be primitive: we do not know how to analyze it into terms that do not themselves presuppose a grasp of parthood. But it also seems to be a notion employed in different senses, among which it is important to distinguish.

First, we can distinguish "thick" and "thin" notions of parthood. In the thin sense, any feature or aspect or property of something—anything we can truly ascribe to it—can be called a part of it. We might say, of a dog that is a hybrid of breeds A and B, and which is currently behaving very much like a member of breed A, that it is displaying "the A part of it." We might casually convey that we want two opposed things by saying "part of me wants X, but part of me wants Y." We might perhaps even talk about a red square in terms of its two parts, "redness" and "squareness" (cf. properties as "logical parts" [Paul 2002]). But the "thick" notion marks a contrast with a thing's mere properties: my height, or my build, or my demeanor, are properties of me but not parts of me. The thick notion seems to involve some or all of the following conditions:

1. The parts of something exist simultaneously with the whole.
2. The parts of something are of the same basic category as the whole.
3. The (discrete) parts of something are existentially independent of each other.[15]
4. A thing is existentially dependent on its parts collectively.

The first condition distinguishes parts of something from what goes into making it. To use an example from Bennett (2011, 288), eggs may often be ingredients in cakes, but an egg is only a part of the cake if, even once the cake is complete, there still exists an egg somewhere inside it. When we divide something into

parts we are identifying constituents of it that currently exist and currently compose it.

The second condition rules out counting properties of an object as parts of it: parts of an object should themselves be objects, though a property might have other properties as parts (e.g., part of being a bachelor is being male). Similarly, events involving me are not parts of me, and while I may "play a part" in an event **(p.25)** like a hockey game, I am not a part of it in the same sense as its first half or first period (other, shorter events). Obviously parts and wholes need not share all categories—parts of a table need not themselves be tables. But the parts of a table should at least be physical objects.[16]

The third condition says that of any two parts (which do not have a further subpart in common), the one might cease to exist without the other automatically ceasing to exist. They are in this respect like any two things of their kind, considered apart from their composing a whole. The four legs of a table, for instance, are existentially independent things just as much as two unrelated tables are: the existence or nonexistence of the one does not logically imply the existence or nonexistence of the other. This stops things like "the-table-as-square" and "the-table-as-wooden" being parts of the table (cf. Fine 1982, 1999), since doing away with the table-as-square (that is, doing away with the table, considered as a square object) would automatically do away with the table-as-wooden (for it gets rid of the table itself).[17]

The fourth condition says that if the parts all suddenly ceased to exist, the whole would also thereby be gotten rid of. If the four legs and the flat top of the table were to vanish, that would simply be the table vanishing. This rules out treating the causal contributors to something as parts of it. For instance, my parents have had a huge role in forming me, both physically and psychologically. (I might say "they are part of who I am.") But they are not parts of me in the thick sense, because they (and all my other formative influences) could cease to exist without me thereby also vanishing.

Note that this condition does not say that the whole could not exist if the parts did not exist; I have many cells as parts right now, but after a few years most of those cells will not exist while I still do, made of a different set of cells. But this is compatible with me presently depending on those cells in the sense that their **(p.26)** instantaneous and complete disappearance would be my own instantaneous and complete disappearance. Note also that while this means that the existence of a whole is grounded in facts about its parts, it does not follow that all the properties of a whole are grounded in properties of its parts (i.e., are mere combinations); though I suspect the latter is true, it is not true by definition.

So the thick sense of parthood requires that parts and whole exist simultaneously and belong to the same category, that parts are existentially independent, and that the whole existentially depends on the parts. In this book I will generally be operating with the "thick" notion of part, and I intend my definitions of "combination" and "anti-combination" to use this notion. This distinguishes combinationism from certain nearby ideas, namely bundle theory and fusionism. By "bundle theory" I here mean the idea of the subject as a whole comprised of mental states, which are not themselves subjects. Since here the "parts" and "whole" are things of different kinds, bundle theory does not by itself count as combinationism—that would require adding the idea that some or all "sub-bundles" within the mind also qualify as subjects. Fusionism is the idea that two or more subjects can "fuse" into a single, more complex subject—after which they themselves no longer exist. (Seager [2010, 2017] and Mørch [2014] both advance fusionism as a possible solution to panpsychism's combination problem.) Here the "parts" and "wholes" never coexist: there is only first the many and then the one, and so in the thick sense we should not call this a compositional relationship.

1.3.4. What Are "Wholes"?

As well as different ways of conceiving of what a thing's parts are, there are different ways of conceiving of wholes. Perhaps the simplest is to think of wholes as just "the many parts counted as one thing" (Baxter 1988, 579). In this sense, to say that two things "compose a whole" is just to group them in thought and name them, to "take them as one" rather than as many. This is the sense of "whole" which fits best with the formal system known as classical extensional mereology (Leonard and Goodman 1940; Lewis 1991, 72–90), with its axioms of unrestricted and unique composition: for any set of things, there is one and only one thing which they compose. Because, after all, for any set of things, we can count them as one thing. Call this very modest sort of "whole" an "aggregate."

Sometimes, however, we seem to think of a "whole" as defined not just by the parts which compose it but by a specific structure they must instantiate, such that it exists only when its parts are structured in that way. A car, for instance, is not simply an aggregate of mechanical parts, because those parts could exist when separated, but the car exists only when they are connected, and connected in the **(p.27)** right way (e.g., Fine 1999; Koslicki 2008; cf. Wilson 1999, 2013). Call wholes like this, which exist only when a certain structure is instantiated, "structure-specific wholes."[18]

Whether we think of wholes as aggregates or as structure-specific does not by itself change the plausibility of combination, i.e., of deriving all their properties from the properties of their parts, and the relations among them. But it does change the shape of the derivation. Because aggregates in some sense just *are* their many parts, taken together, they can be expected to have a property corresponding to every property of their parts. Structure-specific wholes, on the

other hand, might well be thought to *lack* properties corresponding to those properties of their parts which are irrelevant to their defining structure. That is, structure-specific wholes can have a limited measure of "autonomy" in the sense of being unaffected by all the messy details of their parts. According to Wilson (2010, 2013), it is crucial to the "ontological autonomy" of certain wholes that they have a different set of properties than their parts, and in particular have fewer degrees of freedom, that is, require fewer parameters to fully specify their state. Consider, for example, a whirlpool which forms in water, sucking down fresh water while constantly preserving its vortical structure. It seems to be composed of water molecules, each of which weighs something, but it's not clear that it makes sense for us to ascribe to it any weight at all. Or consider the physical parts of a biological cell, which all have their own specific masses, charges, and locations: the aggregate they compose will correspondingly have a very detailed distribution of mass and charge properties at specific locations. But beyond certain parameters, the details are irrelevant to the preservation of the cell's biological structure, and so we might think of the cell itself as having only the property of "mass and charge distributions being within acceptable parameters," and nothing more detailed than that. But this does not mean that the whole has any properties not derived from its parts.

But there is also a strand of philosophical thinking that separates out "real" wholes much more sharply from "mere" aggregates, which demands that to count as a whole, a thing must have some sort of deep unity, something to make it objectively one and not many (see, e.g., Leibniz 1989, 78–80, 85–90; James 1890, 158–160; cf. Merricks 2001; Shani 2010). Call this sort of whole a "true unit." Whereas for something to be an aggregate is just for us to "take it as one," for something to **(p.28)** be a true unit is for it to be such that we *must* take it as one and *cannot* consider it as a set of parts without losing something crucial to it. A true unit is meant to be "more than the sum of its parts," not just in the innocuous sense that the parts can do different things when working together than when working separately, but in the more demanding sense that the whole should do things that go beyond anything the parts could do, working together or separately. In effect, a true unit should be a strongly emergent whole, a whole whose properties escape any accounting in terms of the properties of its parts.[19]

There is room for lots of metaphysical debate about which of these three types of whole exist, what their relationship is, and which category we ourselves might fall into.[20] Since true units are, almost by definition, not open to a wholly compositional explanation, my defense of combinationism must presuppose that our minds are not true units; conveniently, the scientific progress of the past few centuries seems to have proceeded in large part by abandoning the thought that anything in nature is a true unit, perhaps with the (very questionable) exception of elementary particles. But beyond that, I want to remain neutral about how far we should be thinking of composite subjects as aggregates or as structure-specific wholes, and discuss the combinationist project from both perspectives.

After all, part of what motivates combinationism about consciousness is the thoroughgoing compositionality of the physical world, and this involves the explicability of both physical aggregates and physical structure-specific wholes. Chapters 3 and 4 treat subjects as aggregates, while chapters 5–8 treat them as structure-specific wholes. In this way, hopefully, I will catch the right theory somewhere in my net.

1.3.5. What Are "Experiences" and "Subjects of Experience"?

The categories of "experience" and "subject" are hard to define exactly, but easy to gesture at. Whenever there is consciousness (whenever there is "something it is like" for a certain being to be in a certain state, "from the inside"),[21] there is both **(p.29)** some*one* who is having experiences and some*thing* that is happening in or to them. The former might be called a person, soul, self, or, as I will tend to say, a "subject of experience." The latter might be called feelings, mental states, states of consciousness, experiencings, or, as I will tend to say, "experiences."

Going by the way we usually talk about them, subjects are a sort of "thing" or object, which persists through an indefinite period of time, whereas experiences are something like an event, which lasts for some limited duration. Insofar as the one is an object and the other is an event, neither can be a part of the other in the thick sense. Both subjects and experiences, moreover, should be distinguished from experiential properties, properties which are defined by what it is like to instantiate them:[22] on this definition, anything that instantiates experiential properties is a subject, and whenever they instantiate experiential properties they will be "having" some experience or other. But experiences do not seem to be the same things as experiential properties, for the former are particulars (things that occur to specific subjects, at specific times), and the latter are universals (things that can occur to many different individuals on many different occasions). Rather, experiences are particular instances of experiential properties.

The distinction between subjects and experiences is important, because the controversial idea I wish to defend, composite subjectivity, goes beyond the familiar idea that experiences can have other experiences as parts, and even be mere combinations of those parts. It is very natural to think that my "field of experience" (the totality of what I am conscious of at a given moment) can be divided into components, such as my present total visual experience, which is a part distinct from my present auditory or olfactory or emotional experiences. It seems equally unobjectionable to divide my total visual experience into component experiences, like my experience of some particular visible object (e.g., my coffee cup). But while these component experiences might be called "parts of my consciousness," they are not themselves parts of *me*, the subject, because they are not themselves conscious in the way that I am. One can accept the existence of a **(p.30)** component experience of seeing my coffee cup,

without accepting the existence of any conscious subject which is conscious of seeing that cup and nothing else.[23]

Although most writers accept that experiences can be composed of other experiences (see Tye 2003, 20–41, for a dissenting view), there is disagreement about how best to individuate experiences, and how to think about their composition. I do not think there is a single right way to individuate experiences: as Bayne (2010, 24) says, “counting experiences is arguably more like counting the number of objects in a room . . . [than] like counting the number of beans in a dish. . . . The idea that there is only one way in which to proceed is somewhat farcical.” For simplicity I will adopt the following approach: for each feature of what it is currently like to be me (e.g., I feel sad, and I also feel hot), there is “an experience” (e.g., my sadness-experience and my hotness-experience), and one experience is part of another when my having the one includes and subsumes my having the other (e.g., “I feel hot all over” includes “I feel hot in my feet,” so the hotness I feel in my feet is part of my total hotness-experience).

So we can distinguish between experiences composing experiences and subjects composing subjects. But even having subjects composing subjects would not automatically qualify as composite subjectivity in my sense. For there might be a subject that I contain within me somehow—a sentient parasite, a late-term fetus, a second small brain grown in an unusual place—whose consciousness had nothing to do with my consciousness. To rule out this kind of case, I defined combinationism in terms of a part-whole relation where both parts and whole are subjects (because they instantiate experiential properties), and where the experiential properties of the latter are combinations of the experiential properties of the former: a relation between subjects, defined partly in terms of relations between their properties. I will use the terms “component subjects” and “composite subjects” for parts and wholes in this relation. (I will use “they” pronouns for such entities, even when singular; I apologize for the slightly increased ambiguity in some sentences.)

A composite subject might be divided into component subjects along various alternative lines, just as a human body can be divided into many alternative sets of parts. Some divisions (“top-down”) might focus on the distinctive, complicated features of the whole, like person-level psychology. Thus if we observe a conflict between two styles of inference someone employs, we might try to distinguish two inferential systems within them and hypothesize that each of them is a component subject. By contrast, other divisions (“bottom-up”) might focus on the basic, **(p.31)** widely shared, more fundamental features of the whole. If the left half of me (or of my brain) and the right half of me (or of my brain) are conscious subjects in their own right, they would be parts in a division of this sort.

The examples of the split-brain phenomenon and dissociative identity disorder (DID), discussed in section 1.1, nicely illustrate this contrast. The two apparent component subjects in the split-brain patient are divided at an anatomical level: one consists of the left hemisphere, the nonhemispheric brain parts, and (roughly) the right half of the body, while the other consists of the right hemisphere, the nonhemispheric brain parts, and (roughly) the left half of the body (since they share major parts, they overlap). This is a more sophisticated form of division than simply carving the body into spatial chunks, but it can be drawn simply in terms of biological tissue, and is to that extent a relatively "bottom-up" division. The alters of a DID patient, by contrast, correspond to a "top-down" division because to divide them from one another we make reference to high-level features like personality and self-identifications. There is no *a priori* reason to expect the two divisions to line up; the alters are unlikely to be exclusively based in distinct brain areas (cf. Rosenberg 1994).

Any *introspective* division of a subject—a division into "whatever underlies *this* experience" and "whatever underlies *that* experience"—will likely be a fairly "top-down" division, and we cannot assume in advance that it will correspond neatly to any natural division of the brain. The same goes for divisions based on felt motives and feelings, including the divisions that become salient in cases of inner conflict. That is not to say that these introspectively defined subjects are not neuron-based: their physical basis is still some assembly of neurons, but those neurons might be spread widely and eccentrically through the brain. Moreover, each of those neurons or clusters of neurons might also play a role in subserving other functions, and to that extent be included in another part (i.e., it may not be introspectively obvious which parts overlap).

I am not presupposing that all these ways of dividing a human subject will yield component subjects; it is possible that some, all, or none do. But if they did, they would not all yield component subjects of the same kind. Indeed, they might seem to yield parts of such different metaphysical status (e.g., a personality cluster and a hemisphere) that they could hardly *both* be conscious subjects. The question of what kind of thing subjects are will be taken up in more depth in chapter 2.

1.4. Plan of This Book

There are three different ways to object to experiential combination: (i) showing it to be inherently impossible; (ii) claiming that it could not work with certain **(p.32)** particular sorts of parts and wholes we are interested in, given what we know about them; and (iii) pointing out that we have so far no positive theory of it. Call these internal problems, bridging problems, and lack-of-theory problems (the first two terms taken from Coleman 2017, 250–254).

While the first aim to prove that no sort of composite subjectivity could ever exist, the second instead point either to a particular feature of human

consciousness which, they claim, could not be accounted for by any combination of conscious parts, or point to a certain set of putative conscious parts and deny that they are suitable for composing the sort of consciousness that we enjoy. The third sort of objection is that we simply have at present no inkling of how composite subjectivity is supposed to actually work. While this is not in itself an argument against its possibility, it is a reasonable basis for withholding assent from any theory (like CRP) which requires it.

I devote chapter 2 to addressing five internal problems: the subject-combination problem, the unity problem, the privacy problem, the incompatible contexts problem, and the boundary problem. All of these focus, in one way or another, on the metaphysics of the subject and the unity of consciousness, and in chapter 2 I identify the assumptions about subjects and unity that these objections rest on. The rejection of these assumptions provides parameters for a combinationist theory, and the next six chapters then add flesh to the bones. Chapters 3–8 are dedicated to refining, applying, and justifying the abstract claims defended in chapter 2, and thereby to addressing both a number of what I have called "bridging problems" and the more general "lack-of-theory" problem.

Chapters 3–8 actually develop not just one but three theories of mental combination, focused on different issues and tailored to the interests of different constituencies. Combinationism is not in itself a theory of consciousness; rather, it is an idea that can be usefully combined with many different theories. Consequently, I have tried to organize this book so that different parts will be of value to different readers, who can thus skip ahead to the chapters of most interest to them.

First, chapters 3 and 4 present panpsychist combinationism, an account of mental combination addressed to constitutive panpsychists. In these chapters I treat consciousness as a fundamental property, something not explained in terms of anything else but instead baked into the base level of our universe. This means that the most important component subjects will be the microscopic physical parts that make up everything, and in chapter 4 I will consider certain bridging problems arising from the sheer number of microscopic parts we each have.

Second, chapters 5 and 6 present functionalist combinationism, an account of mental combination addressed to philosophers and cognitive scientists interested **(p.33)** in the consciousness of intelligent information-processing systems. Whereas panpsychist combinationism is intended as an account of fundamental reality, functionalist combinationism is not: it is a theory of how the processing of information enables the combination of simpler minds into more and more complex and intelligent ones. Chapter 6 applies functionalist combinationism to a number of real and hypothetical examples, sketching out a

compositional explanation of the structured consciousness associated with brains (whole ones, half ones, and split ones) and with social organizations.

Finally, chapters 7 and 8 present psychological combinationism, an account of mental combination focused on the special concerns of the most complex and sophisticated sorts of subjects, namely “persons” like ourselves. Persons are not just intelligent subjects; they are subjects with a capacity for rational agency and self-consciousness, and here I will examine bridging problems concerning these capacities, as well as the ways that these capacities can break down into inner conflict or dissociation. Finally, in chapter 8 I draw together the ideas of the preceding chapters through a detailed consideration of a particularly challenging thought experiment, that of two persons fusing into one.

Although these chapters contain multiple theories, these theories are not really rivals; they do not inherently conflict with one another, though they are offered in part to allow different readers with conflicting prior beliefs about consciousness to still accept combinationism. They are instead best seen as aimed at different levels of reality: panpsychist combinationism aims at the fundamental level of nature, functionalist combinationism aims at the level of information-processing systems, and psychological combinationism aims at the level of people, with their agency, their problems, and their self-conceptions. Personally, I think all three theories are true, because consciousness is present at all levels, in interestingly different forms, and all of those forms, though in different ways, can combine.

Notes:

(1) Of course, we need to make sure to consider the parts *in their specific relations* to each other: a rock may take up more space than all of its component atoms would, if they were arranged differently. But that is simply to say that the rock takes up more space as it actually is than it would if it were restructured somehow.

(2) The metaphor of “levels” in nature, which I use here and throughout the book, can be defined in a variety of ways (see, e.g., Wimsatt 1976; Marr 1982; Bechtel 1994; Rosenberg 1994; Craver 2015). I use it to mean a set of entities of similar sizes, interacting in distinctive ways that might be studied by a specific science (as sociology studies group-level goings-on, psychology studies human-being-level goings-on, and neurobiology studies neuron-level goings-on) and typically but not always composed of “lower level” entities and composing “higher level entities.”

(3) Here and below I use the phrase “real relation” just to exclude a certain sort of trivialization-by-creative-definition. For instance, consider the relation “composing-a-composite-subject-together-with.” From the fact that two minds stand in this relation to one another, it follows *a priori* that there is a composite

subject they compose. Yet clearly this has not really achieved anything. A “real relation” is one that, unlike “composing-a-composite-subject-together-with,” actually pertains to the relata themselves, so that our understanding of it is not parasitic upon our understanding of what it is meant to explain.

(4) What if P is entirely grounded in, and explained by, just *one* of the *p*s? This arguably satisfies the definition of combination given, but there is clearly a sense in which no actual “combining” is going on. I will call a case like this “trivial combination”: combination that is not trivial is “substantive.” For brevity, all references to “combination” should be read as meaning “substantive combination” unless otherwise specified.

(5) This kind of emergence is sometimes called “strong emergence,” to distinguish it from “weak emergence,” where the whole’s property is, in principle, explained by and grounded in some properties of its parts, but cannot easily, or in practice, be deduced from them (see Bedau 1997; Chalmers 2006b; Wilson 2010; Seager 2017). I am unsure what to call properties that are grounded but not explained, or vice versa.

(6) Other users of this style of argument include Proclus (1963, 163), Avicenna (1952, 47ff.), Descartes (1985, 2:59), Butler (1860), Mendelssohn (2002), Clarke (1978, 3:759), Bayle (1991, 128–134), and Lotze (1894, 158). See also Mijuskovic (1984) and Lennon and Stainton (2008).

(7) While Block maintains that the compositional aspect of these cases is irrelevant to their force, others disagree: David Barnett (2008, 309) has argued that the best explanation for these intuitive judgments is that “our naïve conception of a conscious being demands that conscious beings be simple” (cf. Montero 2017, 225; Coleman 2017, 259n31).

(8) I will use the term “person” for a certain sort of conscious subject, one possessing intelligence, rationality, and self-consciousness comparable to our own; what counts as a “person” will thus be as vague as what counts as “comparable to.” An adult human or an intelligent alien (Vulcan, Kryptonian, or similar) is a person; a dog or a baby is not.

(9) Versions of the same “too-many-minds” objection are also employed in debates over personal persistence, targeting theories on which “we” (persons) are distinct from but share *all* our parts with some other entities (organisms, or bodies).

(10) This is a highly abbreviated version of the “hard problem of consciousness,” related to the “explanatory gap” (Levine 1983; Chalmers 1995; cf. Nagel 1986), supported by the “conceivability argument” (Chalmers 1996; cf. Kripke 1980; Descartes 1985) the “knowledge argument” (Jackson 1982), and the “structure-and-dynamics” argument (Chalmers 2003a), all of which have been debated and

defended at length by others much better than I can do here (e.g., Seager 1995; Lewis 1990; Chalmers 2002, 2009; Dennett 2006).

(11) I use “fundamental” to mean “not grounded in anything else,” in the sense of “grounded” explained in section 1.3.

(12) If an explicit definition is needed, a good first pass is that a “structural property” is either a relational property, or an intrinsic property which consists entirely in relations among a thing’s parts. It contrasts with “absolutely intrinsic”; I take these definitions from Pereboom 2011, 93ff.

(13) Several subtly different distinctions have been drawn here, such as “constitutive” and “emergent” (Chalmers 2015; Goff 2017a, 155), “reductive” and “emergentist” (Goff 2010), “constitutive” and “nonconstitutive” (Mørch 2013), or “constitutive” and “causal” (Mørch 2014). Sometimes this distinction is drawn between different sorts of emergence, such as “weak” and “strong” (Chalmers 2006b), “conservative” and “radical” (Seager 2017), or “weak,” “strong,” and “brute” (Mørch 2014). Adam Pautz, in unpublished work, also distinguishes “reductive” and “primitivist” versions within the constitutive camp. See also Roelofs 2017b.

(14) This requires, but goes beyond, “supervenience”: A supervenes on B if any world with B also has A (“no change in A without a change in B”). But supervenience is not enough: if two independent things were both necessitated by the same thing and could not exist otherwise, they would supervene on each other, but neither would ground the other (cf. Schiffer 1987; Horgan 1993; Wilson 2005).

(15) I use “discrete” to mean “non-overlapping,” “sharing no parts.” This is stronger than “distinct,” which I use to mean simply “nonidentical,” “not the very same thing.”

(16) I don’t know how to rigorously spell out “same basic category,” but I can gesture at what seems important. The physical parts of a table can be sensibly compared to the table—we can ask which is bigger, whether they are the same color, what the ratio of their masses is, and so on. We can say how the table leg would have to change to resemble the table more—how it would have to grow, change shape, etc. But it feels like a joke to compare the table with one of its properties—to ask whether it or its shape or its color is larger, or heavier, or harder. And clearly it makes no sense at all to talk about turning one into the other—there is nothing that could be done to a property to “make it into” an object, or vice versa.

(17) Of course two parts of something may overlap. But in this case we can always identify both a part that they have in common and discrete parts of each that are not shared. For instance, a table has as parts a north-facing half and an

east-facing half; these are not existentially independent because they share a quadrant. But we can identify that quadrant (the northeast) as another part of the table, and also identify the other, discrete, parts that the two halves do not share (namely the northwest and southeast quadrants). To do this with the table-as-square and the table-as-wooden would require inventing odd entities to be the shared part (“the table without any features”) and the unshared parts (“being-square,” “being-wooden”). But these would then violate the second condition.

(18) Note that aggregates do not lack structure (the aggregate of all the parts of my car is, currently, just as structured as my car itself); they are simply not defined by that structure (the aggregate of all the parts of my car could still exist if the parts were disassembled). In this respect I use the term “aggregate” differently from, e.g., Coleman 2012, 139–142, and Morris 2017, 117–120, who apply it only to unstructured wholes like heaps of sand.

(19) Strictly, the historically most influential version of this idea is not naturally put in terms of “emergence,” since the special unit-making power (sometimes called an “entelechy,” “substantial form,” or “soul”) was not typically thought of as “produced by” simpler things coming together. More likely, it was either divinely bestowed from above or passed on by other true units with the power to “in-form” matter and produce further true units. For my purposes it is harmless to call it “emergent,” since I use that term just to mean what is not a mere combination.

(20) Maybe aggregates don’t “really exist”: our talk of them is just a linguistic shortcut for talking about their parts all together. Maybe structure-specific wholes don’t “really exist”: our talk of them is just a linguistic shortcut for talking about how aggregates behave when they are structured a certain way (“car” is just what we call a collection of mechanical parts *when they are connected*). What does it even mean to “really exist,” as opposed to just “exist”? What kinds of questions are these? (Cf. Lewis 1991, 81; Yi 1999; Fine 2001; Wilson 2014.) Into these thickets I shall not venture.

(21) Philosophers disagree about how much of mental life involves this phenomenological sense of “consciousness”—whether it is only sensations and emotions which have “something it is like” to undergo them, or whether thoughts, beliefs, plans, judgments, intentions, etc. do too. They also debate whether any phenomenology associated with thoughts, etc. is distinctively cognitive or just reheated imagery (Prinz 2011; Horgan and Tienson 2002; Kriegel 2007; Bayne and Montague 2011). I will tend to speak in line with the more “liberal” (Bayne 2010, 5–7) position, that many different mental states have phenomenology, but nothing I say will depend on this.

(22) This is narrower than simply “properties which it is like something to have”: if I can desire X both consciously and unconsciously, then there is sometimes

something it is like to have the property “desires X,” but what it is like does not define that property. Cf. Bayne and Chalmers 2003, 30–31, on “phenomenal states” versus “phenomenally conscious mental states.”

(23) Similar things can be said for divisions of the mind into “faculties” or “drives” or anything else that is not explicitly itself a conscious subject. I may divide my mind into will, imagination, and intellect, but that does not commit me to thinking that there is anything it is like for my imagination itself to imagine things.