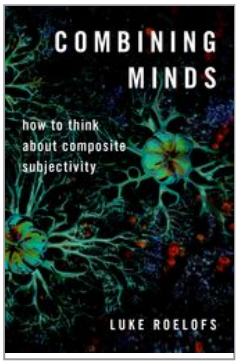

Combining Minds: How to Think about Composite Subjectivity
Luke Roelofs

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

Composite Subjectivity and Intelligent Subjects

Luke Roelofs

DOI:10.1093/oso/9780190859053.003.0005

Abstract and Keywords

This chapter is about how to combine subjects of experience understood in functionalist terms, as systems whose consciousness comes from having a functional structure which supports intelligent behavior. This requires examining how composition relates to the key features of such subjects, including not just their functional structure but the structure of their consciousness, and the systematic coherence between these two structures. The chapter argues that information-integrating interactions are key to connecting the conscious structure and functional structure of the parts, so that they form a whole with even richer structure. This integration can take many forms, even including the social interactions of cooperating subjects in a social group.

Keywords: functionalism, conscious structure, subject of experience, personal identity, information integration, representation

THE HUMAN BRAIN is often called the most complex object in the universe. In this and the next chapter I develop a theory of the role this complexity plays in mental combination. I have called this theory “functionalist combinationism” because it focuses particularly on “functional structure,” the set of ways that different states of a system are caused by, and cause, other states of that system. One popular view in the philosophy of mind is that consciousness itself, and indeed all mental phenomena, are ultimately explained by functional structure; call this “pure functionalism” (e.g., Putnam 2003; Dennett 1991; Shoemaker 2000, 2003). According to pure functionalism, the way to get minds is simply to take some material, which may be in itself completely mindless, and organize it

to implement the right functional structure, to work the right way. Readers attracted to pure functionalism should take functionalist combinationism as an independent and complete theory of mental combination, a rival to and replacement of the panpsychist combinationism of chapters 3 and 4. (Call this version of the theory "pure functionalist combinationism"). Pure functionalism implies that chapters 3 and 4, which treat consciousness as a fundamental property, were misguided right from the start.

(p.150) But many philosophers who are not pure functionalists still think that functional structure is of vital importance in a full understanding of the mind, even if it is not the whole story: call this "impure functionalism" (Lewis 1980; Chalmers 1996, 274ff.). In particular, I think readers attracted to panpsychist combinationism would do well to also accept functionalist combinationism as a supplementary theory, a theory specifically dealing with how the sort of consciousness possessed by human beings and other animals differs from but depends on the more basic consciousness that is fundamental and pervasive in nature. (Call this version "functionalist-panpsychist combinationism.") In this chapter I will lay out functionalist combinationism in a form that is, as much as possible, neutral between pure and impure functionalism, but will at times note how the theory's implications differ between the two.

Like in chapter 3, and unlike in chapter 7, I will think of subjects as substrates of experience rather than personas; that is, I identify the subject of some experiences with the system within which they arise, rather than with any sort of construct out of those experiences. But unlike in chapter 3, and like in chapter 7, I will think of composite subjects as structure-specific wholes, not as aggregates. Subjects which are made of parts are not just those specific parts, considered as one; they are defined by the particular way their parts are put together, and exist only as long as those relations are maintained.

To put a name to this crucial "way their parts are put together," I will use the term "intelligent subject" for conscious beings with the kind of complex functional organization displayed by humans and many other animals. The aim of functionalist combinationism is to understand how there could be intelligent subjects composed of other intelligent subjects, and how in such cases we could relate facts about consciousness in the whole to facts about consciousness in the parts. The next chapter examines four case studies where intelligent subjects seem to combine into other intelligent subjects. The entities concerned include biological organisms, organs, and parts of organs (i.e., human beings, human brains, and subsystems within human brains), as well as social groups made up of such organisms; functionalist combinationism is also intended to cover artificial conscious systems like hypothetical future AIs or cyborgs, and conscious nonhuman organisms, although I do not discuss any such cases in detail.

Panpsychist combinationists have a special reason for interest in functional structure. As well as understanding how intelligent subjects can compose other intelligent subjects, they need to explain how intelligent subjects could be composed by the simpler, *un*intelligent subjects they postulate. This challenge is particularly urgent because panpsychist combinationism implies such a widespread distribution of both subjecthood and phenomenal unity. By defending experience inheritance, the **(p.151)** combinationist may solve the subject-summing problem, but they seem to obliterate the distinction between physical systems which are, and which are not, subjects of experience. Yet our notion of what it means to be a subject of experience is surely partly defined by the fact that it applies to human beings, some animals, and perhaps other creatures (aliens, spirits, robots) that operate in similar ways, but does not apply to clouds, ovens, or trees, and certainly not to arbitrary aggregates like "all the toasters in Norway." Similarly, considerations of elegance and nonarbitrariness may support the view that phenomenal unity holds among all causally connected sets of experience, but this seems to violate the very contrast—between how one person's many experiences are related and how different people's experiences are related—that initially prompted us to formulate the idea of phenomenal unity.

So the functionalist-panpsychist combinationist, having accepted the metaphysical foundations laid out in chapter 3, is under a special pressure to show that our intuitive distinctions can be "reconstructed" upon that counterintuitive, but theoretically elegant, foundation. This is part of a general difficulty for panpsychists that chapter 4 called the "mismatch problem": the world as physics sees it looks to be structured along very different lines from human consciousness. In chapter 4 I offered the informational structure hypothesis as a schematic solution to this problem:

> **Informational structure hypothesis (ISH):** The overall structure manifest in human consciousness corresponds to the structure of information-processing in the human brain, not to its gross physical structure.

For panpsychists, functionalist combinationism serves to flesh out and substantiate the ISH, to explain how we get human consciousness out of the vast expanse of rudimentary consciousness postulated by panpsychist combinationism.

5.1. Defining Intelligent Subjects

Functionalist combinationism focuses on "intelligent subjects," a type of conscious system that is more recognizable as a subject than the panpsychist's microsubjects. In order to have an at least somewhat definite idea of what systems these are, and what is required from a compositional explanation of one, I offer the following rough and tentative definition:

> An intelligent subject $=_{def}$ A conscious subject whose consciousness is structured and whose experiences play the functional roles characteristic of **(p.152)** intelligence in virtue of their phenomenal character, conscious structure, and representational content.

This definition is intended to actually apply to all the sorts of creatures which we are intuitively inclined to regard as conscious subjects (human beings and some or most animals), to not apply to all the sorts of creatures which we are intuitively inclined to regard as definitely not conscious subjects (most inanimate objects), and to be indeterminate in application to the sorts of creatures which we are intuitively uncertain whether to regard as conscious subjects (e.g., snails, worms). Pure functionalist combinationists can fully endorse these intuitions and regard “intelligent subject” as equivalent to “subject”: there are no nonintelligent subjects. Functionalist-panpsychist combinationists will reject these intuitions as applied to consciousness per se, because they regard many nonintelligent things (indeed, all material things in the universe) as conscious subjects. But they can still explain and vindicate these intuitions if they can explain why only some systems are intelligent subjects.

The above definition involves four key components: phenomenal consciousness, consciousness being “structured,” the functional roles characteristic of intelligence, and experiences playing roles in virtue of their character, structure, and content. Call these “consciousness,” “structured consciousness,” “intelligent functioning,” and “coherence” between consciousness and functioning. Pure functionalists hold, in essence, that intelligent functioning is all there is to explain; consciousness and structured consciousness are themselves explained by intelligent functioning, and this guarantees their coherence with it. Thus by their lights the definition given above is somewhat redundant. Because of the primary importance of intelligent functioning, I will spend most of this section elaborating and exploring this notion; in the next section I ask what it would take to ground and explain some system’s being an intelligent subject. But first let me say a little to clarify my notions of “structured consciousness” and “coherence.”

5.1.1. Structured and Coherent Consciousness

I draw the term “coherence” from Chalmers’s (1996, 220) discussion of “the coherence between consciousness and awareness,” where “awareness” means essentially “access-consciousness,” i.e., the “psychological correlate of [phenomenal] consciousness . . . wherein some information is directly accessible, and available for the deliberate control of behavior and for verbal report.”

(p.153)

> **Coherence:** Any state of a subject which is phenomenally conscious will also be access-conscious, and vice versa.

As Chalmers observes, general conformity to coherence is a striking feature of everyday experience for creatures like us: the information that guides our behavior and internal processing is generally the same as the content of our conscious experience, even if exceptions are possible. (For possible exceptions or qualifications, see Chalmers 1996, 221–229.) Thus I have defined intelligent subjects in such a way as to make it partly definitional of them: their experiences must play certain functional roles in virtue of their experiential features.

Note that coherence presupposes a certain sort of consciousness: consciousness comprising multiple distinguishable experiences which differ from each other but interact in sufficiently varied ways to play functional roles, including the role of being access-conscious. This is what I mean by saying that intelligent subjects must have "structured consciousness": if they simply undergo one big undifferentiated blur, or a million distinct but qualitatively identical experiences, or anything like that, then coherence cannot hold them and they cannot be intelligent subjects.

This still does not tell us exactly what it is for consciousness to be "structured." A first attempt might be that consciousness is structured when it is divisible into parts which are then related to one another. But "divisible into parts" is importantly ambiguous. In chapter 4 I argued that we should not assume that just because an experience is metaphysically divisible (i.e., it has elements grounded in different substrates) it will display "phenomenal contrast" (i.e., be experienced by its subject as presenting distinct elements). What I called "phenomenal blending" involved metaphysically divisible experiences that lack phenomenal contrast, where different qualitative elements are seamlessly incorporated into a new quality. For consciousness to be structured requires not metaphysical distinctness but phenomenal contrast; thus a subject whose experiences are all phenomenally blended together lacks structured consciousness.

Moreover, "related to one another" is too weak a criterion. (All my experiences bear to one another the relation "occurring later than wooly mammoths who lived earlier than," but that is not part of the structure of my consciousness.) The sorts of relations that matter are those that make a phenomenal difference to the subject, that are manifest in consciousness. Most obviously, they include two of the "unity" relations distinguished in chapter 2, phenomenal unity and representational unity, but they may well include others. I do not think I can offer a complete account of what these relations are, beyond encouraging the reader to consider their own present experience, and the various ways that its elements hang **(p.154)** together. Of course not all intelligent subjects need have consciousness structured in exactly the way that human consciousness is structured; they need only have enough conscious structure for their consciousness to be capable of coherence with intelligent functioning.

5.1.2. Defining Intelligent Functioning

But what exactly is intelligent functioning? Let me first note, and set aside, the question of how exactly “functional roles” and “functional structure” can be rigorously described. Early functionalists were much enamored with computational formulations, which characterized functional structures as something like computer programs (e.g., Turing 1950; Putnam 2003). But other functionalists have objected to this approach, often preferring models based more closely on the structure of the brain (e.g., Smolensky 1987), or explicitly saying that our ways of formulating functional descriptions will have to evolve in tandem with our understanding of the workings of the mind (cf. Block 1992, 71–74). I have no stake in these debates; what I will say about functional roles is rough enough not to depend on how exactly they are defined.

Whatever functional roles are, which are the ones characteristic of intelligence? I will follow tradition in being somewhat vague on this score. These roles are the sort of thing that we all know on some level—they are what we are noticing and responding to when some creature “seems conscious” to us. But, like the rules of grammar, they are extremely hard to articulate. I will just gesture at a few examples; note that in each case it is required that the interactions in question be stable and reliable across time:

- Intelligent subjects have a set of states (which we might call “beliefs”) which interact with each other so as to ensure that they remain generally “consistent.”
- Intelligent subjects have states of two kinds (which we might call “desires” and “beliefs”) and a set of capacities for action which interact with each other so that whenever the beliefs entail that an action would satisfy a desire, and only then, that action is likely to be taken.
- Intelligent subjects have mechanisms (which we might call “decision-making”) that connect all tendencies toward action (driven by different desires) and prevent multiple conflicting actions from being taken by prioritizing some desires over others.
- Intelligent subjects have a set of states (which we might call “pleasures and displeasures” or “happiness and sadness”) which are set up so as to **(p.155)** automatically give rise to desires either to pursue or to avoid them, and which tend to be produced when it is believed that a desire has been met or frustrated.
- Intelligent subjects have a set of states (which we might call “sensations”) which correlate well with conditions in the external world, which are largely independent of their desires and decision-making, and which tend to influence the formation of new beliefs.

Obviously many refinements, clarifications, and additions are necessary for a full account, but the above is sufficient for my purposes.[1]

Note that to the extent that a system functions intelligently, we can usefully and meaningfully talk about that system's "perspective" or "point of view." Because there are states playing the role of sensations and beliefs, we can use them to characterize "how the world seems" to the system; because there are states playing the role of desires and pleasures, we can use them to characterize "what is important" in that apparent world; because there are a specific set of actions available under a decision-making mechanism, we can make sense of asking what the system "should do," what actions are "called for" by important features of the apparent world.

I point out the special applicability of the term "perspective" to intelligent subjects because I suspect it plays a role in our intuitive willingness to ascribe consciousness to them. Subjects with a perspective in this sense are ones which we can imagine being, and moreover ones whose behavior we can usefully predict by imagining being them (for further discussion see Buchanan and Roelofs 2018). That is, only with intelligent subjects does imaginative simulation start to get a grip, and it seems likely that our impression of a thing's consciousness depends importantly on our ability to perform such simulations—to put ourselves "in its shoes."

5.1.3. The Vagueness of Intelligent Functioning

A key feature of intelligent functioning is that it is *vague*; that is, it admits of borderline cases which neither clearly qualify as intelligent functioning, nor clearly fail to. The reason it is vague is that none of our states plays its functional role infallibly: sometimes I fail to pursue my desires given my beliefs, sometimes my perceptions are biased by my desires, sometimes all my representations fail to be **(p.156)** consistent with one another, and so on. Since we are still intelligent subjects, the phrase "play its role" in the above definition must be read as "usually plays its role." But what counts as "usually"? After all, in some pathological cases people seem to suffer from major failures of intelligent functioning on a daily basis (in psychosis, delusion, hallucination, etc.). But at the same time, there must be some minimum reliability, for otherwise almost anything could be said to be "unreliably intelligent": we could interpret any system's internal states as being "desires" and "beliefs" of various sorts, and as long as it "acted appropriately" on at least one occasion we could then chalk up the ninety-nine other occasions to the system's fallibility. Somewhere in between the ideal of 100% reliability that we never reach and the absurdity of 1% reliability, it begins to be reasonable to call a system intelligent, but plausibly there is no sharp boundary.

There is also vagueness that derives from a requirement of complexity. Some people might have more desires, or more beliefs, or more capacities for sensation than others; this does not mean that they are intelligent subjects and the others are not. But again, there must be some minimum, or everything will count. Consider a rock which tends to fall downward when not supported. We could interpret it as having the desire to reach the earth's center, the belief that moving downward will accomplish this, and very simple capacities for action (move downward) and perception (detect which direction is down). But surely we do not want to count it as an intelligent subject just because its functioning conforms to the above principles when interpreted in these terms. (After all, we could equally well interpret it as desiring to rest on some hard surface and moving downward in desperate hope of finding one.) So some minimum degree of complication seems necessary, but again there seems no way to draw a sharp boundary.

I regard this vagueness in what counts as an intelligent subject as a welcome implication, for it provides the most plausible way of handling a broad range of beings which fall in between what intuition would judge "clearly conscious" and "clearly unconscious"—creatures like worms, insects, jellyfish, plants, and even cells. While there is plenty of scope for genuine discoveries about how such creatures operate, even with complete knowledge of their physiology and behavior it may be impossible to definitively put each of them either in the same category as rocks or in the same category as monkeys. And it doesn't seem plausible that some particular small enhancement or quirk of neurology should "turn on the lights"—that one species of worm, say, should be strictly nonconscious and another, almost identical species be dimly conscious. Thus it seems the most reasonable thing is to admit that for some or all of these creatures, it is semantically indeterminate how to categorize them, just as there are men whose heads **(p.157)** are borderline cases of baldness, neither definitely bald nor definitely not bald but somewhere in between.[2]

(The functionalist-panpsychist combinationist, of course, will here rehearse the arguments from section 3.3 of chapter 3, against the idea that worms and their ilk are borderline cases of phenomenal consciousness per se, for [they will say] consciousness does not admit of analysis into some vague degree on an underlying spectrum. Consciousness is a fundamental property, so either there is something it is like to be X, or there is not. So functionalist-panpsychist combinationists should take a somewhat nuanced view here: everything is conscious, including rocks and worms and monkeys and us, but not all conscious things function intelligently. We and monkeys do, while rocks don't. And because intelligent functioning can be analyzed into various matters of degree [viz., complexity and reliability], there will be intermediate cases which could equally well be counted as the very feeblest sort of intelligent subject, or as almost an intelligent subject but not quite. There is no sharp boundary: the clear but imperfect coherence of a human perspective shades imperceptibly into the

useless, inert, confusion of what-it-is-like-to-be-a-rock, and in between goes through stages that we have no decisive reason to classify on the one side or the other.)

5.1.4. Internal and External Functioning, and Blockheads

A final important note about intelligent functioning is that we can distinguish "external" functioning (what outward responses the system makes to various stimuli) from "internal" functioning (how the system is organized, how its different inner states relate to one another). Our intuitive appreciation of a system's functional structure is, inevitably, primarily driven by its external functioning, since that is what we are most used to seeing. But internal functioning is also important, as we see when we consider a hypothetical system which has the same external functioning as one of us, but a very different, and much simpler, sort of internal functioning. Block (1981) imagines such a being, which has come to be called the "Blockhead." It is a humanoid device controlled by a rudimentary mechanism searching a vast lookup table which contains every possible one-hour-long **(p.158)** intelligent-seeming English conversation. If allowed to search this galaxy-spanning list at superluminal speeds, the device could simulate intelligent conversation (for an hour) merely by finding on its list a "canned" conversation which matches its present one up to the last-received utterance. Here there is intelligent-seeming external functioning without intelligent internal functioning. There are no internal representations that can be combined to form more complex representations, or used for inferences, or stored for later access, or anything like that. There are no goal-states, no process for mediating between conflicting desires or selecting the right means to an end; there is precisely one process, of moving down a list and checking for exact matches. Block thinks, and I agree, that this device is obviously not intelligent; its appearance of intelligence derives from "canned" intelligence, the intelligence exercised by its designers in exhaustively distinguishing the sensible from the nonsensical canned conversations (cf. Jackson 1993).

Functionalist combinationists should not regard creatures like this (let us call them all "Blockheads") as intelligent subjects—they should require both internal and external intelligent functioning. For pure functionalists this makes it natural to say that Blockheads just aren't conscious at all, but panpsychist combinationists cannot say this (since for them, everything is conscious). What I think they should say instead is that while Blockheads have consciousness, they lack structured consciousness and coherence; even if they have some experiences, and behave outwardly like a conscious being, their behavior is not lined up with or governed by particular features of their consciousness, like ours is. Because they lack the right internal structure, what they do is divorced from what it is like to be them, and they are not intelligent subjects.

5.2. Explaining Intelligent Subjects

If intelligent subjects can be composites, they are plausibly structure-specific wholes, not aggregates. They are defined by their parts exhibiting the right features and the right organization, not just by being those parts considered as one. Moreover, the persistence conditions we intuitively ascribe to them seem to make special reference to their structure. My brain and body right now as I write, for instance, are probably made up of less than half the same atoms as they will be by the time you read this, when most of the atoms that now compose them will be spread through the environment as dust and water vapor. So if we thought I was an aggregate of atoms we would seem pushed toward saying that I will then be a widely dispersed cloud of dust and water vapor, and that some new person has taken my place. But we refuse to say this, insisting instead that I continue to exist, and continue to be an intelligent subject, as long as *some* collection of atoms **(p.159)** is continuously instantiating the structure of my brain and body. Thus it looks like we are thinking of me as a structure-specific whole. At each moment, an aggregate of atoms *constitutes* me, but I am not identical to it, since I am constituted by a different aggregate of atoms a moment later. And similar things go for other intelligent subjects, like you, your parents, or your cat.

Combinationists need to show how one intelligent subject could be grounded and explained by others which compose it; functionalist-panpsychist combinationists need moreover to show how intelligent subjects could be grounded and explained by unintelligent microsubjects. So let us first ask: How, in general, does one ground and explain the existence and properties of a structure-specific whole?

5.2.1. Explaining Structure-Specific Wholes in General

Let us start with a nonmental example. Consider a tower made of wooden blocks. All that it takes for this tower to exist is for the wooden blocks to be related a certain way (“arranged towerwise,” i.e., on top of one another). This flows from the “essence” of the tower: towers are essentially tall and thin. This is not itself something to explain; this is just what we mean by “tower.” All that needs to be explained is why the relevant properties are instantiated, and we can explain that without mentioning the tower itself. Because structure-specific wholes are defined by a particular structure, they admit of metaphysical analysis and can be grounded by analysis in whatever accounts for the instantiation of their defining structure.

But note that the tower’s essential property is not “being arranged towerwise,” since it is not thus arranged; rather, it is tower-shaped. And the property of being tower-shaped is not instantiated by any of the blocks. Is this a problem? Of course not, since the tower’s essential property is instantiated by the aggregate of the blocks: it is a corresponding property to their collective property of being arranged towerwise. The something-like-identity between the aggregate and its parts (what in chapter 3 I called SI, the substantive indiscernibility of parts and

aggregate) ensures that the defining property of a tower is instantiated, and the existence of the tower follows. Other properties of the tower, like its color, are likewise explained by the aggregate that constitutes the tower having those properties.

But not all properties of the aggregate are shared by the structure-specific whole. The aggregate, presumably, began to exist when the bricks did—in a factory, made by some workers. But the tower's history is much shorter: it was not made in a factory by workers; it was made on the floor by a child. This history is of course still explained by something about the aggregate, namely the fact that the aggregate came to be tower-shaped on the floor, by the action of a child. But to know which of the properties "offered up" by the aggregate will belong to the **(p.160)** structure-specific whole, we need to attend to its essence, its defining structure. And properties of the aggregate that have nothing to do with this structure, like its history prior to becoming tower-shaped, make no difference to the structure-specific whole. This is why we could not use SI to directly explain the existence and properties of the tower: the tower and the aggregate of blocks that constitutes it at a particular moment are not substantively indiscernible, since there are facts about the aggregate that do not correspond to anything about the tower.

So in general we might say that structure-specific wholes can be explained in two steps: the aggregates that constitute them inherit properties from their parts in virtue of being simply those parts considered as one, and their essence specifies a set of requirements, a "filter," determining both which properties must be instantiated by the aggregates in order for the structure-specific whole to exist, and also which properties of the aggregates are shared by the structure-specific whole and which are not.[3] Moreover, when the structure-specific whole shares properties with the aggregates that constitute it, this is not plausibly regarded as the property being instantiated twice; rather, the two entities share the same token property instance, just as (according to the arguments of chapter 3) the aggregate can share property instances with its parts. For instance, the tower occupies a certain part of space; so does the aggregate; so does one of the blocks, which is part of both; but this part of space is occupied only once, and the instance of the property of occupying it is shared by all three entities.

Suppose we try to apply the same procedure to intelligent subjects, as defined in the previous section. We would need to consider the essential properties of intelligent subjects (which I am taking to be consciousness, intelligent functioning, and structured consciousness, connected so as to secure coherence), and then explain how it comes about that aggregates of matter instantiate those properties.[4]

(p.161) 5.2.2. Explaining Consciousness

Explaining why an aggregate of material parts might instantiate consciousness is, obviously, a big challenge. It is famously difficult to explain why anything at all is conscious, and I will not here attempt to do so. If constitutive panpsychism is true, then things are conscious because they are built out of conscious matter. If pure functionalism is true, then things are conscious because they are organized so as to function intelligently. If theistic dualism is true, then things are conscious because God chose to implant immaterial souls into certain physical bodies. If some other theory of consciousness is true, then its posits explain why things are conscious.

We can, however, ask a more precise question: What role can consciousness in a thing's parts play in grounding and explaining consciousness in that thing? According to panpsychist combinationism, consciousness in the parts very directly grounds and explains consciousness in the whole: wholes (or at least aggregates) share all the experiential properties of their parts. But if pure functionalism is true, then things are more complicated. For the whole to be conscious is for it to exhibit intelligent functioning; for the parts to be conscious is for them to exhibit intelligent functioning. So to see the relationship between consciousness in the whole and in the parts we will need to first consider the relationship between intelligent functioning in the whole and in the parts.

5.2.3. Explaining Intelligent Functioning—Division Relativity

Intelligent functioning, unlike phenomenal consciousness, is in principle very easy to explain. Aggregates inherit the causal powers of their parts (or so I argued in chapter 3), and specific causal structures can then be explained by the relevant theories in physics, chemistry, and biology: intelligent subjects exist because gravity caused planets to coalesce, because carbon can form such long and complex molecules, because evolution selected for greater intelligence in certain environments, and so on. While this part of the explanation is far from easy, it is proceeding apace, and the combinationist can happily rely on it.

What role is played, in this complicated but unmysterious explanation of functional structure, by intelligent functioning in parts of the thing? Certainly it is nothing so simple as automatic inheritance; a whole with parts which function intelligently may not itself function intelligently (consider the solar system and its human parts, for instance). This is because, whereas causal powers are plausibly division-invariant, functional structure is division-relative.

(p.162) Functional structure is division-relative because the functional role played by a particular state is defined in terms of the other states which it interacts with. My beliefs are disposed to mutual-adjustment-to-ensure-consistency with all of my other beliefs, but not with your beliefs. Thus we will

see intelligent functioning when we look at all of my beliefs together, but not when we look at all of yours and mine together.

In the previous section I noted that we could think of intelligent functioning as coming in degrees; we could then say that different levels show different degrees of intelligent functioning. The more you and I share information and cooperate, the closer we come to functioning intelligently as a pair (equivalently, the closer the pair comes to functioning intelligently as a pair). And plausibly I function more intelligently than any of my neurons does considered as a neuron, although perhaps neurons themselves function more intelligently than, say, pebbles, or atoms. Looking at different scales will reveal different degrees of intelligent functioning, some of which will be sufficient to call units on that scale "intelligent subjects" and some of which will not.

5.2.4. Explaining Intelligent Functioning—Sensitivity, Control, and Coordination

In light of this division relativity, intelligent functioning in a thing's parts will contribute to intelligent functioning in the whole only if the relevant parts are "hooked up" to the rest of the whole in such a way that their states can play the same or similar functional roles for the whole as they do for the parts. The simplest such case would be "trivial" combination, where a single part is entirely responsible for the intelligent functioning of the whole. Consider the relationship between a human brain and a human being: the whole functions intelligently largely because the brain, a particular part of it, does so.[5] For example, sometimes the whole human being wants food, detects the presence of pizza nearby using sight, recalls that pizza is food, and walks toward the pizza as a result. When this happens, the brain is performing essentially the same functions: it wants food, detects the presence of pizza nearby using the inputs received from the eyes by the occipital lobe, recalls that pizza is food, and initiates walking toward the pizza by sending signals down the spinal cord as a result. What allows for the functioning of the brain to be shared by the whole human being is that the brain's states both *control* the rest **(p.163)** of the body (there are brain states that can initiate or prevent contractions of the body's muscles) and are *sensitive* to the rest of the body (there are brain states that are reliably caused by stimulation of the body's sense organs). This control and sensitivity allows the brain's intelligent functioning to be simultaneously the intelligent functioning of the whole body.

Control and sensitivity are still key when we move beyond trivial combination and consider wholes with multiple, intelligently functioning parts. To the extent that the states of each part have some measure of control of, and sensitivity to, the rest of the whole, their intelligent functioning belongs also to the whole. Consider here a team of people working together on some cooperative project, like sailing a ship or building a house. Insofar as the perceptions of each are somewhat sensitive to the situation of the whole group, and the voluntary actions of each have some control over what the whole group does, their

individual states can play the functional roles characteristic of intelligence not just for each of them but for all of them.

Control and sensitivity are not quite enough all by themselves, however; there must also be coordination between the different intelligent parts. If they control the whole's behavior but do so independently of one another, for instance, then one is likely to undo what another does, and the whole will display paralysis or self-defeating behavior rather than intelligent functioning. (Consider an army with ten generals who never talk to each other.) The parts must coordinate their activities somehow, whether through a "division of labor" (different parts have primary control of different aspects of the whole's behavior), a "majority vote" (different parts sum or synthesize their respective impulses to determine the whole's behavior), or some other method. We can boil this down to the following principle:

> **Conditional functional inheritance:** If a part of X has an experience which plays a certain functional role, then that experience plays the same or a similar functional role for X to the extent that the part in question is connected to the rest of X such that its experience has sensitivity to, control over, and coordination with other events occurring in X.

The notion of "the same or a similar functional role" is rough and intuitive, like so much else in this chapter. Consider some examples: if a subsystem of my brain works to determine the colors in a visual image and construct a visual color map, then the "map experience" it generates does not play quite the same role in its functioning as it does in mine (e.g., for me it prompts revisions in my beliefs about fruit in my immediate environment, while the subsystem has no beliefs about fruit), but clearly the roles it plays for the subsystem and for me are systematically **(p.164)** connected. Likewise, when someone reports what they think to the other members of their institution to influence an ongoing debate, what is reported does not play quite the same role for them and for the institution: for them perhaps it is a belief, a settled view that they will unhesitatingly act on, while for the group it is one thought among others, perhaps entertained but not yet endorsed. Yet clearly there is something important in common between the two roles. The particular sorts of control, sensitivity, and coordination that a part's experience has will determine how similar the functional roles it plays at the two levels are, as will the roles played by states of the whole's other parts.

5.2.5. Explaining Structured Consciousness and Coherence

Finally, what about explaining the last two components of my definition, structured consciousness and coherence? After all, it is not enough to have both consciousness and intelligent functioning: a subject is an intelligent subject, like one of us, only if its experiences are the states which play the relevant functional roles, in virtue of their experiential features—only if it functions intelligently

because of its experiences. This requires that its experiences be diverse and powerful: it must be capable of many distinct sorts of experiences, which can affect each other differently and can give rise to different behavioral expressions.

Pure functionalists have a ready explanation of coherence: consciousness is nothing over and above the right kind of functioning, so coherence could not fail to be true. But according to panpsychist combinationism, consciousness is in principle separable from any functional notion, including awareness/access-consciousness. Coherence could thus completely fail for some sorts of subjects, even though it is in fact generally true of us, and functionalist-panpsychist combinationists need to give some idea of what explains this. (In section 5.4 I suggest that coherence holds for us because our consciousness gets its structure from *information integration*, which also underlies functional structure.)

What about structured consciousness? As with consciousness itself, I will not attempt the daunting task of providing a general-purpose explanation of how any subject ever comes to have structured consciousness. What I will examine, though, is the more specific question of how a thing's conscious parts can contribute to the structure and coherence of its consciousness. As with intelligent functioning, the easiest case is trivial combination, where a single part does all the work. If a whole has a single conscious part, which controls its overall functioning (as is roughly the case with a human being and their brain), then the structure and coherence of that part's consciousness will also belong to the whole's. The whole inherits all the **(p.165)** experiences of the part, and so its consciousness is structured however the part's consciousness is structured.

But what is much more perplexing is how the whole's consciousness will be structured if it inherits experiences from multiple parts. If a group inherits its members' experiences, or a brain inherits the experiences of its subsystems, how do these "fit together" for the whole? This is, I think, the most challenging question for functionalist combinationism to answer, and so addressing it will occupy the next two sections.

5.3. Structured and Unstructured Consciousness

An intelligent subject with a single conscious part that controls and is sensitive to its overall functioning will have consciousness that is structured the same way as that part's, and which displays the same coherence between consciousness and function. An intelligent subject with no conscious parts will have to get its conscious structure in some noncombinationist way, which goes beyond the scope of this book. But what if an intelligent subject has two or more conscious parts, both of which have structured consciousness that coheres with their own intelligent functioning, and both of which have a significant degree of control over and sensitivity to its overall functioning? Since it inherits their experiences,

the structure of its consciousness must in some way depend on the structure of theirs, but how? That is what this section and the next try to outline.

First, a note on terminology. I said earlier that “structured consciousness” could be understood as “enough conscious structure for consciousness to be capable of coherence with intelligent functioning.” I employed the notion of “conscious structure,” whose relationship to “structured consciousness” should be made clear. In short, conscious structure covers all the relations among elements of consciousness which, when there are enough of them, count as structured consciousness. So unintelligent wholes with intelligent conscious parts, like the solar system, lack structured consciousness but still contain a lot of conscious structure, namely, that exhibited by its various intelligent conscious parts, like me or you.[6]

(p.166) 5.3.1. Explaining Phenomenal Unity

The first question about explaining conscious structure is what explains phenomenal unity, and to this question we have already seen one answer. Panpsychist combinationists, appealing to the arguments of chapter 3, can say that phenomenal unity is irreducible to, and separable from, other relations like access-unity or global consistency. And this irreducible, division-invariant relation pervades the universe: the experiences of any two intelligent subjects are already unified regardless of how exactly they are interacting, but by itself this phenomenal unity does not guarantee representational unity, causal interdependence, or anything like that. What remains to be explained is thus principally representational unity.

Pure functionalist combinationists, by contrast, will want to explain phenomenal unity by reference to the holding of “enough” other relations, in particular functional relations. Phenomenal unity between two experiences means that the composite they form is itself an experience, and according to pure functionalism, whether a state is an experience depends on whether it plays the right functional role. So for pure functionalists, the explanation of phenomenal unity will “piggyback” on the explanation of functional unity given in the previous section, and phenomenal unity will share the division-relativity of functional unity.

So on both versions of functionalist combinationism, the key aspect of conscious structure that still needs to be explained is representational unity, and it is this that the rest of this section focuses on.

5.3.2. The Canvas Model and the Superimposition Model

It will be useful to start by reflecting on two somewhat metaphorical models of how conscious structure might arise from combining multiple structured consciousnesses. Both presuppose phenomenal unity among the combining experiences, since without that there will be no single consciousness that they

combine into, but only a number of distinct and separate consciousnesses in one system. But beyond that assumption they diverge.

One, which is very easy and tempting but ultimately unhelpful, is what I call the "canvas" model. Here we imagine each component experience being "placed" at a certain point in a preexisting space. From my parts I inherit, say, a red-experience and a blue-experience, and so I experience red-here-and-blue-there, like two paint splotches on a canvas. Coleman (2012, 157–158) is the most explicit advocate of this **(p.167)** approach, saying "to take us toward the combination of the panpsychic ultimates . . . the metaphor-model I will appeal to is that of paint patches on a canvas" (157).[7]

The canvas model makes a lot of sense if we assume a correspondence between metaphysical distinctness and phenomenal contrast—between there being two or more elements of our consciousness which are grounded in distinct parts of us, and the experience of being presented with two distinguishable elements. In chapter 4 I identified, and rejected, this assumption: the fact that two elements of my consciousness come from different parts of me does not imply that I will experience them as two distinct things contrasting with each other. On the contrary, I proposed that the default state for a composite experience subsuming two experiential parts grounded in distinct parts of the subject is not phenomenal contrast but phenomenal blending: by default the two subsumed experiences will be present only as contributions to the quality of the composite experience. This proposal, and the rejection of a correspondence between metaphysical distinctness and phenomenal contrast, was crucial to chapter 4's solutions to the palette problem and revelation problem.

A more basic problem with the canvas model is that when we combine paint patches into a picture by putting them at certain locations on our "canvas," a crucial role is played by the canvas itself—but where does the canvas come from? Why does it have particular locations available, arranged in the particular dimensions that it has? This kind of metaphor makes combination seem easy, but only by smuggling something in that does most of the work for us. An adequate model of combination should be one on which the "space" is actually constructed by the components themselves: the paint patches must build the canvas.

For this reason, I prefer to think of the composition of consciousness in terms of the superimposition of layers—less like painting, more like animation using multiple transparent celluloid sheets (cf. Lee N.d., 21–22). Each component, just like the composite, has its own conscious "space," divided and organized in its own way. Insofar as human consciousness is appropriately described as a "phenomenal field," something within which different elements are laid out, some central and some peripheral, some nearer to each other and some further apart, we should regard the component subjects composing humans as having

their own such fields, not as contributing elements which are somehow “fitted into” the whole’s.[8]

(p.168) Philosophers have debated whether human consciousness is “holistic” or “atomistic,” i.e., whether the whole phenomenal field comes first, with particular experiences as merely abstracted aspects of it, or the parts come first, with the whole being merely a bunch of them woven together (Gurwitsch 1964; Searle 2000; Dainton 2010; Bayne 2010; Chudnoff 2013; Koksvik 2014). One effect of accepting the superimposition model in place of the canvas model is that even though combinationists are committed to a metaphysically “atomistic” view where independent component experiences come together to form the phenomenal field, they can combine this with “holism” in the following sense: none (or not all) of the experiential elements that we can introspectively distinguish are metaphysically independent of the whole field. The different “layers,” different component fields, are more fundamental than the total field, but the elements appearing in the total field need not be—they can be mere abstracted aspects, because they each represent intricately combined contributions from many component fields.

5.3.3. Phenomenal Blending and Indeterminacy

This “superimposition” model immediately raises the question of how the layers are “aligned.” Because each component subject has its own conscious field, but there is no preexisting composite field to organize them, there seems to be nothing to decide how the internal structure of one field should map onto that of the others. Does the attentional focus of one component map onto the attentional periphery of another? Does the visual field of one component map onto the auditory experiences of another? Even if two visual fields are lined up, should the top of one map onto the top of the other, or onto its bottom, or onto its left side or center? It seems that unless something can fix the component fields into particular relations, the composite’s field will be underdetermined; we have no reason to suppose it to be one way rather than another, out of all the ways that its component fields might be aligned.

Rather than treating this underdetermination of conscious structure as a problem, however, I propose to embrace it. It is a virtue of the superimposition model that it predicts such underdetermination as the default outcome, because this allows it to make sense of conscious *un*structuredness. In chapter 4 I hypothesized the correlative processes of radical confusion and phenomenal blending, whereby multiple conscious elements appear as an undifferentiated quality that incorporates its ingredients qualitatively without their appearing distinctly. Here I can present the same idea in a new connection: phenomenal blending is what you get when multiple fields of consciousness are superimposed without anything to fix their alignment. In such a situation, composite experience **(p.169)** is equally well envisioned as any of the possible ways to superimpose its component fields, and so is not well described by any of

them alone. Because there is no basis for singling out one of these patterns, the composite's actual experience must be undifferentiated. If functionalist-panpsychist combinationism is true, then the composite does undergo this undifferentiated experience, despite its undifferentiatedness; if pure functionalist combinationism is true, then the composite lacks consciousness altogether, since according to pure functionalism entirely unstructured consciousness is no consciousness at all.

Note in particular that phenomenal blending and phenomenal contrast are, on this account, division-relative. The parts might experience a phenomenal contrast between two or more distinct elements in their conscious fields, but the composite experience of the whole blurs these together so thoroughly that the whole experiences no phenomenal contrast. Or, more precisely, the whole experiences those various contrasts undergone by the parts, but in a confused fashion, so that it cannot discern either.

Experiences often have representational content; they are about something. So it is worth noting what this undifferentiated mode of combination might mean for representation. I think the most natural thing to say is that the content of the consciousness of unstructured composites is "indeterminate" among all the many ways to relate its components' contents, although this demands two qualifications. First, this kind of "indeterminate consciousness" is very different from the status I earlier ascribed to creatures like worms and bacteria, of being "indeterminately" intelligent subjects. In that earlier case, the point was that a representation applied from outside (our concept of an "intelligent subject") is not fully and precisely defined, and so when applied to certain conscious beings will not yield a determinate truth-value. But the point now is that if the conscious state itself has content, that content will not only fail to be fully and precisely defined, but massively fails even to approximate precision.[9]

The other key qualification is that we should not describe the consciousness of unstructured composites as "indeterminate" with respect to its phenomenal character. There is a particular, objective way that it feels to be such a composite **(p.170)** (namely, a blend of all its components' experiences), and to call its phenomenal character indeterminate would seem to deny this. To put it another way, I follow the common opinion among philosophers that indeterminacy is an essentially representational phenomenon, and so can pertain to how we represent something or to how it represents the world, but not to how it feels.

5.3.4. The Role of Phenomenal Binding

In order to get structured consciousness with contrasting elements, instead of simply a blend, something must serve to fix the alignment of the superimposed layers, so that particular elements in one component field are experienced by the whole as connected with particular elements in another. On the canvas model this role is taken by the canvas, the "space" into which component

experiences are placed, but I have rejected any such canvas. So instead what fixes the relations between elements must be some sort of relation between the component fields themselves. We might metaphorically speak of this relation as "pinning" two layers together at a certain point; to have a label for the relation which accomplishes this, I will call it "phenomenal binding" (taking the term from Woodward [2015, 151ff.], drawing on the extensive literature in psychology on the "binding problem"). Phenomenal binding both makes the content of the composite experience more determinate and gives it a "structured" phenomenal character, in which distinct and contrasting elements are available to be distinguished by the subject.

Plausibly neither of these is accomplished all at once, though; for most creatures, indeterminacy in content is somewhat reduced, but only somewhat,[10] and patterns remain somewhat "blurry." When we are tired, disoriented, or in the first half-second of being startled, our experience is blurrier than normal in the sense I have in mind. It is not quite homogeneous, for there is still a sense of many specific possibilities, but our lack of a sense of which one is actual gives it a relatively undifferentiated character. For example, when we are disoriented and do not know which direction is the one we had been heading, it seems to me that our experience of our surroundings blurs together all the live possibilities (that it was *this* direction, that it was *that* direction, etc.) to produce an overall experience which "feels indefinite." The more unstructured a composite is, I suggested, the blurrier **(p.171)** is its experience on average, and thus the closer it is to experiencing a single big blend. And even our most highly structured experiences, when we are maximally alert and confident, confidently monitoring many complex tasks and stimuli, retain some small measure of blurriness "at the edges." When we scrutinize them closely enough, virtually all of our everyday thoughts and words turn out to have some degree of indeterminacy in their content; functionalist combinationism says that this is not surprising, because radical indeterminacy is the default condition of all composite subjects.

5.4. What Is Phenomenal Binding?

But what is this relation I have called "phenomenal binding"? I do not know exactly, and will not try to deduce it from first principles. But if I am right that the default outcome of combining phenomenal fields is an undifferentiated blur, and that no preexisting "canvas" can "slot" them into a structure, then there must be some relation among phenomenal fields which gives rise to conscious structure in the composite experience they form.

Whatever exactly it is, this "phenomenal binding" needs to be considered under three aspects: its objective aspect, how it is manifest externally, detectable by the objective methods of natural science; its within-subject aspect, how it is manifest in the consciousness of the whole, who undergoes multiple phenomenally bound experiences; and its between-subjects aspect, how it is

manifest in the consciousness of each part, who undergoes an experience that is bound with another experience which they do not undergo. In this section I consider what seems plausible to conjecture about these three aspects of phenomenal binding.

Note that these aspects may well not be such that each of them strictly necessitates the others; for instance, whatever the objective aspect is, we might imagine a zombie world, devoid of consciousness, where objectively equivalent physical relations obtain. And whatever the between-subjects aspect, we can imagine a Cartesian world, populated with sharply separate immaterial souls, where individuals have experiences with just the same character as those which, in our world, are phenomenally bound, but which are not bound because of the utter separateness of the Cartesian souls. I do not pretend to have a full grasp of how to think about and relate the different aspects of the phenomenal binding relation whose existence is implied by the theory developed so far, and so the following section is best taken less as the presentation of a completed account and more as an attempt to work through what we can reasonably say about phenomenal binding's three aspects, and to identify and resolve some of the tensions that arise in doing so.

(p.172) 5.4.1. Phenomenal Binding as Integrated Information

We can start with the objective aspect: What observable differences between physical systems might correspond to the presence or absence of phenomenal binding structure? I think the most attractive proposal is that phenomenal binding correlates with *integration of information*.

I am far from the first to suggest that informational structure is crucial to complex consciousness (see, e.g., Baars 1988; Chalmers 1996, 284–292, 2017, 209–210; Gabora 2002); it seems empirically obvious that the processing of information in the brain is systematically connected to the structure of consciousness. That was chapter 4's "informational structure hypothesis," and this link between conscious structure and informational structure seems particularly important to explaining the coherence between conscious structure and functional structure. For me to be access-conscious of something, it is clearly a requirement that information about it be broadcast widely to different subsystems of my mind, so that it can guide various cognitive and behavioral tasks. So to understand why conscious structure is coherent with access-consciousness, and more broadly with intelligent functioning, it makes sense to look at the structure of information.

For the idea of *integrated* information I am specifically indebted to Giulio Tononi's (2004, 2008, 2012; Tononi and Koch 2014; Oizumi et al. 2014) "integrated information theory" (IIT) of consciousness. The key insight is that information as usually defined is unsuitable to account for consciousness, because a system's consciousness is intrinsic to it: each of us experiences what

we do independently of what anyone else thinks about it. Yet the traditional mathematical conception of information (drawn from Shannon 1948) as "differences that make a difference" makes information dependent on an observer, and thus extrinsic. That conception says that one entity carries information about another when knowing the state of the first allows an observer to deduce the particular states of the second, or at least to reduce their uncertainty about it somewhat: more information means a greater reduction in uncertainty. Considering integrated information, however, allows us to characterize a system's *intrinsic* informational properties. A system integrates information to the extent that differences in one part make a difference to the other parts, so that the system is effectively its own observer.

More precisely, information is integrated to the extent that differences in one part not only make a difference to other parts, but make *different* differences from what other differences, or differences in other parts, would make—particularly when these influences run mutually, with each part both affecting and being affected by the others. Tononi nicely illustrates the idea by contrast with cases where there is information but a lack of integration of it. A photograph, for instance, **(p.173)** stores a large amount of information, allowing that information to flow from the photographed object to whoever looks at it. But each point in the photograph is indifferent to the other points, so the information in the photograph is not integrated at all (Tononi 2012, 57–59). Likewise, a pair of books contains the information in the one book and that in the other, but neither book's contents make any difference to the other's, so that although having two books means more information, it does not mean any extra integration of that information. For the information to be integrated would require something like each book's being updated, edited, filled with citations or corrections, based on the content of the other; only then does each page reflect the influence of, and thus integrate information from, both books.

When information is integrated, an investigator examining one part of a system could, in principle, deduce significant things about the state of many other parts; after all, we routinely do this by listening to verbal reports, which reflect activity all across the brain even though they are proximally produced by particular parts of the motor cortex. Moreover, the system acts as its own investigator; it will proceed differently based on these subtly different results, where each particular event both reflects a distinctive set of causes and also produces a distinctive set of effects. Consider in particular that a particular brain event (e.g., that which initiates my voluntarily raising my arm) typically reflects a huge number of other brain events: all the perceptions, emotions, memories, beliefs, desires, etc. that either did make or could have made a difference to whether and how I performed it. Moreover, this decision then feeds back to all those others, potentially making a difference to them.

Information integration underlies intelligent functioning, but it is not the same thing. To function intelligently, a system must integrate at least a fair amount of information (with the exception of “Blockheads,” which have only external, not internal, intelligent functioning), but a system might integrate information without functioning anything like an intelligent subject. A device might be built, for instance, that takes in and organizes vast amounts of information, but has nothing remotely like desires, beliefs, decision-making, perception, etc. If pure functionalism is true, then all this information integration would involve no consciousness of any sort; by contrast, functionalist-panpsychist combinationism implies that such a system would have a richly structured consciousness that was nevertheless utterly alien to ours, and consequently impossible for us to intuitively recognize as conscious (cf. Aaronson 2014; Tononi 2014; Buchanan and Roelofs 2018). On either view, information integration is a necessary but not sufficient condition for intelligent functioning, and thus for intelligent subjecthood. It follows that if information integration is also phenomenal binding —i.e., is also what accounts **(p.174)** for the specific structure of our consciousness—then we should expect a general congruence between functional structure and conscious structure: thus we have (at least the beginnings of) an explanation for coherence.

Tononi attempts a rigorous mathematical quantification of information integration as “phi,” but I will not tie myself to this way of developing the above ideas. Phi is calculated based on the “cause-effect information” of particular parts of a system relative to other parts, where a part has greater cause-effect information when both the set of possible states of the other parts which might have caused it and the set which might result from it are more selective (i.e., differ more from the maximum-entropy set of possible states, where we are maximally uncertain of which state the other parts are in). A system generates more “conceptual information” the more cause-effect information its parts have relative to each other, and phi is then the degree to which a system generates “irreducible” conceptual information, in the sense of generating more than any two parts we divided it into would generate if modeled as independent of each other.[11]

While the attempt at precise quantification is valuable for empirical purposes, there is no need for functionalist combinationism to commit to it. Indeed on a theoretical level it introduces new puzzles about its meaning: it is simply not clear what it means to have a higher or lower “quantity of consciousness” or “degree of conscious structure” (cf. Pautz 2015). How structured is your consciousness right now? Perhaps you can say “very,” and perhaps you can roughly compare it to other occasions, but it would be faintly comic to think there is a precise numerical answer. (Compare questions like “How structured is this car?”)

Another reason for combinationists not to buy IIT wholesale is that it is sharply anti-combinationist, due to its “exclusion postulate,” according to which, within any hierarchy of nested systems and subsystems, only one layer can be conscious (see Tononi 2012, 59–68; Tononi and Koch 2014, 6; Oizumi et al. 2014, 3, 9–10). Whichever entity has the highest value of phi is conscious, and any system which contains it as a part, or is contained within it, lacks consciousness entirely even if it has significant phi. By this principle, a neuron kept alive in a petri dish would be conscious, but neurons in a living human brain are not conscious simply because they are part of a higher-phi system, the brain, whose consciousness excludes theirs. And if a human society ever became sufficiently integrated that its phi value exceeded that of an individual human, all the members of that society would **(p.175)** be instantly but undetectably zombified, deprived of consciousness just by being part of a conscious whole.[12]

For both the above reasons (the lack of need for precise quantification, and the exclusion postulate), I wish to take on what I see as the guiding insight behind IIT, not the theory itself. Functionalist combinationism needs only the key idea that systems which integrate information do so intrinsically, not relative to an observer, because they function as their own observer. Instead of following Tononi’s method of calculating a phi-value and equating it with something like “quantity of consciousness” or “quantity of conscious structure,” I prefer to say simply: each interaction among a whole’s conscious parts, in proportion as it integrates information, effects some degree of some form of phenomenal binding among the phenomenal fields of those parts. This very loose claim obviously leaves a lot of room for further investigation, elaboration, and refinement: after all, functionalist combinationism is more a sketch of a theory than a theory.

Similarly, I depart from Tononi in that, instead of privileging the particular level where a maximum of integrated information might be found, functionalist combinationism treats information integration as determining conscious structure wherever it occurs. This allows me to use “information integration” for something division-invariant, even though Tononi’s notion is very division-relative. In Tononi’s sense, for instance, a pair of noninteracting human beings has zero phi, because it is not a maximum of integrated information: it integrates no information at all compared to one division of it into parts (namely, into person-1 and person-2), since no information would be lost by treating those parts as independent (since the state of one carries no information about the state of the other). But a combinationist need not say that the pair has “zero information integration.” Combinationists need not care about *maxima* specifically, because they have no need to identify one specific level as the right level for consciousness to the exclusion of all others. They can instead point to the fact that as well as one division into parts that carry no information about one another, the pair also has a great many divisions into parts which carry lots of information about one another (e.g., into “left-half-of-person-1+left-half-of-person-2” and “right-half-of-person-1+right-half-of-person-2”). Even though the

two-person system is not integrating **(p.176)** more information than certain of its parts, it is nevertheless a system within which a great deal of information is being integrated. It contains that information integration without being itself a maximum of integrated information in the same way that it contains the conscious structure of its parts without itself having structured consciousness.

5.4.2. Phenomenal Binding as mutual Co-Presentation

So I suggest that the externally observable aspect of phenomenal binding is most likely information integration. But what is trickier to understand is what phenomenal binding means for the component subjects involved—to understand its between-subjects aspect. In particular, here we face in particularly strong form the dilemma posed by the boundary argument and the incompatible contexts argument from chapter 2. That dilemma, recall, was roughly this: either we think of conscious unity as relative to subjects, so that two experiences belonging to different component subjects can be unified relative to the whole, but disunified relative to those two parts ("relative unity"), or else we think of it as a relation among experiences independently of what subjects they belong to, connecting two experiences whether we focus on the composite subject that experiences them both or on the component subjects that experience only one each ("absolute unity").

In the first case we must deny "interdependence" (the idea that being unified by itself changes the phenomenal character of experiences), on pain of a single experience ending up with multiple incompatible phenomenal characters from being unified simultaneously with different sets of experiences relative to different subjects. In the second case we must deny "boundedness" (the idea that my experiences can be unified only with other experiences of mine, not with experiences I do not have).

So far we have seen that different options make more sense for different sorts of unity. Causal interdependence is an absolute relation, so for it combinationists should deny boundedness, but functional unity and global consistency are subject-relative, so for them combinationists should deny interdependence. In chapter 3 I argued that phenomenal unity, if it is taken to be a fundamental relation, should also be absolute, like the fundamental physical relations, and so for it functionalist-panpsychist combinationists should deny boundedness. Pure functionalist combinationists, since they take phenomenal unity to depend on some sort of functional unity, should instead take it as subject-relative and thus deny interdependence.

What about representational unity? Unfortunately, both boundedness and interdependence seem *most* compelling when we are considering representational **(p.177)** unity. When I try to focus on the representational unity between some of my experiences, to identify what it is about their unity that goes beyond their merely being experienced together, what I end up with is

that each experience in some sense “connotes” or “references” the others. What I mean by this is that in undergoing one of them, I am made aware of the presence of the others and how those others are related to it, and when I focus attention on one of them, it already “points me on” to the others as potential foci for attention. Each has the others in its background; by having one experience, the subject is already given an inkling of the others.

For example, when I see a cat, I can distinguish visual experiences of its various parts, can distinguish the colors from the shapes, can distinguish these from the emotional experience of affection it evokes in me, or the conceptual experience of recognizing it as an animal of a certain type. Yet the distinguishing seems in a certain sense artificial: scrutinizing any part for its individual content seems to reveal not a self-contained atom of meaning but things like “affection—*for the thing I see*,” “orangeness—*of that shape there*,” “a leg—*of that body*.” Each contains a reference to the other experiences.

This is why it seems unpalatable to deny interdependence. The problem for the combinationist is that they must either deny interdependence or boundedness, and boundedness is *also* very compelling for representational unity. How can the sort of mutual indication, the sort of internal pointing-to each other, that representationally unified experiences display, be made sense of as a relation that might hold between the experiences of distinct subjects? How can a subject have an experience that stands in this relation to an experience they do not have?

I do not think, however, that these two intuitively plausible principles are actually on a par. I think that the intuitive appeal of boundedness is largely illusory, and that reflection on certain common experiences can undermine it very effectively. By contrast, I do not see an equally good way to deny interdependence: even if not all pairs of representationally unified experiences exhibit the sort of “interpenetration” that interdependence claims, it seems that many do. For this reason, I think combinationists should deny boundedness for representational unity.

The appeal of boundedness is that it can at first seem hard to make sense of one subject’s experiences reflecting the specific character of some other experiences in the way described just above, without it following that those other experiences belonged to that subject. This would seem to require that the experiences of the one subject convey to them, just in being experienced, the presence of the other experiences, but do so without actually providing the full character of those other experiences. This would seem to require that the first experience have, baked right into it, not just the positive sense that the other experience exists and has certain **(p.178)** features, but also the negative sense that the other experience has other features, that its character goes beyond what the first experience conveys. There must be a sort of “negative”

phenomenology, a sense of the other experience as having a nature that is *not* fully captured in the first experience's indicating it.

Fortunately, this sort of experience—being aware of something, and thereby aware of the presence of other things that we are *not* aware of—is actually a pervasive feature of everyday human consciousness. It is most obviously present in visual perception: our visual experience of a three-dimensional object is always an experience of it *as* having concealed sides and a concealed interior that we cannot presently see (Clarke 1965; Noë 2005; Nanay 2010; Briscoe 2011). It is also present, I believe, in mind-perception: our immediate awareness of another person's expressive behavior is an experience of that behavior *as* expressing something that is not fully revealed to us, namely their mental state (Smith 2010, 2015; Church 2016; Roelofs 2018; Chudnoff 2018; cf. Krueger 2012). And similar phenomenology is arguably present also in cognitive experience (Jorba 2016) and in the experience of trying to remember something (James 1890, 251ff.). This intermingling of positive and negative, whereby we are aware of one thing as continuous with something else we are aware of not being aware of, and thus as indicating that something else without fully revealing it, has been given a plethora of labels: "adumbration," "amodal perception," "horizon(tality)." But I will adopt the term "co-presentation" (see Michotte et al. 1991; Husserl 1970, 1982, 2001; Merleau-Ponty 1962; cf. Kelly 2004).

I think mutual co-presentation is the best tool for letting us deny boundedness while accommodating the intuitions that support it, allowing representationally unified experiences to convey each other's presence to their subjects without thereby fully providing each other's character.

To see more clearly how this would work, consider the structure of an ordinary case of perceptual co-presentation, e.g., me looking at a table. Here the front of the table co-presents the back: that is, I have an experience which directly presents one thing (the front) as part of a larger whole (the table), which is presented as having aspects that are not directly presented (the back). The proposal defended in different forms by myself, Smith, Church, and Chudnoff is that when I interact with someone and, as we might say, "read" their mental states in their bodily actions, I likewise have an experience which directly presents one thing (their bodily movements) as part of a larger whole (the process of them having a mental state and expressing it), which is presented as having aspects that are not directly presented (the mental state itself). The suggestion I am making now is that representationally unified experiences each directly present themselves, to their **(p.179)** subject, as parts of a larger whole (the composite experience), which is presented as having aspects that are not directly presented (the other experience).[13]

Co-presentation can be more or less informative about what is co-presented, and it seems natural to relate this to information integration. So let us suppose that the more information is integrated between two component subjects, the more specific is the “something more” which each experiences some element of their own experience as continuous with. This is like the progression from seeing a distant black blob as “the visible portion of some object” to seeing it as “the head of some creature”—independently of what is actually revealed (still a black blob), it comes to tell us more about what is concealed (that it is a creature with a head) through a richer understanding of how the two relate (as head-to-creature, not just part-to-whole).

For instance, to have an experience representing P-or-Q might perhaps involve having one experience representing P and another representing Q, each of which co-presents the other as “an alternative.” The experience representing P need not be so informative as to indicate *what* the other represents; it might only demonstratively refer to it as “that alternative to P,” just as my visual experience might not indicate what the rear side of the coffee cup looks like, but only that it is “the opposite side of what I directly see.” An experience can be informative by indicating only how what it co-presents relates to what it presents.

Which forms of informative co-presentation, among which simpler contents, correspond to which representational relations, is obviously a huge topic. But I am not trying to give a systematic account of human representational capacities, only to sketch out how combinationists might think about them.[14] The key thing is that informative co-presentation allows a subject to incorporate references to another’s experiential content into their own, without fully grasping or deploying that content themselves. We might use another metaphor, namely that of citation: when one text cites another, its content is enriched without what does that enriching being actually in the text. Many authors citing each other may lack the skills or even the concepts to understand what the others do, but can still jointly create an intellectual edifice richer than any could create alone.

(p.180) 5.4.3. Phenomenal Binding as Representational Unity

So the objective aspect of phenomenal binding seems likely to be something like information integration, and the between-subjects aspect seems likely to be something like mutual co-presentation. Finally, the within-subject aspect is just the familiar relation of representational unity. That is, we already know something about phenomenal binding for the whole, because we enjoy structured consciousness. Of course there are many ways that experience can have structure, and so there must be many variants of phenomenal binding. This is not surprising, given how we have characterized it: it connects elements in different phenomenal fields to one another, and so may take on as many different particular forms as there are different kinds of elements in phenomenal fields.

But the most prominent sort of structure exhibited by human consciousness is representational unity of various forms. In vision, for instance, there has been extensive empirical investigation of “feature binding,” which relates our independently processed impressions of color, shape, and motion.[15] In linguistic thought there is the link that connects predicates to their subjects, and clauses in complex sentences to each other.[16] In both cases we see a relation which welds two representations, of “something F” and “something G,” into a representation of “something F and G.” It seems likely that the particular forms it takes (e.g., visual feature-binding and predication) reflect its conditioning by different complex systems (the human visual and linguistic systems); in simpler subjects it presumably takes a less sophisticated form, though we should not expect to be able to form a positive idea of its specific features.[17]

It is very plausible that representational unity presupposes phenomenal unity: two experiences cannot together form an experience with complex representational content if they do not even form an experience. This means that what combinationists say about phenomenal unity will ramify into what they say about representational unity, and thus into what they say about phenomenal binding.

(p.181) Consider first the (somewhat simpler) position of functionalist-panpsychist combinationists. They take phenomenal unity to be a fundamental relation, and thus treat it as division-invariant (and pervasive in nature). This is convenient, if I was right in the previous two subsections to treat both information integration and mutual co-presentation as division-invariant relations. For then the functionalist-panpsychist combinationist can say that, in the actual world at least, phenomenal binding is a division-invariant relation manifest in three ways: physically as information-integration, for component subjects as mutual co-presentation, and for composite subjects as representational unity. Although these three things might come apart in zombie worlds, Cartesian worlds, and other strange worlds, that fact is fairly inconsequential for what happens in this world.

Pure functionalist combinationists are in a more complicated situation. They tie phenomenal unity (like consciousness itself) to functional unity, and functional unity is division-relative: what roles states play in a given system may diverge from what roles they play in parts of that system or in a more encompassing system. Consequently phenomenal unity is also division-relative: two experiences might form a composite experience relative to one system but not another. Since phenomenal unity is a precondition for representational unity, this also applies to representational unity. But the intuitive plausibility of interdependence militates in favor of seeing representational unity as division-invariant. How can this tension be resolved? I am not entirely sure, but here is a suggestion.

I think pure functionalist combinationists ought to address this tension by distinguishing two relations: a division-invariant relation involving information integration and mutual co-presentation, and a division-relative relation which consists in the former relation plus phenomenal unity, yielding representational unity. Let us call the former “pre-phenomenal binding” and the latter “phenomenal binding”: pre-phenomenal binding is a matter of how two experiences affect each other, both physically and phenomenally, while phenomenal binding is just that same interaction considered relative to a system in which the two experiences together have the right functional profile to form a single composite experience.

That implies that we get representational unity relative to a given subject when (1) two experiences are integrating information and mutually co-presenting (both division-invariant facts), and (2) those two experiences are phenomenally unified relative to that subject. Strictly, interdependence is denied, and boundedness accepted, for representational unity, but only in the sense that making two pre-phenomenally bound experiences into representationally unified experiences (by considering them relative to a subject for whom they play the right functional **(p.182)** role to be phenomenally unified) does not change the individual phenomenal character of each. The spirit of interdependence is vindicated, because making two experiences pre-phenomenally bound in the first place (i.e., having them integrate information so as to mutually co-present) *does* change their phenomenal character, and whenever we undergo two representationally unified experiences, we can recognize this interdependence, this interpenetration, between them.

5.5. So How *Do* Intelligent Subjects Combine?

The foregoing discussion, particularly in section 5.4, got decidedly complicated, and the reader has my apologies for that. So let me try to summarize the key proposals made in this chapter, so as to generate a few principles that might make up the working core of functionalist combinationism.

Functionalist combinationism is concerned with “intelligent subjects,” which are characterized in both functional and phenomenal terms:

> An intelligent subject $=_{\text{def}}$ A conscious subject whose consciousness is structured and whose experiences play the functional roles characteristic of intelligence in virtue of their phenomenal character, conscious structure, and representational content.

Because this analysis of intelligent subjecthood is available, we can ground and explain the existence of an intelligent subject just by showing how the key features (consciousness, conscious structure, intelligent functioning, and coherence between the last two) come to be instantiated. Functionalist combinationism is concerned not with the general question of what explains and grounds the existence of intelligent subjects but with the much more limited

question of what role in that explanation and grounding could be played by the consciousness of component subjects.

Functionalist combinationism is compatible both with pure functionalism, which says that once intelligent functioning is in place all the rest follows (guaranteeing also the coherence between conscious structure and intelligent functioning), and with impure functionalism, which accepts the importance of intelligent functioning but takes something else to be necessary in order to add consciousness to the mix. In particular, panpsychists may accept panpsychist combinationism, from chapters 3 and 4, as their ultimate account of consciousness, but add functionalist combinationism as a "second layer" accounting for the special features of intelligent subjects.

(p.183) 5.5.1. Conditional Experience Inheritance

There is enormous disagreement about how consciousness and conscious structure can be explained, while the explanation of intelligent functioning is pretty simple in principle, though extremely complex in practice. So let us start with that:

> **Conditional functional inheritance:** If a part of X has an experience which plays a certain functional role, then that experience plays the same or a similar functional role for X, to the extent that the part in question is connected to the rest of X such that its experience has sensitivity to, control over, and coordination with other events occurring in X.

This does not yet say that the whole system is conscious, just that it has a state playing (something close to) the functional role of a conscious state. The two versions of functionalist combinationism then add different further principles to explain consciousness: pure functionalists can say that consciousness is reducible to having states playing the right functional roles, while panpsychists can say that systems inherit consciousness from being made of conscious matter. Either option allows us to move from conditional functional inheritance to conditional experience inheritance:

> **Conditional experience inheritance:** If a part of *X* has an experience which plays a certain functional role, then X shares that experience, with the same or similar functional role, to the extent that the part in question is connected to the rest of X such that its experience has sensitivity to, control over, and coordination with other events occurring in X.

This principle applies most straightforwardly to cases of trivial combination, when a single intelligent subject guides the overall functioning of a system that contains it. In such a case, CEI says that the system is also the subject of that part's experiences—that, for example, my brain and I share our consciousness. The more interesting case is when multiple parts of a system are conscious, and their experiences share control over, and sensitivity to, its overall workings, and

their experiences are sufficiently integrated with one another. In this case, the system inherits all their experiences, subsumed in a composite experience which each of them experiences only part of.

Note that conditional experience inheritance does not require that the conscious parts whose experiences are inherited are *intelligent* subjects; if unintelligent subjects are possible, then as long as their experiences play some functional **(p. 184)** role, however simple, those experiences could be inherited by the whole given the right functional embedding. A whole with experiences that play simple roles for its parts might connect them in such a way that they can play *more* complex roles in its overall functioning; that is, an intelligent subject might inherit experiences from unintelligent parts, as panpsychists say we do.[18]

5.5.2. The Three Faces of Phenomenal Binding

The structure of the whole's consciousness reflects the structure of, and relations among, the consciousness of its parts. My preferred metaphor for this is to think of each part's phenomenal field superimposed on the others, with points in the fields (i.e., experiences) sometimes pinned to each other (phenomenally bound), and the fields indeterminately lined up otherwise. In the limiting case of zero information integration (in a being which would not qualify as an intelligent subject at all) the composite experience is simply a phenomenal blend, a single quality involving no phenomenal contrast, no discernible structure. Metaphorically, the fields are completely unfixed in their alignment, as though spinning around over one another so that the structure of one simply blurs out the structure of the other.

I labeled the relation which brings the parts' experiences into a definite structure "phenomenal binding"; I have suggested that this relationship's within-subject manifestation in representational unity, that its between-subjects manifestation is mutual co-presentation, and that its objective aspect is information integration. But pure functionalist combinationists face a specific difficulty in trying to understand this relation. If pure functionalism is true, then phenomenal unity requires functional unity, which is division-relative, even though representational unity seems to make a genuine difference to the unified experiences, seemingly requiring it to be division-invariant (at least if the incompatible contexts argument is valid). This tension pushed me, in the previous section, to suggest that if pure functionalist combinationism is true, then although representational unity itself (which requires phenomenal unity) is division-relative, there is a division-invariant relation (provisionally labeled "pre-phenomenal binding") which involves information integration and mutual co-presentation, and which yields representational unity (relative to a particular division) as soon as phenomenal unity (relative to that division) is present.

(p.185) We can put these complications aside if we focus on cases where two experiences are functionally integrated enough for phenomenal unity relative to both parts and whole. In such cases, both sorts of functionalist combinationist can say the following: for the whole which experiences both, there is a representationally unified complex of experiences; for the parts which undergo only one or the other, there is an experience which co-presents another, by presenting its given content as one aspect of something not-fully-given. Corresponding to the increasing definiteness of the composite field is the increasing informativeness of the co-presentation present in the component fields. The parts' richer impression of each other's state, the whole's experience of one thing characterized by multiple aspects, and the integration of physical information across the system are, so to speak, three sides of the same coin.

5.5.3. Inclusionary and Exclusionary Approaches

Pure functionalist combinationists can leave matters there: composites inherit those experiences that are embedded in the right functional relations, but no others. Their phenomenal fields comprise only experiences that are both phenomenally conscious and access-conscious. Call this the "exclusionary" model of combination: any experiences in the parts that are not causally connected so as to be access-conscious are not inherited, and so not phenomenally conscious for the whole at all. This agrees with the spirit of Kant's (1997, B131-2) famous remark that "the 'I think . . .' must be able to accompany all my representations; for otherwise . . . the representation would either be impossible or else at least would be nothing for me."[19]

Functionalist-panpsychist combinationists face one last puzzle, however, which they might resolve in a similarly "exclusionary" way, or in a contrasting "inclusionary" way. The puzzle concerns the relationship between the consciousness of intelligent subjects and the consciousness of the aggregates that constitute them. The latter is governed not by conditional experience inheritance but by chapter 3's simpler principle of experience inheritance: it comprises all the experiences of the aggregate's parts, regardless of their functional role, and hence regardless of their accessibility. Does the intelligent subject too undergo all these inaccessible experiences?

(p.186) One option for functionalist-panpsychist combinationists is to say no, to insist that since the intelligent subject is not identical to the aggregate, it need not undergo the same set of experiences. The intelligent subject undergoes only the access-conscious experiences, and no others; the aggregate undergoes both the access-conscious experiences and also the rest. This view of the intelligent subject's consciousness could be called "exclusionary," just like the pure functionalist's view.

But this response might seem to rest too much weight on the distinctness of the intelligent subject and the aggregate. Structure-specific wholes are not strictly identical to what constitutes them, sure, but they're not truly distinct either. The tower of blocks is not quite the same thing as the mass of blocks, but it's also not quite another thing.

So it is worth noting another option for functionalist-panpsychist combinationists: the "inclusionary" approach. This would say that experience inheritance holds for intelligent subjects, implying that many intelligent subjects (those with conscious parts whose experiences are not functionally integrated into the whole's intelligence) experience a sort of "phenomenal overflow" (Block 1995, 2005, 2011; Rosenthal 2007): they are phenomenally conscious of things that they cannot cognitively access.

What would it be like to be such a subject, with many more phenomenally conscious experiences than one is access-conscious of? I think the claims made so far in this section suggest that these isolated experiences will form a sort of unattended, inarticulate "background consciousness," which all structured consciousness appears against.[20] This is because they are, almost by definition, not integrating information with the rest of our experiences, and so not phenomenally bound to any of our other experiences. In our metaphor of superimposing layers, these elements are not tied down enough to have any determinate place in the composite field—they are present, but equally present everywhere, and to that extent make little difference anywhere in particular. If two or more experiences were both isolated in this way, they would both occupy the background and moreover would blend together, as though to put a very faint lens filter onto our view of the world.

The inclusionary option is an extension of the radical confusion hypothesis from chapter 4. There I argued that there could be phenomenal overflow with respect to the internal structure of each access-conscious experience. Your feeling of pain when you prick yourself is certainly access-conscious: according to panpsychist combinationism this feeling is in fact composed of trillions of components **(p.187)** (the experiences associated with the microscopic brain events that add up to that macroscopic brain event), which are not access-conscious individually though they can be accessed as a mass.

Even an exclusionary approach would not deny that the pain is experienced by the whole (since it meets the right functional criteria), and experiencing the pain means experiencing the microexperiences that compose it. What is now at stake is experiences in the parts of something that are not, even in mass, access-conscious for the whole. For instance, consider the microexperiences associated (according to panpsychism) with all the events in my body that are not part of brain functioning (like events in blood cells, skin cells, the fluid in the brain's ventricles, and so on). The exclusionary option is to say that these experiences

make no contribution to the consciousness of the whole. The inclusionary option is to extend the radical confusion hypothesis to say that these experiences are confused with the whole's entire phenomenal field.

This raises the phenomenological question: Do we experience such a "background consciousness"? I cannot be sure: at times it sounds like an accurate description of my experience, but at other times I am doubtful. And by definition, it will be all but impossible to conclusively detect, since it makes so little functional difference. Rather than come down on one side or the other, I will simply note various places in the next chapter that certain experiences will be informationally and functionally isolated, and will thus either not belong to the composite subject, or form an inarticulate conscious background for it.

Panpsychist combinationism aimed to be a theory of fundamental reality, subject to the standards appropriate to such theories: parsimony and uniformity of ultimate laws and *a priori* entailment of other laws. Functionalist combinationism does not aim for fundamentality: functional structure is by definition nonfundamental. Consequently the standards of success are somewhat different. It is still important to seek simple, elegant principles that are not needlessly complex or arbitrary. But nonfundamental principles can achieve elegance in ways that would be unacceptable for fundamental principles: they can include "other things being equal" clauses, to allow for an open-ended range of exceptions to a given generalization; they can restrict themselves to applying only under specific conditions, and employ simplifying concepts, even if those conditions or those concepts are themselves very hard to specify simply in fundamental terms.[21]

(p.188) Consequently, I will make use of ideas (like "intelligent functioning," "information integration," and "conscious structure") that I cannot give explicit, reductive definitions: all we should demand is that these ideas should be no less clear, intuitive, and illuminating than familiar psychological and philosophical ideas (like "sensory modality," "instrumental reasoning," "propositional content") that we already use to think about human minds. That is, my hope is not that functionalist combinationism will be a complete account of intelligent subjects, but only that it can be usefully incorporated into our ongoing, slow but steady efforts in understanding them.

Notes:

(1) Note that I have not included capacities like language use or reflective self-consciousness, nor have I made any special mention of reasoning, problem-solving, or abstract thought, all of which are prominent in some definitions of the word "intelligent." My aim is to capture what is distinctive about a wide range of animals, not what is distinctive about humans or any subset of humans.

(2) Note that it may often be useful, in thinking about these intermediate cases, to treat “intelligent functioning” as gradable, so that functioning can be more or less intelligent. This removes some of the vagueness of “functions intelligently,” construed as a predicate that does or does not apply, just as we can remove the fuzzy boundary between “is bald” and “is not bald” by simply rating each head on a scale of “degrees of hairiness.” But this will not remove all vagueness, since there are many different factors involved (in both intelligent functioning and baldness). Any particular scale will depend on picking a particular way to weight these factors, out of many equally valid weightings, giving rise to indeterminacy about the ordering of certain cases (e.g., is the man with fewer but thicker hairs “balder” than the man with more but thinner hairs?).

(3) Is the structure-specific whole really a distinct thing from the aggregate of blocks? Wouldn’t that have the strange result that two distinct physical objects occupied the very same space? Yet it seems contradictory for them to be the same thing, since one came into existence earlier than the other. These paradoxical puzzles have been discussed extensively (see, e.g., Geach 1967; Gibbard 1975; Bennett 2004), and I make no attempt to solve them here. I hope simply to show how experiential combination is no *more* problematic than physical combination. Readers are welcome to regard either the structure-specific whole or the aggregate as a mere linguistic fiction, not a real entity.

In particular, readers might want to “cut out the middleman,” dispensing with the aggregate and saying simply that a structure-specific whole exists when a set of parts collectively instantiate a corresponding property to its defining properties (e.g., a tower exists when some objects collectively instantiate “being arranged towerwise,” because that is a corresponding property to the tower’s defining property of “being tower-shaped”). But this is not a substantive change, since the whole point of SI is that the aggregate is just the parts considered as one. To drop it from the explanation in favor of the parts collectively is just to drop it and replace it with itself.

(4) There is an ambiguity in “the essential properties of intelligent subjects.” I mean the properties which are essential to something’s being an intelligent subject; these might not be the same as the properties which are essential to a certain intelligent subject’s being itself. The two might come apart if an intelligent subject could cease to be an intelligent subject while remaining itself, as we might think is the plight of a patient in a persistent vegetative state.

(5) “Largely” here does not mean “entirely”: much of the brain’s processing is influenced by features of and feedback from the body, in such a way that assigning sole responsibility for intelligence to the brain is arguably a distortion. Nevertheless an oversimplified idea of the brain-body relationship is a useful illustration of the point being made here.

(6) Here is an analogy. A cell can be said to be “biologically organized,” meaning something like “organized in the way characteristic of biological organisms.” But a compost heap that contains that cell is not biologically organized (showing that “being biologically organized” is division-relative). Nevertheless, there is something we can say of the compost heap: it contains biological organization (namely, it contains that of the cell and its other biologically organized parts). After all, someone looking to study biological organization could do much worse than looking at a compost heap. So we can distinguish something division-invariant (“biological organization”) from something division-relative (“biological organizedness”). My distinction between “structured consciousness” (division-relative) and “conscious structure” (division-invariant) is analogous.

(7) If we are more interested in propositional representations, the analogue might be to imagine a great list of propositions, or (perhaps equivalently) an extremely long conjunction: if part of me thinks “it’s hot” and part of me thinks “I’m excited,” I end up thinking “it’s hot and I’m excited.” (Cf. the idea of phenomenal unity as closure under conjunction: Tye 2003, 36–40; Bayne 2010, 47–72).

(8) It is a tricky question what literal sense can be made of the popular phrase “phenomenal field.” In Roelofs (2014b) I defend one interpretation, where “distances” correspond to degrees of causal interdependence.

(9) Indeterminate reference is a better model for the indeterminacy I am ascribing to unstructured composites, rather than indeterminate application of predicates. For example, many people have thoughts about “the outback,” but as Lewis (1986, 212) notes, “it’s vague where the outback begins . . . because there are many things [precise regions of desert], with different borders, and nobody has been fool enough to try to enforce a choice of one of them as the official referent of the word ‘outback.’ ” Does a given person’s thought that “the outback is hot” refer to outback_{4356}? Does it refer to outback_{4876}? There is no way to say that it refers to one but not the others, so the most reasonable thing to say is that it refers indeterminately to all of them: its content is indeterminate between “outback_{4356} is hot,” “outback_{4876} is hot,” and all the others. In something like this way, it seems to me, we must suppose that unstructured composites, if they represent anything, indeterminately represent a baffling array of different contents.

(10) This fits with the metaphor of “pinning”: putting a pin through two layers ensures that the two elements the pin goes through are co-located in the composite field, but since the two layers can still move relative to each other, the alignment of other elements remains indeterminate. Of course, if the layers were imagined as two-dimensional and perfectly rigid, then all indeterminacy could be removed by just two pins. But we should probably not imagine them as two-dimensional and rigid. (Indeed we should be agnostic about how far any such

metaphorical picture matches the actual features of microexperience.) In Roelofs (2014b) I argue that the human phenomenal field is a deforming one, with an open-ended number of dimensions.

(11) This is a deliberately simplified summary of an intricate mathematical structure, which admits of multiple subtly different types of “phi” and has evolved through multiple versions. I have tried to capture the core idea that remains constant through all versions; interested readers should consult Tononi 2004, 2008, 2012; Tononi and Koch 2014; Oizumi et al. 2014; and the website www.integratedinformationtheory.org.

(12) The stated motivation for the exclusion postulate is Ockham’s razor: given consciousness at one level, it would be superfluous to also postulate consciousness at other levels. But this reasoning makes sense only if consciousness at the other levels was something over and above consciousness at the highest-phi level, which is precisely what combinationism denies: the whole is conscious, but its consciousness is not something new in addition to the consciousness of the parts. By analogy (and recalling section 3.5 of chapter 3) I think it would be absurd to deny that composite objects have any mass, on the basis that mass for the composite would be a superfluous addition to the mass of its parts: its mass is nothing over and above theirs, and so “comes for free.”

(13) Unlike in the perceptual case, of course, they do not present themselves through a distinct experience that represents them (that would obviously lead to an infinite regress of experiences presenting experiences); rather, they are presented to the subject just by themselves. That is, this suggestion assumes that it makes sense to see experiences as “self-representing” (or “self-presenting,” if one prefers), i.e., the idea that whenever we have an experience (e.g., seeing a green plant) we are simultaneously made aware of two different things: the external object which the experience represents (the plant) and the experience itself. For defenses of this idea see Strawson 2015; MacKenzie 2007; Zahavi and Kriegel 2015; for some critical evaluation see Schear 2009.

(14) See Mendelovici (2018) for a similar point: the challenge of explaining the mind’s various forms of structure should not be seen as a special problem just for panpsychists (or other combinationists).

(15) See, e.g., Treisman and Gelade 1980; Duncan and Humphreys 1989, 1992; Treisman and Sato 1990; cf. the problems raised for adverbial theories of intentionality in Jackson 1975.

(16) See particularly Soames 2010, where a mental act of predicating is appealed to in order to solve the long-running debate on the “unity of the proposition.”

(17) I am assuming for the sake of argument that at least some conscious states have representational content intrinsically; if this is false, and conscious states either lack content or get it entirely in virtue of factors or processes (e.g., functional organization or evolutionary functions) which have nothing specifically to do with consciousness, then the explanatory burden on the combinationist is correspondingly lightened. Note that this implies that I am primarily discussing “narrow” rather than “broad” content: if the content of a state depends partly on extrinsic factors (like causal history), then a theory of consciousness will account only for those aspects of content that are independent of those extrinsic factors.

(18) Compare Lycan (1995, 40) on homuncular functionalism: “We explain the successful activity of one homunculus, not by idly positing a second homunculus within it that successfully performs that activity, but by positing *a team* consisting of several smaller, individually less talented and more specialized homunculi—and detailing the ways in which the team members cooperate in order to produce their joint or corporate output.”

(19) Other sentiments in this ballpark include Rosenthal’s (1986, 329) principle that “conscious states are simply mental states we are conscious of being in” and Chalmers’s (1995, 202) “detectability principle,” that “where there is an experience, we generally have the capacity to form a second-order judgment about it.”

(20) There is some affinity here with the *alayavijñana*, “background consciousness” or “storehouse consciousness,” posited by certain Buddhist philosophers in the Yogacara school (Asanga 1992).

(21) For example, an elegant biological principle might be something to the effect that “species that produce large numbers of offspring invest less care in each one, and tend to prosper in unstable environments, while species that produce small numbers of offspring invest more care in each one, and tend to prosper in stable environments” (cf. MacArthur and Wilson 1967; Pianka 1970). But this principle would not fare well by the standards applied to fundamental theories, because there are numerous large or small exceptions, and because the conditions under which it applies (e.g., ones where populations of reproducing organisms can exist) and the concepts it employs (e.g., “parental investment”) would be very hard to specify simply in fundamental terms.