CHAPTER 1

Mind-Body Theories and the Emotions

William Jaworski

Problems for an Ontology of Emotions

Emotions are reckoned to have a variety of characteristics (Goldie 2000). Among other things we feel them, and our feelings have intentionality or directedness: I am afraid *of* something; I am angry *about* something; I long *for* something. But what exactly are the states that have these characteristics? When an individual experiences an emotion, what is that emotional state? Is it a physical occurrence or not? If it is, how is the psychological description of it related to the physical description of it? Does 'anger' always refer to the same type of physical state, or can it refer to different types of physical states under different circumstances? Is it possible to give an exhaustive account of the emotion in purely physical terms? If not, if the experience is not a physical occurrence, how is it related to the physical occurrences that seem to accompany it – to events in the limbic system of the brain, for instance? In what follows I'll survey some representative answers.

One way to understand various positions on the ontology of emotions is to see them as responses to mind-body problems – persistent problems in understanding how thought, feeling, perception, and other mental phenomena fit into the universe as described by our best science. To many philosophers and scientists it seems plausible that the behavior of everything in the universe can be described and explained exhaustively by physics. Yet we evidently have capacities, such as our capacities to think, feel, and perceive, that it is difficult to imagine could ever be described and explained by physics alone. Mind-body problems are expressions of this difficulty. One example is the problem of emergence.

The problem of emergence is the problem of explaining how lower-level physical or physiological occurrences can generate or produce higher-level mental phenomena such as conscious emotional experiences. Consciousness did not always exist in the physical universe. Neither does consciousness exist in all parts of the universe. Many philosophers and scientists take these observations to indicate that very specific physical conditions must be in

place in order to cause or produce conscious phenomena. But what conditions are those, and how do they do it? How is it that, say, the movements of tiny particles in my brain can give rise to the rich emotional experiences I have? The following claims make it difficult to see how:

(1) We have conscious emotional experiences.
(2) We are composed of physical particles.
(3) The properties of a composite whole are determined by the properties of the particles that compose it.
(4) Physical particles do not have conscious emotional experiences.
(5) No number of nonconscious particles could combine to produce conscious emotional experiences.

Each of these claims is plausible on its face. It seems obvious that we have conscious emotional experiences as claim (1) says. Claim (2), moreover, seems well supported empirically; we seem to be composed of the same materials as everything else in the physical universe, and our best physics suggests that those materials are microscopic particles. Many examples seem to illustrate claim (3). I have the mass I have, for instance, because I am composed of physical particles with smaller masses that collectively add up to my bigger mass. Likewise, I have the position and velocity I do because the particles composing me are located in such-and-such a place and are moving with such-and-such a velocity. Change their position and velocity and you succeed in changing mine. Given the range of properties that are like this, it's not implausible to suppose that all the properties of composite wholes are determined by the properties of the particles composing them. It seems, moreover, that the behavior of those particles can be described and explained exhaustively by physics. We don't need to invoke a psychological or even a biological vocabulary to describe and explain what they are and what they can do. This lends some support to claim (4). There are also, it seems, good reasons to endorse claim (5). One particle by itself does not have the power to produce conscious experiences. If it did, then consciousness would have emerged much earlier in the universe's history than we think it did, and it would also be more widespread – even rocks and tables could be conscious. But if one particle by itself does not have the power to produce conscious experiences, then it is difficult to see how any number of nonconscious particles could combine to produce conscious experiences.

Each of the foregoing claims (1)–(5) is therefore plausible. Yet claims (1)–(5) are jointly inconsistent. Claim (1) implies that we have conscious emotional

experiences, yet claims (2)–(5) imply that we do not. The claims cannot all be true; at least one of them must be false, but it is not clear which.

Eliminative Physicalism, Substance Dualism, and Panpsychism

Mind-body theories offer to solve mind-body problems in various ways. Eliminative physicalists, for instance, look to resolve the problem of emergence by rejecting claim (1). A psychological vocabulary is the by-product of a defective way of trying to describe and explain human behavior, one that will eventually be displaced by a complete physical description of that behavior. As a result, it is false to say that we have conscious emotional experiences. There are no such things. Substance dualists, on the other hand, look to reject claim (2). According to them, we are nonphysical beings that are only contingently attached to bodies; we are not bodies ourselves. Hence, we are not composed of physical particles or stuffs. Panpsychists, for their part, reject claim (4). They claim that fundamental physical particles are endowed with thoughts and feelings just as we are. As a result, there is no question of how those particles might come together to produce conscious emotional experiences. Conscious emotional experiences exist everywhere in the universe, right down to the fundamental physical level.

Eliminativism, substance dualism, and panpsychism are nevertheless marginal positions. Eliminativism is marginalized because it denies what seems to many people to be the manifest fact that we experience anger, joy, sadness, and other emotions. Substance dualism, on the other hand, is marginalized because it doesn't do justice to the phenomenology of emotional experience – its visceral character in particular. William James described that character in a well-known passage:

> What kind of an emotion of fear would be left, if the feelings neither of quickened heart-beats nor of shallow breathing, neither of trembling lips nor of weakened limbs, neither of goose-flesh nor of visceral stirrings, were present, it is quite impossible to think. (1884: 193–4)

If James is right, then emotions are essentially embodied; it is impossible for us to experience them apart from particular bodily changes, and if that is the case, then one of substance dualism's central claims about mental phenomena cannot be true of the emotions: they cannot exist apart from a body. Descartes himself, it seems, felt the pressure of this observation, and posited a third class of properties to try to account for the visceral phenomenology of emotional experience. In addition to mental properties and physical properties, he said, there were also properties pertaining to the

union of mind and body – properties that included the emotions (Cottingham 1985).

Panpsychism is marginalized for its part because it is so counterintuitive to think that quarks, leptons, and other fundamental physical particles might have mental lives as rich as our own. Our best empirical accounts of these particles do not invoke a psychological vocabulary; we can describe, explain, and predict their behavior without positing thoughts or feelings.

In what follows I will focus on views that deny either claim (3) or claim (5). Views that deny (5) can be divided into three categories: physicalist theories, dual-attribute theories, and neutral monist theories. Views that reject (3) include hylomorphism.

Reductive and Nonreductive Physicalism

Physicalism claims that everything is physical; everything can be exhaustively described and explained by the most empirically adequate theories in current or future physics (Lewis 1983; Jaworski 2011, 2016). Physicalism has been the most popular framework for addressing mind-body problems since the mid-twentieth century. If we put to one side eliminativist theories, physicalist accounts of the emotions fall into two categories: reductive and nonreductive. The paradigmatic reductivist theory is the psychophysical identity theory.[1] It claims that mental states are identical to physical states and that we will discover these identities empirically in something analogous to the way we discovered that water is H_2O (Lewis 1966, 1972; Armstrong 1968, 1970; Jaworski 2011). Water might initially have been defined as, say, the drinkable stuff that filled rivers and lakes. By studying that stuff empirically, scientists were able to discover that it was H_2O. They were thus able to conclude that water was identical to H_2O. According to identity theorists, something analogous will be true of the emotions. Mental states in general might be defined initially by their typical causes and effects. Anger, for instance, might be defined as the state that is typically caused by suffering injustice and that typically causes such-and-such feelings and physiological changes. According to identity theorists, it is possible in principle to discover what the state with these typical causes and effects is. It is thus possible in principle to identify anger and any other emotion with some type of physical state.

One influential challenge to the identity theory has been the multiple realizability argument (Putnam 1967; Block and Fodor 1972; Fodor 1974;

[1] Behaviorism is the other kind of reductive physicalism (Jaworski 2011: 103–11).

for discussion see Jaworski 2011: 131–3).[2] The argument claims that it is possible for a mental state, such as anger, to be correlated with more than one type of physical state. If anger can be correlated with brain state B in humans and with a different type of physical state in Martians, then anger cannot be identical to brain state B. A thing cannot exist without itself, so if anger can exist in a Martian who doesn't have brain state B, then anger cannot be identical to brain state B. Identity theorists have responded to this argument in a number of ways (Kim 1972; Lewis 1980; for discussion see Jaworski 2011: 134–6). The most popular line of response claims that our current taxonomy of mental and physical states will undergo a revision in the future. Even though our current taxonomy posits a single type of state that goes by the name 'anger', that term instead can refer to multiple different types of states: anger-in-humans, anger-in-Martians, anger-in-robots, and so on. By analogy, people once took 'jade' to refer to a single type of mineral, but over time they came to recognize that it referred to multiple different types of minerals such as jadeite and nephrite. It is thus a mistake to speak of jade in general, and in the same way it is a mistake to speak of anger in general. There are instead many different types of states which have been grouped under the one confused heading of anger, and each of these species- or kind-specific types of states is identical to a physical state: anger-in-humans = physical state A, anger-in-Martians = physical state B, anger-in-robots = physical state C, and so on.

Far and away the most popular way of accommodating multiple realizability, however, has not been to rehabilitate the identity theory, but to abandon it in favor of some kind of nonreductive physicalism. The most popular theories of this sort combine physicalism with functionalism (Putnam 1967; Block and Fodor 1972; Fodor 1974; for discussion see Jaworski 2011: 136–49). Functionalism claims that psychological descriptions are abstract descriptions that ignore the physical details of a system and focus simply on inputs to it, outputs from it, and internal states that correlate the two. When we say that Eleanor is enjoying the play, we are saying that she is in some internal state that correlates certain sensory inputs to her with certain behavioral outputs from her. The psychological description does not specify what that internal state is; that's what makes the description abstract. This abstractness allows functionalists to accommodate the possibility that a single kind of mental state might be correlated with different kinds of physical

[2] For a detailed survey of the literature on multiple realizability and reductionist responses to it, see my 'Mind and Multiple Realizability' in the Internet Encyclopedia of Philosophy (www.iep.utm.edu/mult-rea/).

states. A human like Eleanor might differ physically from a Martian like Gabriel, but the physical differences between them needn't appear at an abstract psychological level of description. At that level, Eleanor and Gabriel might be indistinguishable; both might correlate inputs with outputs in a way we would describe as enjoying the play. Consequently, the same type of emotional state, enjoying the play, could be correlated with different types of physical states: one type of physical state in a human, and a different type of physical state in a Martian.

Nonreductive physicalism of the aforementioned sort has faced several challenges (Jaworski 2011: 149–64). Some of these have targeted physicalism, others functionalism, and yet others the combination of the two. When it comes to the emotions specifically, physicalist theories have faced challenges accommodating both the intentionality of emotional states and their felt qualities. Physicalists have typically tried to account for intentionality in terms of theories of mental representation (Fodor 1987; Dretske 1988; Jaworski 2011: 90–1). The simplest theories of this sort claim that mental representation consists in causal covariation between features of the environment and internal states of the organism or representing system. Suppose that my nervous system has a component, *c*, that is capable of being in two states, ON and OFF, and that these states covary with the presence of something red: *c* turns ON when it encounters something red, and it is OFF otherwise. Because *c*'s ON/OFF state covaries with the presence of redness, it indicates the presence of redness in something analogous to the way smoke indicates the presence of fire. Because fire reliably causes smoke, the presence of smoke typically conveys the information that there is fire. Similarly, because redness reliably causes *c* to be ON, *c*'s being ON typically conveys the information that something in the environment is red. To have an internal representation of redness is thus to have an internal component that is activated if and only if something in the environment is red.

Simple covariation theories nevertheless have difficulties accounting for the specificity of intentional content – for what makes a representation of or about this particular object, property, or event as opposed to some other (Fodor 1987). This is especially clear in the case of more sophisticated mental states such as emotions. Anger, let us suppose, is typically triggered by something insulting. According to a simple covariation theory, my anger should be reliably caused by insulting episodes. If I'm an irascible person, however, my anger might also be reliably caused by episodes that aren't insulting at all. What is it, then, that makes my anger a state that is about insults instead of a state that is about insults-or-non-insults? Simple

covariation theories do not provide a satisfactory answer to this question. As a result most physicalists endorse more sophisticated theories of mental representation such as Fred Dretske's (1988).

Dretske's theory of mental representation combines causal covariation with designed or naturally selected functions. On Dretske's view, the components of a system perform various jobs or functions within it. In the case of artifacts, those functions are determined by a designer, someone who constructs the system with an eye to having its components contribute to an overall task. In the case of natural systems such as human organisms, the functions performed by various components are determined by natural selection. It plays a role in the development of organisms analogous to the role of a designer in the production of artifacts. Natural selection has assigned to various sensory organs or subsystems the functions of supplying organisms with information about the external environment and their other internal states – functions they perform by having internal states that covary with features of the environment and with other internal states.[3] Physicalists argue that the causal relations connecting states of the nervous system to features of the environment, as well as the natural selective mechanisms that determine something's function, can all be given an exhaustive physical description and explanation. If they are right, then mental representation can be exhaustively described and explained in principle by physics, and if intentionality can be understood in terms of mental representation, then it too can be accommodated within a physicalist framework.

Qualia-based objections to physicalism argue that physicalism cannot account for the qualitative aspects of experience, or qualia (Jackson 1986; Chalmers 1996, 2002; Jaworski 2011: 83–8). According to qualia theorists, there is something it's like to feel anger or fear or jealousy, and what it's like to experience one of these emotions is different from what it's like to experience another.[4] This what-it's-likeness, say qualia theorists, cannot be physical. One reason, they say, is that it seems possible for two physically indistinguishable individuals to have different qualia. What it's like for you to experience anger might be different from what it's like for an exact physical duplicate of you to experience anger. When your duplicate experiences anger, it might feel to him or her the way enjoyment feels to you. Or possibly when your duplicate experiences anger it might feel like nothing;

[3] Prinz (2004) defends a theory of the emotions along these lines.
[4] Not all theories of emotions agree on this. Theories like James Russell's (2003) claim that feelings are not intrinsically emotional. Feelings are instead neutral episodes that can be mere moods or that can be components of emotions if suitably directed.

your duplicate's anger might not have any qualitative dimension at all. If qualia were physical, qualia theorists argue, then necessarily you and your duplicate would have to have the same qualia, for you and your duplicate have all the same physical characteristics. But since it is possible for you and your duplicate to have different qualia, it follows that qualia must not be physical; they cannot be accommodated within a physicalist framework.[5]

A different challenge to providing a physicalist account of the emotions targets the combination of physicalism and functionalism. It claims that a theory of this sort ends up evacuating psychological discourse of any real explanatory content (Kim 1998; Jaworski 2011: 161–4, 169–76). If this is true, nonreductive physicalism implies that the real reasons why we behave as we do have nothing to do with our emotional states. To appreciate the argument, consider an analogy. Let us suppose that the behavior of fundamental physical particles can be exhaustively described and explained by physics. Suppose, however, that we decide to call particles 'J-particles' exactly if I'm giving a thumbs-up. In an instant I can transform every particle into a J-particle simply by raising my thumb, and conversely, I can transform every J-particle into a non-J-particle simply by putting my thumb down. These so-called transformations, however, are not real. In bringing it about that every particle is a J-particle, I do not alter in any way how those particles behave or the reasons why they behave as they do. The exhaustive account of their behavior and the reasons for it continue to belong to physics alone; it has nothing to do with J-particle discourse. The worry about nonreductive physicalism is that it implies something analogous about psychological discourse: if it is true, psychological discourse has no bearing on what things do and why they do it. If physicalism is true, then everything can be exhaustively described and explained using the vocabulary of physics. It is possible to describe the same physical things using a different vocabulary, including an abstract psychological one, just as it is possible to call the same particles 'J-particles'. But whatever vocabulary one chooses to describe physical things, if physicalism is true, the exhaustive account of what those things do and why they do it is still given

[5] Physicalists have responded to the qualia challenge in several ways. Some deny that qualia exist. Dennett (1992), for instance, argues that the very concept of qualia is incoherent (Jaworski 2011: 219–28). Others, such as Heil (Chapter 2, this volume), argue against the possibility of physical duplicates having different qualitative experiences. Yet others argue that even though qualia theorists insist that qualia constitute the raw data of experience, qualia are in fact posits of a dubious theoretical framework that fails to capture the phenomenology of lived perceptual experience (Noë and O'Regan 2002; Noë 2004). Other physicalists have challenged the claim that qualia cannot be accommodated within a physicalist framework. Dretske (1995), for instance, argues that it is possible to give an account of qualia in terms of mental representation.

by physics alone. Psychological discourse thus seems to be as irrelevant to that account as J-particle discourse; it is bereft of any real description or explanatory value. In fact, the problem looks even worse for nonreductive physicalists. They insist, after all, that the categories of psychological discourse do not correspond in a straightforward way to the categories of physical theory. A type of psychological state like anger could correspond to many different types of physical states. But if the physical categories are the ones that give us an exhaustive account of reality, and psychological categories don't correspond to those, then it is difficult to see how psychological categories could correspond to reality at all. If nonreductive physicalism is true, therefore, it is difficult to see how talk of thoughts, feelings, perceptions, and the like could provide any kind of explanation for human behavior.

Dual-Attribute Theories and Neutral Monism

Like most physicalist theories, dual-attribute theories claim that nonconscious particles could combine to produce conscious emotional experiences, but unlike physicalist theories, they deny that the account of how the production happens can be given exhaustively in physical terms. Rather, when conditions are right, physical things produce nonphysical attributes or properties, ones that cannot be described purely in terms of physics.[6]

The most popular dual-attribute theories are varieties of emergentism or epiphenomenalism. Both claim that physical states produce nonphysical mental ones; they differ over whether those nonphysical states can exert any causal influence over physical states in turn. Emergentists say they can; epiphenomenalists say they can't. What concerns us here, however, is the claim they share, namely that physical states produce nonphysical mental ones. How exactly is this supposed to happen? Critics argue that dual-attribute theorists have no satisfactory answer.

Some dual-attribute theorists, for instance, posit brute psychophysical laws. There are laws of emergence, they say, that are every bit as basic as the

[6] Dual-attribute theories are sometimes called forms of property dualism to contrast them with forms of substance dualism. But this use of the term 'property dualism' can be misleading since substance dualists are committed to property dualism as well (Armstrong 1968: 11). Likewise, the label 'dual-aspect' has also been used instead of 'dual-attribute', but this too can be misleading since 'aspect' suggests that, according to dual-attribute theories, the mental-physical distinction is merely a matter of how things appear to us. The term 'attribute' expresses more clearly the ontological nature of the dualism that dual-attribute theorists endorse.

laws governing purely physical interactions. Consequently, the world just is a place in which physical events produce mental events in accordance with basic laws. Critics argue, however, that psychophysical laws do not really address the problem (Strawson 2006; Jaworski 2011: 224–33). Knowing that there is a psychophysical law linking pain to brain state B might explain why this or that instance of pain is correlated with this or that instance of B, but it does not explain why there is a correlation between instances of pain and instances of B in the first place. It does not explain how the rapid vibrations of countless tiny physical particles can generate the felt qualities of fear or anger; it simply says that the generating happens in a regular, lawlike way. According to critics, however, some explanation is needed for why there would be laws connecting things as different as brain states and qualitative experiences. Without such an explanation psychophysical laws are every bit as mysterious as the psychophysical correlations they are introduced to explain.

Dual-attribute theorists are also free to endorse panpsychism: mental phenomena such as conscious emotional experiences are present at every level of reality, including the fundamental physical level. As we've seen, however, panpsychism is a marginal position because it is so counterintuitive to think that fundamental physical particles have mental states like our own. To soften the counterintuitive implications of panpsychism, some dual-attribute theorists endorse panprotopsychism, the claim that fundamental physical particles do not have mental states like ours, but rather protomental states – simpler precursors of the mental states we have (Chalmers 1996: 153–5; 2002: 265–7). According to panprotopsychists, when fundamental physical particles combine to form atoms or molecules, they give rise to protomental states that are more sophisticated, and when atoms and molecules combine to form neural tissues or different parts of the brain, they give rise to mental states that are more sophisticated still, and so the combinatorial process goes until combinations of lower-level items eventually give rise to the rich, sophisticated mental states that constitute our own mental lives.

Critics argue, however, that panprotopsychism only solves the problem of emergence by replacing it with a different problem: the problem of saying what protomental states are and how they combine to produce more sophisticated mental states. We know what thoughts, emotions, and perceptions are, but what are protothoughts, protoemotions, and protoperceptions supposed to be? Panprotopsychists can respond that protomental states are theoretical postulates, that they are precisely the states of lower-level things that in combination produce familiar mental states. But how exactly is any