1

Introduction

Philosophical History and the Problem of Consciousness

The history of analytic philosophy, if viewed as more than a repository for superseded theory, could provide the basis for a transformation in the problem of consciousness with which philosophers of mind are currently grappling. Philosophers of mind seldom discuss or investigate, more than cursorily, the history of the interrelated concepts of mind, consciousness, experience, and the physical world that they rely upon in their theorizing. But these concepts in fact emerge from some of the most interesting and decisive philosophical struggles of the analytic tradition in the twentieth century. Historically, these struggles and their results set up the philosophical space in which contemporary discussion of consciousness moves, defining and delimiting the range of theoretical alternatives accessible to participants in the discussion of the explainability of consciousness and its relation to our understanding of the physical world.

Most contemporary philosophical discussions of consciousness address the question of its explainability in terms of objective, scientific description or the question of its ontological reducibility to objective, scientifically describable phenomena. Philosophers often raise these questions, moreover, against the backdrop of the thought that consciousness has certain properties or features that may make it especially resistant to scientific explanation and description. Paramount among the features of consciousness usually cited as problems for its explanation or reduction are its *privacy, subjectivity, ineffability, phenomenality, immediacy*, and *irreducibly qualitative character.*[1] These features

or properties are typically taken as problematic for one or both of two "naturalistic" programs of explanation: either *physicalism*, which holds that a successful explanation of consciousness accounts for it as wholly physical, or *functionalism*, which holds that a successful explanation accounts for conscious states as functional states of the brain or person. The debate about the reality and reducibility of these special features of consciousness, having developed over the 1980s and 1990s, shows no sign of being resolved, and indeed, it is unclear what sort of consideration, empirical or philosophical, might decisively settle it.[2] But historical analysis offers to reinvigorate the debate, bringing it to a greater richness and philosophical depth. It does so by showing that each of the determinate notions used in these various types of arguments to characterize (or to contest the characterization of) the specific properties of consciousness, and the forms of explanation appropriate to understanding them, in fact originate in the historical context of bygone philosophical theories and concerns, often seemingly quite distant from those of philosophers who apply those notions today.

Broadly speaking, several of the main aspects of the contemporary discussion of consciousness – in particular, the discussion of its alleged privacy, ineffability, and subjectivity – first arise historically from tensions present in analytic philosophy's longstanding attempt to describe the relationship between linguistic meaning and experience.[3] Historical analysis elucidates this attempt, revealing its underlying form and clarifying its significance for today's debate. Characteristically, analytic philosophy is a *linguistic* inquiry. For the purposes of historical reconstruction, it can be defined as a specific tradition in terms of its determinative and unique attention to language and its logic, and this attention determines the historical and contemporary form of its inquiry into the nature of experience. In particular, analytic philosophy typically investigates the *conceptual and logical structure of language* in order to understand experience and to explain its relationship to objective knowledge about the physical world. From around the turn of the twentieth century, the explanatory projects that would define analytic philosophy of mind sought to elucidate the epistemology and ontology of our knowledge of the objective world on the basis of reasoning about the *structure* of experience or consciousness, the total pattern of the logical or conceptual interrelationships of its basic elements.

One of the inaugural innovations of analytic philosophy was to tie this explanatory project to a program of *linguistic analysis*, whereby the structure of experience is specified by means of a clarification of the logical relationships between *propositions*, both those immediately describing experience and other, more highly conceptual and interpretive ones. Within this program, the analysis of experience, consistently identified with the analysis of the *language* of experience, is the analysis of the logical and conceptual structure of this language, of the network of the syntactic and semantic interrelationships of the terms and sentences that describe, explain, and express experience. The goal of analysis is then the identification and description of this structure of relations. But from the beginning of the analytic tradition, the basic elements of experience figure as the indefinable *relata* of this network of relations, the elements that can be described and explained only by reference to their semantically and conceptually relevant interrelations, and never in themselves. This configuration – in which consciousness is constantly understood as immediate content, and objective language and explanation as relational – has, despite changes in detail and emphasis, continued to characterize the discussion of the problem of consciousness to the present, through the various shifts in doctrine and method that the analytic inquiry into experience has undergone over the twentieth century.

A *structural* or *structuralist* explanation (in the sense in which I use these terms in this study) is one that accounts for particular items by *locating* them in a broader structure of *relations* of one kind or another.[4] Structuralist explanation typically operates by *first* characterizing the nature of the system of interrelations in which a type of events or objects stand, and *then* explaining particular items by locating them within this system. Thus defined, structuralist explanation is an exceedingly general explanatory practice. As we shall see, for instance, it subsumes many forms of *semantic* explanation whereby words, concepts, or meanings are explained in terms of their logical or semantic roles in a language, as well as most forms of *causal* explanation that explain particular objects or events in terms of their position in a structure of causes and effects. The explanatory projects most prominent in the contemporary debate about consciousness are themselves versions of structuralism.[5] *Physicalism* or materialism, for instance, is the doctrine that every real phenomenon can be described and explained in

terms of basic physics. It operates explanatorily by locating each puzzling phenomenon within the total pattern of relations that physics can capture, typically a pattern of *causal* relations that is conceived of as exhaustive of reality. *Functionalism* is the doctrine that mental states, including states of consciousness, are completely explainable in terms of their functional interrelationships with other mental states and physical states. Understanding mental states as definable in terms of these interrelationships, it always explains them by locating them within a total pattern of relations.[6] These explanatory projects, as we shall see in the chapters to follow, themselves have a rich and hidden conceptual history in the analytic tradition, one that entwines them inseparably with the problems of experience and consciousness that they were developed to solve. Historical analysis, by exposing this conceptual history, shows the extent and depth of the entwinement of structuralism with the problems of explaining consciousness, suggesting new possibilities for the understanding and resolution of these underlying problems.

Not *all* forms of explanation, however, are structuralist in this sense. Consider, for instance, *genetic* explanations (that explain things in terms of their origins and histories of descent) and *narrative* explanations (that explain by situating particular things or events within a larger narrative story). Though these other forms of explanation might refer to or make use of larger contexts or unities – a specific history, for instance, or a broader narrative – they do not function primarily, as structuralist explanations do, by locating items within a larger pattern of interrelations of a particular kind. If the point of explanation generally is to produce intelligibility of one kind or another, we can recognize these alternative forms of explanation as producing different kinds of intelligibility and understanding in each of the domains in which they are felt to be most appropriate.

In this introductory chapter, I argue that the history of philosophy provides a genuine *explanation* for the much-discussed resistance of consciousness to contemporary structuralist (primarily, physicalist and functionalist) accounts, and that this explanation, if properly understood, could help to bring the contemporary debate to a greater level of methodological richness and sophistication. Historical analysis of concepts is a species of conceptual analysis, and conceptual analysis explains by revealing the underlying conceptual determinants of

patterns of use and description. By unearthing and evaluating the original arguments made for positions that have played a determinative role in structuring our contemporary concepts, historical investigation can remind contemporary philosophers of the original reasons for using concepts of mind and explanation in the ways that we do today. This points the way to a richer and more fruitful discussion, by recommending an explicit reconsideration of these often-forgotten or obscured reasons. Thus conceived, the historical explanation for the intractability of consciousness to physicalist description does not stand in any deep tension with other, more usual explanations for the problem – for instance, that consciousness fails to supervene on the physical or that there is an explanatory gap between our concepts of the physical and our concepts of consciousness.[7] Instead, it contributes to the clarification of these and other descriptions of the problem by clarifying the concepts of consciousness and explanation that they involve.

I

In order to begin to cast the light of historical interpretation on the contemporary discussion of consciousness, it is reasonable to investigate the origin and descent of the interrelated network of concepts that we use to characterize consciousness and the philosophical issues surrounding it. We can make an illuminating beginning by considering the concept of *qualia*. It is in the form of the question of qualia that many investigators today address the question of the explainability of consciousness. In the contemporary literature, qualia are variously thought to be incapable of physicalist or functionalist explanation, resistant to (but capable of) physicalist or functionalist explanation, or, owing to the unclarity or theoretical uselessness of the concept, nonexistent.[8] Argument about the explainability of consciousness, indeed, in many cases amounts simply to argument about the meaning of this concept. Significantly, though, the concept itself has a lengthy and seldom-explored lineage in the discourse of analytic philosophy. Investigation of this lineage provides insight into the philosophical sources of the main features and uses of its contemporary version. The full story of the descent of the concept of "qualia" in the twentieth century would require a detailed study of its own. But the outlines

of an explanation for some of the most significant contemporary uses of the term can already be drawn from an examination of some of the earliest uses of the term in the philosophical discourse.

The philosophical uses of the term "qualia" (and the singular "quale") in English trace back at least as far as the writings of C. S. Peirce, who used the term as early as 1867 to describe the immediate or given elements of experience. For Peirce, qualia (often used as cognate to "qualities") were already the most basic constituents of the totality of sensory experience, the ground of what he called Firstness or immediacy.[9] Drawing on Peirce, William James used the term beginning in the 1870s to denote the "irreducible data" of perception, for instance, the whiteness that is one and the same when I perceive it in today's snow and yesterday's white cloud.[10] These items, James argues, are the same no matter where in experience they occur; and they comprise an irreducible set of posits that must, perhaps along with the atoms of physics, be ultimate philosophical data. James's qualia, accordingly, set an utmost limit to the philosopher's project of analysis or rational inquiry, a limit beyond which only speculation can pass.

The most direct early influence on the contemporary debate, though, runs from the epistemology of the phenomenalist pragmatist C. I. Lewis. In the context of his attempt to distinguish the "given element in experience" from the interpretive element placed upon it by conceptual reasoning, Lewis was among the first to use the term "qualia" in substantially the same way it is used by theorists today:

> Qualia are subjective; they have no names in ordinary discourse but are indicated by some circumlocution such as 'looks like'; they are ineffable, since they might be different in two minds with no possibility of discovering that fact and no necessary inconvenience to our knowledge of objects or their properties. All that can be done to designate a quale is, so to speak, to locate it in experience, that is, to designate the conditions of its recurrence or other relations of it. Such location does not touch the quale itself; if one such could be lifted out of the network of its relations, in the total experience of the individual, and replaced by another, no social interest or interest of action would be affected by such substitution. What is essential for understanding and communication is not the quale as such but that pattern of its stable relations in experience which is what is implicitly predicated when it is taken as the sign of an objective property.[11]

Writing in 1929, Lewis already grants qualia the essential properties of immediacy, subjectivity, and ineffability that often characterize them

today. In the context of his reasoning about the properties of qualia, contemporary arguments for their existence and properties would be quite at home. As they were for James and Peirce, qualia are, for Lewis, the raw material or underlying substance of our rich and conceptually articulated experience of the world. But for Lewis, qualia are also explicitly *private* items. The ineffability of a particular quale outside its behavioral and relational context means that it is, outside this context, in a certain sense particular to its owner. No one else can possess or even understand my quale itself, for there is no way that I can communicate its intrinsic character to another. All that I can communicate is its place in the global pattern of relations that stands as its only objective sign.

There is also, though, an important contextual difference between the way in which Lewis uses the term "qualia" and its use in most of today's discussions. For instead of basing his conception of qualia on general intuitions or demonstrative thought experiments, Lewis articulates his conception of qualia from within the constraints of his global project of reconstructive analytic epistemology. For Lewis, qualia are the end points of epistemologically illuminating analysis. With their exhibition, we complete our analysis of any complex experience by distinguishing clearly between its interpretive, conceptual elements and that part of the experience that is genuinely "given," immediate, noninterpretive, and unconstrained by conceptual categorization. Aside from their role in this epistemological project, qualia have little significance. Indeed, Lewis says, they are abstractions, for our given experiences always come to us structured and formed, and their elements can be determined only by a process of analysis.

The setting of Lewis's concept of qualia within the larger theoretical project of reconstructive epistemology has historically important consequences for his use, and subsequent uses, of the concept. One consequence is that Lewis's notion of qualia has explicit *semantic* implications that contemporary uses of the concept usually lack. For Lewis ties conceptual interpretation to meaningful expression; it is only by conceptually interpreting a "given" element of experience that we gain the ability to communicate *about* that experience.[12] Consequently, Lewis's qualia are strictly indescribable. Strictly speaking, there is no possibility of describing an isolated quale, and there is no language for expressing the properties of individual qualia out of the context of their relationships with other qualia and conceptual interpretation. It

is these patterns of relationship that we do in fact communicate about when we discuss qualia. About the qualia *themselves* we can say nothing, even though we may continually exhibit them to ourselves.[13]

Nor can we, according to Lewis, even *conceive* of an isolated quale. It is ultimately to a *relational* description – a description of their place in relation to a total network of other qualia, external causes, and behavioral effects – that all thought about qualia must relate.[14] For Lewis, then, qualia are real but indescribable, except insofar as we can locate them within a relational structure. It is only in virtue of the quale's having a particular place in a total pattern of relations that it can be referred to at all. Thus, Lewis makes qualia linguistically identifiable only by reference to their positions within a complex relational structure, whose relata we are in no position to characterize independently of that structure.

II

Lewis's conception of qualia as describable only in virtue of the network of their relations, and indescribable in themselves, may at first seem quite uncongenial to contemporary uses of the notion. But even if this implication of indescribability is not always present in contemporary uses of the concept of qualia, the notion of qualia as primary contents set off against a total network of relations nevertheless bears direct relevance to the contemporary discussion of the problem of consciousness. The image of Lewis's original distinction between content and structure appears in David Chalmers's 1996 description of the root of the problem of explaining consciousness physically:

> Physical explanation is well suited to the explanation of structure and of function. Structural properties and functional properties can be straightforwardly entailed by a low-level physical story, and so are clearly apt for reductive explanation. And almost all the high-level phenomena that we need to explain ultimately come down to structure or function: think of the explanation of waterfalls, planets, digestion, reproduction, language. But the explanation of consciousness is not just a matter of explaining structure and function. Once we have explained all the physical structure in the vicinity of the brain, and we have explained how all the various brain functions are performed, there is a further sort of explanandum: consciousness itself. Why should all this

structure and function give rise to experiences? The story about the physical processes does not say.[15]

Chalmers's complaint articulates a picture of the underlying difficulty with the explanation of qualia that will be recognizable even to those who disagree with it. Accordingly, it is reasonable to begin with this consensus in seeking a historically minded account of the problem. Most importantly for the historical analysis, Chalmers's description of the problem turns on a central distinction between physical description, conceived as exclusively structural and functional, and basic experiences or qualia, conceived as resistant to this sort of description.[16] There is, Chalmers suggests, something direct and immediate about consciousness, something that makes it resist description in terms of structural relationships of concepts and functional relations of properties. It is in these terms, and according to these intuitions, that Chalmers goes on to describe the problem of consciousness as the "hard problem" of explaining the arising of *experience*, distinguishing this problem from the various "easy problems" of psychological explanation, all of which amount to problems of structural or functional explanation.[17] Consciousness is resistant to these kinds of explanation precisely because it is something different, something whose immediacy and directness will not be explained even when *all* the functions and structures in the world are accounted for.

Chalmers's intuition of the simplicity, directness, and immediacy of qualia characterizes both contemporary and older uses of the term. But along with this conception of qualia, Chalmers also gestures toward a conception of scientific explanation that is, in broad terms, shared by physicalists and antiphysicalists in the philosophy of mind. In particular, Chalmers conceives of the realm of physicalist (and, in general, scientific) explanation as a realm of *structural* and *functional* explanation, and he protests that such explanation does not suffice to explain the arising of consciousness. In so doing, he exploits a general conception of the metaphysical structure of the world that is congenial to physicalism and held in common by a variety of contemporary theories and theorists. According to this picture – what Jaegwon Kim has called the "layered model" of the world – reality consists ultimately of elementary particles, or of whatever basic units of matter our best physics tells us everything else is composed of, in *causal* relationships

to one another.[18] Accordingly, higher-level entities such as molecules and cells are arrangements of the underlying units, and their properties can be deduced (at least in an idealized sense) from the relations of the underlying units. This makes for a unified *logical* structure of explanation in which all of the causally relevant properties of entities described by the specialized sciences, including psychology, can, in principle, be explained in terms of, or reduced to, relational properties of the underlying units. This logical structure of explanation makes physicalist description essentially relational, for the explanation of a phenomenon adverts either to its compositional relationship to its constituents or to its causal or functional relationships with other phenomena.[19] Given this picture, a characterization of the structural and functional position of a phenomenon is all that the physicalist description has to offer. Reference to nonstructural or nonfunctional intrinsic properties plays no role.

In the underlying motivations of this picture of the world can be sought the underlying motivations of the contemporary discussion of consciousness as a problem for scientific description. The broadly physicalist picture, though, itself has a detailed and important philosophical history; and significantly, this history is not completely distinct from the history of the concept of consciousness to which Chalmers appeals. Historical analysis and reflection reveals the extent to which the conception of consciousness as inexplicable by structural or functional means, and the conception of those means themselves as presupposed in the current discussions, are joined in their origin and philosophical foundations. The philosophical history of the underlying distinction between basic elements of experience and structural or functional description can, in fact, be traced to one of the founding texts of analytic philosophy, Carnap's *Der Logische Aufbau der Welt*:

> Now, the fundamental thesis of construction theory (cf. s 4), which we will attempt to demonstrate in the following investigation, asserts that fundamentally there is only one object domain and that each scientific statement is about the objects in this domain. Thus, it becomes unnecessary to indicate for each statement the object domain, and the result is that *each scientific statement can in principle be so transformed that it is nothing but a structure statement.* But the transformation is not only possible, it is imperative. For science wants to speak about what is objective, and whatever does not belong to the structure but to