CHAPTER 1

Cognitive Kinds

O the mind, mind has mountains; cliffs of fall
Frightful, sheer, no-man-fathomed.

– Gerard Manley Hopkins, "No worst, there is none. Pitched past pitch of grief."

Is there no way out of the mind?

– Sylvia Plath, "Apprehensions"

1.1 Introduction

What is the landscape of the mind? That is the question I aim to tackle in this book. This is an inquiry into the basic components of our mental makeup: What kinds of objects, states, capacities, events, processes, and other entities constitute the stuff of our mental life? As the book's title indicates, the scope is not the mental in general, but the cognitive realm in particular, which I take to be a subset of the mental or psychological realm. Although I will not attempt to demarcate the limits of the cognitive in detail, in what follows, I will attempt to say what characterizes cognitive phenomena, as opposed to other aspects of the mind and brain, later in this chapter (see Section 1.5, as well as Section 2.6). The inquiry is grounded partly in metaphysics and ontology, the philosophical investigation of the building blocks of the universe, and partly in the sciences, empirical research into the workings of the human mind. Since this is a book written by a philosopher, the latter is represented not in the form of original research but by means of distillations of recent empirical work on various mental items and an attempt to synthesize empirical work from different disciplines and subdisciplines. Integrating this empirical work with philosophical argumentation requires paying attention to the relevant literature in cognitive science, including psychology in its various branches (cognitive, developmental, social, and so on), linguistics, neuroscience, computer science, and related disciplines. Given the voluminous amount

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

of work in these areas, it may seem presumptuous to take it all in, and I certainly do not aim to give a comprehensive account of the mental landscape. Instead, I plan to focus on a small number of paradigmatic cases. Of course, this type of integrative project also requires the careful philosophical work of making distinctions, clarifying concepts, and justifying claims with arguments. In this introductory chapter, I intend to lay out some of the philosophical groundwork that supports the argumentation that follows in later chapters. In particular, I plan to spell out the approach to ontology that I intend to take, and specifically the account of categories and kinds that I will adopt, which is naturalist, non-reductionist, and realist (as I will go on to explain).

Inevitably, when one investigates the mind these days, the brain is never far behind. Some would say that the entities constituting the mind are none other than those that comprise the brain, and that we are well on our way to discovering what these are. But despite the fact that there is indeed an intimate connection between psychological and neural entities, I will try to provide reasons for thinking that they are not one and the same and that the categories that pertain to one may not apply to the other. Though the focus will be on mental or psychological entities, their connections and relations to neural entities will often be invoked. To anticipate somewhat, one of the main themes of this book is that there is not always an identity – whether type or token – between psychological and neural constructs, and furthermore, that the validity of a psychological construct does not reside in its coincidence with a neural structure, mechanism, or process. In the neurosciences, there is currently considerable debate and a notable absence of consensus about how mental and neural entities relate to one another. Neuroscientists run the gamut, from those who advocate extreme reductionist positions that posit a “grandmother cell” (see Gross 2002) or a “Jennifer Aniston neuron” (Quian Quiroga, Reddy, Kreiman, et al. 2005) and locate cognitive functions (even particular concepts) in specific brain regions or populations of neurons, to those who preach anti-reductionism and excoriate “blobology,” the alleged identification of areas of neural activation with particular psychological capacities, primarily based on regions identified by neuroimaging technology (Poldrack 2012). In subsequent chapters, I will try to provide reasons for thinking that though we are finding and will continue to find many significant correlations between brain structure and cognitive or psychological function, we should not expect a wholesale identification of one with the other. Indeed, I will argue that we will not always be able to identify psychological functions with neural activity,

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

whether or not this activity is localized in specific neural structures. As I mentioned in the previous paragraph, the emphasis will be on *cognitive* ontology rather than psychological ontology more broadly. All the case studies to be discussed involve cognition in some way, as opposed to affective, perceptual, sensory, or experiential aspects of mentation. The aim is not to give an exhaustive catalogue of the contents of the mind (if that were even possible) but rather to focus on a range of significant examples of categories that involve cognition, examine the case for admitting each into our ontology, and draw some general conclusions about the kinds of entities that we should posit in cognitive science and on the grounds for doing so. After this first programmatic chapter, each of the rest of the book's chapters tackles one or a small number of candidates.

1.2 Naturalism about Kinds

In investigating mental objects, states, capacities, events, processes, and other entities, we are usually investigating types not tokens, that is, not unique particulars, but types or *kinds* of them. Specifically, we are interested in which of these types or kinds are real or "natural," or in standard philosophical parlance: *natural kinds*. Many contemporary discussions of natural kinds base their notion of kinds on the essentialist account first sketched out by Putnam (1975) and Kripke (1980). Instead, I will anchor the account of kinds that I will be deploying throughout this book in a nineteenth-century tradition that is more closely aligned with a naturalist philosophical outlook. According to the naturalist tradition that I will be tapping into, empirical science is our best guide to the kinds that exist in nature, rather than a priori considerations from metaphysics or philosophy of language. This attitude originates with the discussion of scientific classification that is prominently represented in the works of Whewell and Mill, and indeed in their mutual influence. Though Mill is often credited with initiating the discussion of natural kinds (or just plain "kinds," as he called them) in modern philosophy, even a casual reader of Mill's *A System of Logic* (1843/1882) cannot help but notice the considerable debt to Whewell's *Philosophy of the Inductive Sciences* (1840/1847). Despite significant differences in their overall philosophical positions, Whewell being a neo-Kantian rationalist and Mill a staunch empiricist, there is much that they agree on when it comes to kinds. Whewell and Mill both regard science as the guide to uncovering kinds in nature and think that scientific taxonomy aims at discovering kinds. Moreover, they are both concerned with the rational grounds for scientific classification and are keen

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

to understand the differences between "natural" and "arbitrary" scientific classification schemes. They both see kinds as the basis for inductive inference and regard science's search for kinds as a quest to come up with categories that would lend themselves to empirical generalizations and natural laws. As Whewell writes: "The object of a scientific Classification is to enable us to enunciate scientific truths: we must therefore classify according to those resemblances of objects … which bring to light such truths" (1840/1847, 486). Whewell also thinks that classification must not be based on any resemblances whatsoever but on what he calls "natural affinity," which requires us to classify things on the basis of properties that generally cooccur with other properties (1840/1847, 542). Moreover, he repeatedly states that "the great rule of all classification" is that "the classification must serve to assert general propositions" (1840/1847, 495). Mill endorses this emphasis on "general propositions" or "general assertions" and goes on to say that "the very first principle of natural classification is that of forming the classes so that the objects composing each may have the greatest number of properties in common" (1843/1882, 879). Hence, for both Whewell and Mill, the aim of scientific classification is to group things together based on shared cooccurring properties, so that the categories that result enable us to make valid scientific generalizations.

While the naturalist tradition that originates with Whewell and Mill provides the main philosophical inspiration for the account of kinds that I will be operating with in this book, there is one respect in which I will part company with this older tradition. These philosophers are not very clear when it comes to the metaphysics of kinds. They seem to think that uniformities in nature are the basis for successful scientific generalization and inference, but they do not fully explicate the nature of these uniformities. Venn (1889/1907) criticizes Mill for distinguishing between two kinds of uniformity in nature: uniformities of sequence (which are causal) and uniformities of coexistence (which are brute). By contrast, Venn thinks that many of the uniformities of coexistence identified by Mill are actually causal in nature (though he does not think that *all* uniformities in nature are causally based). Still, he holds that uniformities are what enable us to use natural kinds in inductive inference in science. According to Venn (1889/1907, 94), uniformity "is the objective counterpart or foundation of inferribility …." Inductive inferences are based on uniformities and are therefore dependent on the existence of kinds in nature, which reflect these uniformities. Thus far, I agree with Whewell, Mill, and Venn. But by contrast with them, I will assume that uniformities in nature are due to regular and stable connections between causes and effects, and that these

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

causal relations are the metaphysical bases of scientific induction and epistemic practices.[1] This assumption is also shared by many contemporary naturalist philosophers. It is prominent in Boyd's account of natural kinds and it is exemplified in what he calls the "accommodation thesis": "Kinds useful for induction or explanation must always 'cut the world at its joints' in this sense: successful induction and explanation always require that we accommodate our categories to the causal structure of the world" (1991, 139). Boyd also speaks of "the accommodation of inferential practices to relevant causal structures" (2000, 56).[2] This is also a central feature of Kornblith's (1993, 35) account of natural kinds: "It is precisely because the world has the causal structure required for the existence of natural kinds that inductive knowledge is even possible." This link between the epistemology of categories and the ontology of kinds is characteristic of a naturalist attitude to metaphysics, which holds that our metaphysical inquiries should be guided by our best epistemic practices as exemplified in the considered classification schemes of our best scientific theories. Among at least some contemporary naturalist philosophers, the causal structure of the world is the ontological basis for the successful epistemic practices of science.

Contemporary naturalist philosophers think that the causal uniformities in nature, even those discovered by the basic sciences, are rarely if ever ironclad or exceptionless, and this implies that the properties associated with natural kinds are loosely clustered rather than invariably associated with one another. Moreover, as I have already mentioned, the properties that cluster in kinds are not just sets of properties that happen to cling together, since they are associated as a result of causality. Accordingly, rather than view kinds as mere clusters of properties, I have proposed that they be conceived as "nodes in causal networks" (Khalidi 2013; 2018). According to this "simple causal theory" of natural kinds (cf. Craver 2009), certain properties or conjunctions of properties that are causally connected with others in systematic ways can be considered natural kinds. Sometimes we identify the kind with just one of the properties in a causal chain or network,

[1] There may be some uniformities in nature that are brute and not causally based, particularly at the most fundamental level. But I will assume that these are not at issue in a discussion of cognitive ontology. For further justification, see Khalidi (2013; 2018).

[2] Elsewhere in the same paper, Boyd emphasizes the ways in which natural kinds are "practice-dependent" and relative to human interests, and it is not easy to reconcile this attitude with his accommodation thesis. On the view that I favor, human interests serve only to select certain causal structures and processes to focus on, they do not somehow shape or modify them (except in cases in which humans are themselves part of the causal process – see Section 1.5).

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

but at other times we draw a wider circle among a number of them and consider that set of properties to be the kind. Either way, we are identifying properties that are causally conjoined to others, rather than mere clusters of properties. This causal account of kinds is somewhat less restrictive than that proposed by Boyd, who considers kinds to be property clusters that are held in homeostasis by causal mechanisms – though he sometimes relaxes these conditions and gestures toward something like a simple causal account. Thus, the simple causal account is distinct from a strict version of Boyd's account, which requires a specific causal mechanism to keep the cluster of properties in equilibrium (or homeostasis). I have questioned the strict version on two grounds. First, in many cases, there is nothing that can properly be called a causal mechanism that holds the properties together – they may instead be held together functionally or relationally, as we shall see in later chapters. Second, the properties involved are not always in a state of equilibrium – they may be repeatedly instantiated through the action of independent causes.[3] A simple causal theory of kinds can also be usefully distinguished from an essentialist one, at least on many versions of essentialism. Though essentialists also tend to think that natural kinds are discoverable by science, they usually place additional conditions on natural kinds, which I think are at odds with scientific taxonomy. There are four ways in which this account of kinds differs from many essentialist ones. First, the properties that are associated with each kind are causally linked, but they can consist in a loose cluster rather than a set of properties that are both necessary and sufficient for kind membership. Second, the causal properties may be functional or relational rather than intrinsic. Third, the properties involved do not have to be microstructural, as some essentialist philosophers tend to insist. Fourth, the simple causal theory does not claim that these properties are associated with the kind in question across possible worlds or with modal necessity, as essentialists usually hold.

Another significant point of agreement in the naturalist tradition that stems from Whewell, Mill, and Venn is that natural classification schemes and the kinds that they identify can be found across the sciences, including the human sciences. These philosophers tend to see considerable continuity from chemistry to mineralogy to biology to psychology and the social sciences, especially when it comes to the importance and feasibility of uncovering kinds. This attitude seems less prevalent among contemporary

[3] These claims are further justified in Khalidi (2013; 2018). I have also proposed that natural kinds can be represented by means of directed causal graphs. Although I have not worked out this proposal in detail, in such representations, natural kinds correspond to highly connected vertices in directed causal graphs.

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

philosophers, at least some of whom think that it is a truism that natural kinds pertain to the natural sciences. Hence, it may appear oxymoronic to talk about natural kinds in the cognitive sciences. Given that the terminology of "natural kinds" is misleading, especially in the context of the human sciences, I will be talking mainly of "kinds" or "real kinds" instead of "natural kinds," especially given that the very existence of the expression "natural kind" seems to be a historical accident. As Hacking (1991) has pointed out, the terminology of "kinds," which Whewell and Mill used, gave way to "natural kinds" as a result of the writings of Venn. But Venn seems to have taken himself to be using Mill's expression, since he credits him with introducing the term – despite the fact that Mill apparently never used it. Venn (1889/1907, 84) writes: "he [i.e. Mill] introduced the technical term of 'natural kinds' to express such classes as these." It is unclear whether Venn simply misremembered Mill's terminology or whether he deliberately modified it. Either way, we are now saddled with an unfortunate expression, which is misleading on at least two counts. The first reason that the expression "natural kind" is deceptive is that it tends to set up a misguided contrast between the natural and *artificial*. In many scientific domains, there are strong candidates for kinds that have the "trail of the human serpent" over them and may reasonably be considered artificial (especially in the Anthropocene era). Whether we are dealing with synthetic chemicals, genetically engineered organisms, or artificially intelligent systems, scientists now study a range of entities that are the result of human intervention (if not wholesale invention), yet apparently no less real or objective than their supposedly "natural" counterparts. But the terminology of "natural kinds" would encourage us to dismiss the kinds to which these entities belong. The second reason the expression is misleading (which is more important for these purposes) is because the adjective "natural" suggests an affiliation with the natural rather than the social sciences, and it threatens to sideline categories that have a social or human dimension. When it comes to the cognitive sciences, which straddle the biological and psychological sciences, this is especially pernicious, since it tends to privilege the former over the latter, perhaps suggesting that neural kinds are more objective than psychological ones.

Here, it may be objected that the philosophical apparatus of real kinds may not be the right lens through which to view cognitive science. It may be thought that kinds are more at home in sciences like botany or mineralogy, where the paradigmatic individuals are well-defined concrete particulars (individual plants, mineral samples), with clear spatiotemporal boundaries. In cognitive science, though there are some fairly neat individuals such

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

as human persons (and other creatures), which are often classified into kinds (e.g. *schizophrenic*, *bilingual*), the individuals can also be cognitive modules, cognitive capacities, mental states, mental processes, and other entities, so it may not be as useful to think of such entities as belonging to kinds. I would reply simply by stressing the indispensability of taxonomy to any scientific discipline or subdiscipline. Whenever we theorize about any domain, it is inevitable that we classify the items that populate that domain and that we do so in nonarbitrary ways. Classification, in turn, presupposes dividing a domain of entities into types or kinds. Moreover, as I will try to show, although some of the items classified in cognitive science are not best thought of as individuals, but states, capacities, events, processes, and so on, they are also divisible into kinds. Hence, there is no need to think of classification as pertaining exclusively to a domain in which concrete particulars with well-defined spatial boundaries are the main items of interest.

This brief sketch of a naturalist theory of kinds and its underlying metaphysics will have to suffice for now. More details will emerge as we survey a number of candidates for cognitive kinds in subsequent chapters.

1.3 Ontological Matters

In recent philosophical and scientific discussions of cognitive ontology, it is common to read that "ontology" is used differently by philosophers and others, namely psychologists, neuroscientists, and perhaps most prominently, computer scientists. I believe that this claim is not wholly justified. There are perhaps some differences in emphasis and nuance in the usage of these disciplines, but this is not a case of sheer polysemy. The main difference may be that computer scientists (in particular) are interested in how domains are taxonomized without great regard to how they *ought* to be taxonomized, and without a commitment to the domain's actually containing the entities that are posited by the taxonomic or classificatory system. Philosophers, on the other hand, tend to be interested in the *ought* and in the underlying structure of reality. As emphasized in Section 1.2, naturalist philosophers tend to think that our current, mature, scientific taxonomic systems are our best (defeasible) guides to that underlying structure. In other words, they derive an *ought* from an *is*.[4] This is warranted on the

[4] This is a stark and provocative way of putting it. For a more nuanced account of the relationship between scientific practice and philosophical theory, see Khalidi (2013), where I lean on the notion of "reflective equilibrium," first introduced by Goodman (1954/1979).

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

assumption that science aims at discerning that structure. In later chapters, we will encounter challenges to that assumption, on the grounds that some investigators are not just aiming to discern the causal structure of the world, since their inquiries are shaped by non-epistemic norms (especially in areas like psychiatry). I will put such concerns to the side for now, and will take them up in some subsequent chapters (but see also Section 1.5).

If we are naturalists, then talk of "ontology" is closely related to talk of "taxonomy" or "classification" – provided we think that science aims primarily at classifying entities in such a way as to discern the causal structure of the world, and is guided in doing so by epistemic goals. When viewed thus, there does not seem to be an equivocation or ambiguity in the use of the term "ontology" and related expressions. If we bear in mind that "ontology" should not be used as a synonym for "taxonomy" or "classification scheme," but rather to denote the metaphysical structure that is described by a taxonomic system or classificatory scheme, then some of the differences in usage can be cleared up. This caveat is also relevant to the use of terms like "kind" and "category." These two terms (and related ones) are often used interchangeably, by philosophers and cognitive scientists alike, but I propose to distinguish them, as follows. A *kind* should be understood to be an entity in the world, which can be conceived of as a collection of particulars or set of entities (nominalist reading), or an abstraction, such as a universal that is immanent in particulars (realist reading). Meanwhile, a *category* pertains to our conceptual, theoretical, or linguistic framework and practices; it is the concept of a kind.[5] In other words, a kind pertains to ontology whereas a category pertains to taxonomy. Here again, the two notions are closely related, since (on a naturalist understanding) the aim of scientific inquiry is to devise categories that correspond to all and only the kinds.

Once we distinguish ontology and taxonomy, along with kinds and categories, we should take care not to embrace a view that has been derided as the "third dogma of empiricism" (Davidson 1973). According to this dogma, we can somehow confront our kinds with our categories directly to determine whether they are in alignment, as we might compare a map of the landscape with the terrain itself for accuracy. The problem with this way of thinking is that we have no access to the "terrain" that is not mediated by our "map" (which is why the cartographic analogy is so misleading). We access the world via the categories of our taxonomies and hence, we cannot step outside of them to see how well they align with the

[5] To mark this distinction, I will generally italicize *kinds* and put CONCEPTS in small caps.

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

world itself. But that does not mean that we have no way of determining whether and how well our categories delineate the kinds. Since we devise these categories to describe the world, we can determine how well they enable us to generalize, explain, predict, and so on. Depending on their efficacy in fulfilling our epistemic goals, we infer that they have or have not latched on to the causal structure of the world, in line with the naturalist picture outlined in Section 1.2. This view has been articulated lucidly by Child (2001, 38), who writes: "in classifying things by reference to their causal powers, or their causally significant composition, we classify things in ways that reveal the way the world works." As long as human inquiry is able to achieve this goal, the specification of an ontology is not beyond our reach.

I argued in Section 1.2 that classification and the identification of kinds is based on identifying properties that are associated with other properties, and these properties are so identified because they are causally related. But what about those properties themselves? How do we identify the most basic properties in our ontology, and might we have settled on a different system of kinds if we had started with a different set of properties? This is an old philosophical conundrum and I cannot pretend to give a satisfactory answer to the question in the scope of this book (see e.g. Goodman 1954/1979; Lewis 1983). It is true that real kinds are grounded in shared properties and that these properties may be considered the unjustified posits upon which the whole theoretical edifice is built. If the properties to which we humans are attuned are just reflections of our parochial perceptual and cognitive abilities and do not reveal real features of the universe, then you might say that we have no reason to believe that the kinds that we identify expose the real joints in nature. In cognitive science, such properties might include basic behavioral ones involving motion, force, space, and time (e.g. eye movements, button presses, looking times, reaction times) or more abstract intentional ones (e.g. expressed preferences, discrimination between stimuli). But there may be a way of overcoming certain skeptical doubts about these baseline properties. For there is an indirect vindication of our choice of baseline properties in the identification of kinds that enable us to make generalizations, which in turn help us to explain and predict the entities in question. As I argued in the previous section, these epistemic desiderata are themselves causally based, so the choice of properties is ultimately upheld by our ability to use them to understand the causal structure of the world. Unless causality is itself an illusion, or a mere reflection of our inadequate and distorted perceptual and cognitive endowments, our choice of properties in cognitive science is

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

at least indirectly supported by the fact that the kinds that they constitute allow us to discern the causal structure of the mind.[6]

Here, I should also emphasize that on the ontological framework I am adopting, there is no strict divide between properties and kinds. Although kinds are usefully thought of as loose collections or clusters of properties, some of the properties that we take to be associated with kinds can be further decomposed into or identified with more basic properties (perhaps also in the sense of loose clusters). Some properties are best thought of as complex ones that can be understood in terms of more basic properties (e.g. in physics, the property of being an electron decomposes into other properties, namely having a certain mass, charge, and spin). It is true that this way of thinking about properties and kinds purports to justify them with reference to each other, but circularity of this sort seems unavoidable in defending our most basic choice of parameters and dimensions. Moreover, it is evident from reflection on the history of science that the properties and kinds that we identify on the basis of preliminary observations and investigations are not always those that persist after subsequent inquiry. Therefore, it is not as though we are locked into our initial choice of properties no matter how misguided. This may not be enough to assuage skeptics or anti-realists, but I will have more to say about realism in the following section.

There is a further ontological question about properties and kinds, briefly alluded to earlier, that I would like to bracket as far as possible in the context of this book. That has to do with the traditional metaphysical debate between nominalism and realism, which pertains to the underlying metaphysical reality of properties and kinds. Are they universals (metaphysical realism) or collections of individuals (nominalism)? Although I am committed to the existence of properties and kinds, I am not wedded to a particular metaphysical understanding of them. They may be best identified with universals, whether immanent or transcendent, or they might be better understood as sets or collections of particulars. Moreover, if they are identified with universals, it may be that kinds are universals in their own right in addition to properties, or they might just be concatenations of property universals. We need not resolve these questions for the

[6] What if causality itself is not objective? Causal connections are the holy grail in science, as witnessed by the fact that scientists are generally intent on disentangling causation from correlation. I am therefore assuming that causal connections are something like the skeletal frame of reality. Moreover, if there is no *unique* causal relation, but a plurality of causal relations (see e.g. Godfrey-Smith 2010), then there would be no single metaphysical basis for real kinds but a plurality. I will set this possibility aside for the purposes of this book.

purposes of this work, since I doubt that it matters much which position we adopt when it comes to cognitive ontology, just as long as we acknowledge the existence of properties and kinds.

Finally, on the topic of general ontology, I will not try to settle a question as to the identity of the broadest ontological categories. I have referred already to such entities as individuals (or objects), states, processes, events, capacities, and mechanisms, and have indicated the possibility of the existence of others. Such posits are sometimes considered the basic "categories" of ontology. Some philosophers posit a more austere basic ontology. For example, Lowe (2006) posits a "four-category ontology," in which the categories are individual substances (objects), substantial universals (kinds), property or relation instances (modes), and non-substantial universals (properties or relations). Other philosophers think there are other ontological categories that are not equivalent to one of these four, such as processes or events. I will not try to settle these metaphysical disputes, since the focus here is on the applied question of cognitive ontology. In the cognitive sciences, many entities are best conceived of as *individuals* (e.g. person), others as *states* (e.g. belief, pain), others as *processes* (e.g. learning, episodic memory retrieval), and yet others as *capacities* (e.g. language, semantic memory, mindreading). For other cognitive entities, there may be some uncertainty as to which broader ontological category they fit into. For example, we might well wonder whether *fear* is best understood as a *state* or a *process*. In some cases, there may be closely related entities that fit into more than one ontological category. For example, *episodic memory* is sometimes thought of as a *state*, *process*, or *capacity*, or indeed, all of the above – in which case the term "episodic memory" would be equivocal (for discussion, see Section 5.2). These questions will be explicitly addressed when it comes to some of the cognitive entities to be discussed in later chapters. In this book, it will emerge that the broader ontological categories that seem indispensable to cognition are those already mentioned: *individual*, *state*, *process*, and *capacity*. The cognitive kinds to be discussed in subsequent chapters all seem to belong to one of these four basic ontological categories. But I will not try to come up with an exhaustive list of broader categories comprising all cognitive entities. I will also not take a stand on whether some of these broader categories should be understood in terms of others, or whether some of them are more fundamental than others. For example, some metaphysicians would advocate reducing all processes to sequences of temporally ordered events, while others would consider processes to be fundamental and would argue for understanding events in terms of processes, while yet others would favor explicating all

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

other metaphysical categories in terms of processes (thus advocating what has been called a "process ontology"). Since this is not a work in general ontology, I will not try to resolve these issues or take a definite stand one way or the other. If there are theoretical and empirical reasons for considering a certain cognitive entity to be a process rather than a state, I will try to indicate them; if some entities can be understood in more than one way, I will try to justify that claim. Throughout, I will just assume that cognitive entities may fit into a number of different broader categories and will give reasons for considering the particular entities to be discussed to fit into one or more of these overarching categories (e.g. process or capacity), without attempting to give an exhaustive list of such categories or trying to determine which of them (if any) are fundamental.

1.4 Reductionism

There is a very strongly entrenched picture of the domains investigated by science that portrays them as constituting something like a strict hierarchy, from that of elementary particles at the bottom of the hierarchy to social entities like nation-states and economic markets at the top. The idea is that the universe is arranged in a number of layers or levels ranging from the smallest and most fundamental to the largest and most complex. The picture is sometimes referred to as the "layer-cake view," though it is perhaps better described as the "inverted pyramid view," to convey the idea that the more fundamental domains are more austere in terms of the number of entities they posit and the laws or regularities that govern them, while the less fundamental ones are more prolific, not to say profligate. This picture is sometimes traced back to a classic paper by Oppenheim and Putnam (1958), but its prevalence and appeal would seem to go beyond the influence of any one particular source. Although it is undeniably the case that the universe consists both of micro- and macro-entities and that the latter are composed of the former, this picture is misleading as an account of the most salient divisions in nature and the manner in which the different parts of the universe fit together.

What is wrong with the view that reality is arranged hierarchically in a series of domains, the entities of each domain being composed of those in the domain below it, and with the accompanying idea that the theories that purport to describe and explain each domain are reducible to those that describe the domain below it? There are at least three respects in which the view is highly misleading, or at least prejudicial to certain philosophical positions regarding the relationship between the sciences,

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

in particular those that study the mind and brain. First, it implies that the crucial dimension distinguishing scientific domains is that of size or spatiotemporal dimension. Second, the layer-cake picture gives the impression that there is a neat compositional hierarchy in which the domain studied by each scientific discipline or subdiscipline decomposes into and depends entirely upon the one immediately prior to it in the hierarchy. Third, and most importantly, the layer-cake picture encourages a reductionist account according to which the key to understanding each scientific domain lies in the domain that precedes it. It promotes the view that to understand what is really going on in chemistry, one must turn to physics, and to understand what is really happening in biology, one must revert to chemistry, and so on.[7] I will now put forward considerations against each of these claims in turn.

On the layer-cake view, the layers of the cake are often thought to correspond to spatial or spatiotemporal dimensions, with lower levels consisting of micro-entities that operate over very small timescales, while higher levels are composed of entities at increasingly larger spatiotemporal scales. However, the causal systems described by different scientific disciplines or subdisciplines span different spatiotemporal scales. The theories of fundamental physics describe elementary particles as well as galaxies (and many systems in between). To use an example from the cognitive sciences, if we are investigating the psychopharmacological effects of lithium compounds on stabilizing the moods and altering the behaviors of people suffering from mental illness, we may need to refer to the role of lithium ions in increasing the release of neurotransmitters at the molecular scale while also describing the behavior of individuals in macroscopic terms. The causally relevant processes in nature often occur in transverse sections that cut across spatiotemporal dimensions.

A second problem with the layer-cake view is that the domains of scientific theories or disciplines cannot be considered to be arranged in a strict mereological sequence or dependence hierarchy, since there is often only partial overlap between one domain and another and they cannot be conceived of as related compositionally (see Figure 1.1 for a crude attempt to illustrate the difference between the layer-cake view and an alternative picture). The inquiry just described, which relates to the

[7] In fact, Oppenheim and Putnam (1958, 15) explicitly support their reductionist picture by referring to the relation between psychology and neuroscience: "it has proved possible to advance more or less hypothetical explanations on the cellular level for such phenomena as association, memory, motivation, emotional disturbance, and some of the phenomena connected with learning, intelligence, and perception."

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

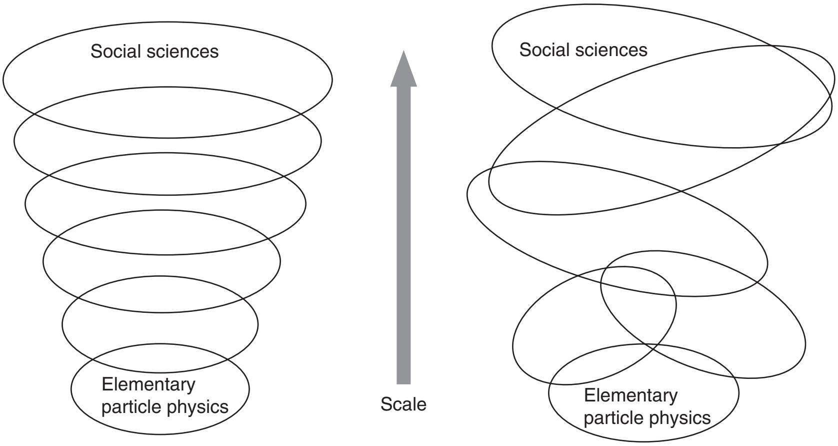

Figure 1.1. Alternative to the "layer-cake" view: The standard "layer-cake view" of scientific domains (left) can be contrasted with a picture (right) according to which scientific domains both crosscut one another and cut across spatiotemporal scales.

psychopharmacological effects of lithium, may be at cross purposes to another neuroscientific inquiry dealing with color processing in the visual system, for example. Both inquiries can be conceived of as investigating causal systems, but the systems involved are not wholly contained within one another. Even though in this case, both systems seem to depend entirely upon the neurobiology of the brain, I will argue later in this section that this impression is mistaken.[8] Rather than a series of ordered hierarchical levels, each of which depends on the next smaller level, an alternative picture is one in which the universe consists in a set of relatively "closed systems," each of which is causally integrated and somewhat causally cordoned off from others. The terminology of closed systems is derived from the physical sciences, but it should not be taken literally, since causal systems are not hermetically sealed. In the following chapters, I will loosely apply the expression to causally integrated systems in which the causal inputs are largely known and other external influences can be safely ignored or bracketed. Sometimes these coincide with scientific

[8] This does not just pertain to the relationship between the cognitive and neural domains. As noted by Rabin (2018), it is not clear how to order disciplines like geology and psychology in terms of dependence relations: both seem to depend on chemistry (and I would add, physics), but neither depends on the other. Similarly, Epstein (2009; 2015) argues that social facts do not depend entirely on facts about individual persons, or even those facts plus facts about their local environments and practices.

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

disciplines or subdisciplines, but at other times they comprise scientific research programs. This way of thinking of scientific domains is closely allied with a conception elaborated by Weiskopf (2017b), building partly on work by Simon (1969/1996) and Haugeland (1998). Weiskopf contrasts two views of the relationship between the complex systems described by scientific models:

> One way sees hierarchies mereologically, in terms of size and spatial containment relations, so that a system is decomposed into subsystems that are literally physical parts of it. … An alternative, however, is to define hierarchies in terms of the *interactional strength* of various components rather than their spatial relations. (2017b, 11–12; original emphasis)

My main (minor) dissent from these remarks consists of resisting the urge to posit a hierarchy at all. If interactional causal strength (and relative causal isolation) is what delimits a domain described by a scientific theory or model, there should be no expectation that scientific domains will constitute an ordered series or sequence, and hence there is no question of a hierarchical arrangement. This anti-hierarchical position also agrees with the attitude expressed by Woodward (2017, 40 n.4), who writes that he takes a "very deflationary" understanding of "levels," according to which "levels talk is just a way of expressing claims about explanatory or causal relevance and irrelevance," which "does not carry any suggestion that reality as a whole can be divided into 'layers' [or] levels on the basis of size or compositional relations"[9] Ideally, "levels" discourse should be entirely replaced with talk of causal systems or domains, but this discourse is entrenched in both philosophy and science and not easily expunged. I will use it occasionally in subsequent chapters, bearing in mind the caveats just mentioned, but I will usually replace it with talk of domains or causal systems.

The third problem with the layer-cake view is the association with reductionism. Classically, reductionism was understood in terms of a complete translation of the theoretical terms of one theory into the theoretical terms of another (Nagel 1961). But one need not hold reductionists to such a high standard. My objection is not just aimed at this classical view

[9] These views are also reminiscent of Wimsatt's prescient account of "levels" (first published in 1994). He writes that levels "*are constituted by families of entities usually of comparable size and dynamical properties, which characteristically interact primarily with one another, and which, taken together, give an apparent rough closure over a range of phenomena and regularities*" (2007, 204; original emphasis). Wimsatt conceives of levels largely along compositional lines, though he also argues for the existence of more complex and non-compositional "causal thickets," which I will mention in Section 1.5.

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

but to weaker notions of reductionism as well, which hold merely that lower-level explanations are more fundamental than higher-level ones, or that one has not fully explained or understood a particular domain unless one has explained it in terms of a lower-level domain. The criticisms I have made of the layer-cake view already suggest that we cannot expect that each domain will be wholly explainable in terms of a domain that is "lower" in the hierarchy, since the whole notion of a hierarchy is problematic. If one thinks of the domains of different theories or disciplines as being partially overlapping, depending on the causal integration and isolation of the entities under investigation, then that threatens not just a Nagelian reduction, but any form of reductionism that posits that each domain is best explained by the next lowest domain in the hierarchy. It might be objected that as long as there is one all-encompassing lowest domain, currently thought to be the domain of elementary particles, it should in principle be possible to reduce each domain directly to that fundamental base. But the reduction of special-science domains like cognitive science directly to fundamental physics is not a serious prospect.

In the rest of this section, I will try to bolster the case for denying even a weak version of reduction when it comes specifically to cognitive science. The usual way of resisting reductionism leans on the claims of multiple realizability and multiple realization. These topics have been extensively addressed in recent philosophical debates and I will not try to do justice to those debates. In brief, multiple realizability claims that cognitive and psychological categories are individuated differently from neural and biological ones. Since the former are typed in terms of their functions or causal roles but the latter are typed in terms of their structures or mechanistic properties, and since the same functions can be performed by mechanisms with very different structures, they cannot be type identical. The argument proceeds at a very abstract level based on what we know generally about functions and structures. Moreover, it concludes only that the kinds of psychology are multiply realiz*able* relative to neurobiology, not that they are actually multiply realiz*ed*. For all this argument claims, in our actual universe, the same psychological and cognitive functions are always achieved by the same neurobiological structures. The argument from multiple realization is subtly different. It relies on direct empirical evidence and bases its conclusion on purported cases in which the same cognitive state or process is subserved by different neural mechanisms. If one can point to actual cases in which the very same type of psychological function is performed (in different individuals, species, or systems, or indeed in the same individual on different occasions) by genuinely different types of

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

neurobiological (or other) structures, then that blocks a type–type reduction. In a way, multiple realizability is the stronger argument since it is supposed to apply regardless of the empirical evidence and cannot be refuted by showing that alleged cases of multiple realization do not in fact support the desired conclusion. But in another way, multiple realization is more powerful, since any genuine finding of different structures realizing the same function would seem to clinch the case and render the abstract argument superfluous. Usually, the two go together, since multiple realizability is a plausible explanation for multiple realization, and multiple realization provides corroborating evidence for multiple realizability. Though these theses have had convincing advocates, they have also had vocal opponents, who claim that neither thesis has been demonstrated. Both theses have been exhaustively debated in the philosophical literature and I have nothing to add to those debates except to endorse the arguments that have been made in favor of multiple realization and multiple realizability.[10] I will not try to reiterate them or advance them here because I think that one can go even further in an anti-reductionist direction, as I will try to explain. In the rest of this section, I will try to provide an argument as to why the cognitive domain is not likely to neatly decompose into the neural domain.

Recent advances in neuroimaging have opened the door for scientists to investigate the workings of the brain in real time, while experimental subjects are engaged in various cognitive tasks. At one point, this development held out the hope for assigning specific cognitive functions to brain regions in a fairly direct and straightforward way. But despite the fact that the past few decades have revealed a great deal about the workings of the brain and how it serves to implement the functions of the mind, there is mounting evidence that the links between neural structure and cognitive function are rather more circuitous and indirect than has often been assumed. At least as things currently stand, the vaunted "structure-to-function" mapping is

[10] For a comprehensive defense of multiple realization using examples from vision science, see Aizawa (2017), where he argues that multiple realization results in a kind of autonomy for higher-level taxonomies, in the sense that there need not be an isomorphism between higher- and lower-level taxonomies. He also argues that this is compatible with the idea that there are important interactions between higher- and lower-level sciences, which constrain the development of their respective taxonomies. Multiple realization and realizability have also been vigorously disputed by some philosophers, for example, Polger (2002; 2009) and Shapiro (2004), but I will not try to address their arguments directly. Sober (1999) is also sometimes interpreted as an argument against multiple realizability, but his position seems more nuanced. He writes: "The reductionist claim that lower-level explanations are *always* better and the anti-reductionist claim that they are *always* worse are both mistaken" (Sober 1999, 560; original emphasis). But I do not take the anti-reductionist import of multiple realizability to be that lower-level explanations are always worse, just that they sometimes capture causal patterns that are not captured at the lower levels.

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

far from a one-to-one correspondence. To pick one example out of a hat, a circumscribed and well-defined part of the brain such as the hippocampus, which is a bilateral structure in the medial temporal lobe (MTL), has long been thought to serve the functions of episodic memory and spatial navigation, and historically, this was based primarily on evidence from lesion patients and animal models. But neuroimaging evidence suggests a far more expansive functional repertoire. According to recent work, there is evidence that the hippocampus plays a role in aspects of perception, attention, working memory, language, and semantic processing, "all of which were originally believed to be outside the domain of hippocampal and MTL function" (Kwan, Craver, Green, et al. 2015; see also references therein). It is tempting to think that there may be some common function or set of functions that is common to all these cognitive capacities, which is the unique cognitive function performed by this particular neural structure. But in this and many other such cases, there does not appear to be a common cognitive function, at least not one that can be specified in terms of our existing cognitive categories.[11] Furthermore, many cognitive functions or capacities are themselves subserved by numerous other brain regions and structures. Hence, what is emerging from these noninvasive neuroimaging technologies is evidence for a many-to-many relationship between neural structure and cognitive function. One of the most promising attempts to understand and explain this relationship is the theoretical framework of "neural reuse" (Anderson 2010; 2014). Neural reuse says that brain structures are "used for diverse purposes in various task domains" (Anderson 2014, 9). This means that when a population of neurons is involved in different cognitive tasks they do not always perform the same function. The cognitive function of a neural population is not an intrinsic property of that particular group of cells but depends on the patterns of interaction between those biological entities and others, in a way that is not yet fully understood in contemporary neuroscience. Neural reuse is therefore a theoretical proposal that urges neuroscientists to take into account not just "neural real estate" but the interrelations between sets of neurons in attempting to understand the relationship between brain and cognition.

[11] Some neuroscientists have advocated a thoroughgoing revision of our cognitive ontology in an attempt to locate the cognitive functions served by different neural regions or networks (e.g. Poldrack & Yarkoni 2016), but this effort has yet to yield convincing results. Anderson (2014, 128) sometimes argues that neural structures or regions may have lower-level functional profiles, which he labels "neuroscientifically relevant psychological factors" (NRPs), though I take it that these are not full-fledged psychological or cognitive functions.

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

If the support for neural reuse were limited to neuroimaging evidence based on fMRI technology, there would be an obvious way to resist the conclusion. It would be open to a skeptic to say that the reason that we are not always able to find a direct link between neural structure and cognitive function is that our current technologies are as yet unable to zoom in precisely enough or that we have not been careful enough to distinguish different areas within a supposedly multifunctional region. In the face of evidence of multifunctionality, proponents of "localization" sometimes adopt a "divide and conquer" strategy, which posits that different functions are being performed by different subregions of the same region (e.g. Saxe, Brett, & Kanwisher 2006; see also McCaffrey 2015b). To return to the example of the hippocampus, it may be thought that the diverse cognitive functions associated with this brain structure may on closer inspection turn out to be subserved by different parts or areas of the hippocampus. Even though the average volume of each half of the hippocampus on the right and left sides of the brain does not exceed a few cubic centimeters, that tiny volume includes tens of millions of neurons, and there would seem to be ample opportunity for different populations of neurons to specialize in different tasks. But Anderson (2014, 30–34) and other researchers have presented various considerations to support the conclusion that reuse persists no matter how fine-grained our investigative techniques. For example, he cites evidence that single neurons in the roundworm *C. elegans* sometimes perform both motor and sensory functions and that other individual neurons even participate "in generating completely opposite behavior as a result of alteration of the neuron's sensitivity, physical connections, and functional connectivity by various chemicals and genes (effects known collectively as neuromodulation)" (Anderson 2014, 32). This and other types of evidence would undermine the "divide-and-conquer" strategy favored by deniers of neural reuse.

Multiple realization and neural reuse can be seen to be mirror images of one another, as a quick comparison will show. According to multiple realization, there is a one-to-many relationship between mind and brain. One has to be careful about stating this claim, since it is not always clearly distinguished from a substantively different claim. Proponents of multiple realization are not merely saying that some cognitive capacities are subserved by a number of different brain regions working in concert. If that were all there was to it, multiple realization would be no threat to reduction. Reductionists could simply reply by saying that the unit of interest is not the neural region but the set of regions or neural network, and that when these units are properly individuated, they might well be put in a

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

one-to-one correspondence with cognitive functions. Rather, multiple realization is saying that different neural structures can perform the same cognitive function, as it were, in parallel.[12] This is not a matter of different neural structures teaming up within the same system, but the same type of region or network or other configuration doing so in different systems (e.g. species, individuals, or even individuals-in-context), despite the fact that they are not type identical. Multiple realization blocks reductionism because it says that the relationship between mind and brain is one-to-many. Neural reuse, on the other hand, declares the relationship to be many-to-one, since it holds that there may be multiple cognitive functions that are subserved by a single neural region or structure. In some ways, this claim is more puzzling than that of multiple realization, since it seems to be denying something like a basic principle of metaphysical supervenience, namely that there can be no mental difference without neural difference. But as can be gleaned from the brief sketch of neural reuse above, the claim is that the functions of neural structures do not attach to them intrinsically, but are partly a result of extrinsic factors such as their anatomical and functional connectivity, as well as modulation by genes and chemicals. Hence, neural reuse is saying that cognitive function does not pertain to populations of neurons but arises out of complex interactions in the brain. In principle, this does not block reduction outright since if neuroscientists are able to identify all the relevant interactions and modulations, it may be possible to find the neural correlates of cognitive functions, even though these are not simply identical to regions or networks. Some of the neural factors may not be such as to be readily identifiable using existing neuroimaging techniques, but they are presumably identifiable in other ways. However, I will now put forward what I take to be a more principled obstacle to a reduction of the cognitive domain to the neural domain.

If held jointly, the claims of multiple realization and neural reuse issue in the assertion that the relationship between cognitive function and

[12] For instance, Price and Friston (2005, 262) do not sufficiently distinguish the two senses of a one-to-many mapping in the following passage:

> Functional neuroimaging data preclude a one-to-one mapping in two ways. First, attempts to manipulate a ‘single’ cognitive process (e.g. semantics) often elicit a distributed pattern of activation over many areas (i.e. a one-to-many mapping from function to structure). Second, the same brain region, or set of regions, may be activated by tasks with different cognitive processes (i.e. a many-to-one mapping). In short, there is a many-to-many mapping between cognitive functions and anatomical regions, with a range of cognitive processes emerging from different patterns of activation among a limited number of brain regions.

Here, the first case they cite, of a one-to-many mapping, is not a case of multiple realization as it is usually understood.

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

neural structure is many-to-many. It is not uncommon for neuroscientists to explicitly endorse a many-to-many mapping in recent years (e.g. Price & Friston 2005; Poldrack & Yarkoni 2016), especially as neuroimaging evidence has piled up indicating a convoluted relationship between structure and function. While I think there is strong evidence to support both claims, multiple realization and neural reuse, I will try to make the case against mind–brain reduction a little stronger. One of the broad aims of this book, to be supported by looking more closely at specific case studies in subsequent chapters, is to argue for a many-to-one relationship between mind and brain, but not just for the reasons provided by the advocates of neural reuse. In addition to the fact that specific neuronal populations appear to be implicated in diverse cognitive functions due to neural connectivity, neuromodulation, and other relational factors that are internal to the brain, I will argue that another reason that neural regions do not always subserve the same cognitive functions is that these functions are individuated in part with reference to the surrounding context of the thinker and the thinker's history. Since cognitive functions are often individuated externalistically or anti-individualistically, while neural structures, mechanisms, and processes are usually not so individuated, I will argue that this blocks the possibility of a correspondence between cognitive categories and neural ones, even taking into account the subtleties of brain chemistry and electrophysiology (see also Khalidi 2017; 2020). The claim is not that neuroscience never has occasion to individuate its constructs relationally or etiologically, but rather, the point is that even when it does engage in relational or etiological individuation, the neuroscientifically salient relata or causal histories do not always coincide with those relevant to cognition.[13] In subsequent chapters, I will try to show that some of the principal categories that we use to explain and predict cognitive phenomena are individuated with reference to contextual and etiological factors that are not generally invoked in understanding the workings of the brain, and this leads to a mismatch between the taxonomic systems of, say, cognitive psychology and neuroscience. The claim is not just a causal

[13] For a defense of etiological individuation in neuroscience, see Garson (2011), specifically when it comes to ascribing functions based on evolutionarily selected effects rather than synchronic causal roles. However, he also acknowledges that non-etiological individuation "appears to be more consistent with neuroscientific practice, which is more preoccupied with structural and functional decompositions of complex abilities than with speculation about evolutionary histories" (Garson 2011, 549). Amundson and Lauder (1994) put forward a general defense of ascribing synchronic causal role functions in various areas of biology, such as comparative anatomy, and at least some of their arguments would seem to apply to neuroscience.

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

but an individuative one. It is not merely that many cognitive states are both effects and causes of environmental factors (since many neural states are too). Rather, cognitive entities are often individuated in part on the basis of such factors, whereas neural entities are usually not so individuated, and this is what leads to a many-to-many mapping between neural and cognitive taxonomic categories. This results in crosscutting systems of categories, ruling out even a local or limited reduction.[14]

In one sense, it should not be surprising that there can be a many-to-many mapping among neural structures and cognitive functions. That is what we find in other scientific domains when we are investigating structures and functions, particularly in the biological sciences. It is widely accepted that different biological structures can perform the same function, as when arthropod eyes and vertebrate eyes perform the function of sight and "display similar psychophysical phenomena" (Weiskopf 2011, 236; cf. Stinson 2016). It is also relatively uncontroversial that the same biological structure can perform different functions, as when bird feathers are exapted from performing the function of thermoregulation to serving the function of flight (cf. Ereshefsky 2012). In general, then, the structure-to-function mapping in biology can be expected to be many-to-many, rather than one-to-one, or even one-to-many. This also holds for brain structures and cognitive functions in particular. When it comes to multiple structures performing the same function, advocates of multiple realization have brought forth various examples and considerations to support the conclusion that the same cognitive function can be performed by different neural structures in different species, in different individuals, and even in the same individual at different times (see e.g. Aizawa & Gillett 2009). As for multiple functions being performed by the same structure (i.e. multifunctionality), the arguments for neural reuse emphasize the fact that it is not just structures that account for functions, but also the relations between structures and the ways in which those structures are modified chemically and by other means. To further support the claim of multifunctionality, I would point to the fact that a thinker's environment and causal history can enter into the individuation or identification of the cognitive function being performed. In some cases, the same neural

[14] Although there has been discussion of the tension between the externalist and internalist individuation of mental states at least since the 1970s, many philosophers have assumed that individuation in a science of psychology, as in neuroscience, is always internalist or individualist (Stich 1985; Fodor 1987), with few exceptions (notably, Burge 1986). By contrast, I will go on to argue that the specific type of externalism that I will be arguing for is widely attested in the study of cognition and that it creates obstacles for a general type reduction among the kinds of psychology and neuroscience.

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

structures can be implicated in different cognitive functions because those functions are individuated with reference to environmental and etiological factors. This means that cognitive kinds are often typed differently than neural kinds. Moreover, this applies not just to neural *structures* but to the *totality of neurophysiological facts*, since these are not generally typed with reference to a thinkers' environment and causal history. Although some neural objects, states, processes, mechanisms, and other entities are sometimes individuated environmentally and etiologically, much of neuroscience investigates the brain in relative isolation from the environment and from its evolutionary history. Even when neuroscience is interested in identifying functional properties, the environmental and etiological factors that are relevant to neuroscientific taxonomy are not always the ones that are pertinent to cognition.[15] Hence, multi-functionality (or the one-to-many mapping between neural and cognitive categories) is supported not just by neural reuse but by the individuation of cognitive kinds with reference to certain distinctive environmental and etiological factors. This type of multi-functionality goes beyond considerations drawn from neural reuse and suggests that the one-to-many mapping applies to neural categories in general, not just to structural categories, such as brain regions or neural networks (e.g. *hippocampus*, *default mode network*). Moreover, if cognitive facts do not just supervene on neural facts, this would rule out not just a type reduction, but an extremely weak dependence of cognitive facts on neural facts. It would also mean that neural generalizations and explanations cannot be expected to wholly account for cognitive entities. This is not a limitation of neuroscience, just a result of the fact that different disciplines and subdisciplines investigate different causal systems, as indicated earlier.

1.5 Realism

Philosophers sometimes make a distinction between classification on the basis of nature's own divisions and classification on the basis of human interest. But there would seem to be something a little misguided about

[15] Craver (2013) defends the importance and centrality of a non-etiological conception of function to neuroscience (and the physiological sciences more generally). For example, he points out that a type of molecule can be described as a neurotransmitter, regardless of "the developmental or evolutionary origins of the molecule in question" (Craver 2013, 137). Garson (2019) disputes that there is a useful non-etiological conception of function in the biological sciences. For my purposes, however, what matters is that the environmental and etiological factors relevant to neural functions are often different from those relevant to cognitive functions.

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

this alleged contrast, at least if our interest lies in making valid distinctions. If epistemic conditions are good and if we are not simply cognitively incapable of understanding some aspects of the world, then our considered theories can be expected eventually to converge on the causal structure of the world. These theories comprise the classification schemes that purport to reveal the real kinds that exist in the world. I have argued elsewhere that our epistemic interests are geared toward discovering causal structures and that as long as inquiry is guided by these interests (as opposed to moral, political, aesthetic, or other interests), then there is no conflict between classifications that serve human interests and those that aim to uncover the divisions in nature (Khalidi 2013). This should apply to the human mind no less than to any other corner of the universe. This is how the naturalist picture sketched out in previous sections meshes with realism about scientific categorization and the search for real kinds. In the previous section, I defended an anti-reductionist position regarding cognitive entities, and in this section, I will try to show that this position is compatible with realism about cognitive kinds. I will do so by attempting to respond to two challenges that might be thought to confront an anti-reductionist realist position about cognitive kinds: mind-dependence and self-reflexivity. But before doing so, I will briefly address a question that pertains to realism not just in the cognitive domain but in other special-science domains.

The categories in the fundamental sciences, such as elementary particle physics or elemental chemistry, seem quite distinct: their individuation conditions can be clearly delineated and they can be easily distinguished from one another. These categories may not be as crisp as many philosophers have traditionally assumed, yet they appear to have a "naturalness" that categories in many other sciences appear to lack. Meanwhile, categories in the biological, psychological, and social sciences, do not seem to be as clearly delineated and they do not appear to be as distinct from each other. Moreover, I think it is safe to say that this is not just a feature of the categories in these respective sciences but of the kinds themselves. In other words, it is not an artifact of our taxonomic practices in these sciences but a property of the underlying kinds. How can we account for this difference? One way to think about it is to relate it to the nature of causal relations in the different domains. Though causal relations appear to be deterministic at the macro-level, there are many more intervening causes in the causal systems that occur in the domains of the special sciences. This difference has largely to do with the asymmetrical relationship between the micro-level and macro-level, since domains that tend to depend constitutively, causally, and otherwise, on other domains tend to be more complex

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

and involve many competing causal processes. I mentioned in Section 1.4 that scientists try to isolate "closed systems" in their areas of inquiry, but this is easier done in a test tube than in the psychology lab. That is not just because one can spatially isolate the causal process of a chemical reaction more fully than that of a participant performing a cognitive task, but also because there are fewer other causal processes that might interfere with it. When it comes to the chemical reaction, the temperature in the room may do so, as may the level of humidity, and perhaps the lighting, among others, but the potential intervening factors are far fewer than those that might intrude on the psychology experiment. When it comes to the latter, in addition to the preceding factors, we might also mention the facial expression of the experimenter, her tone of voice, and music from the hallway, not to mention what the participant dreamt the night before, as well as his heart rate, psychiatric diagnosis, cultural background, gender, and level of education, among many others. On the conception of kinds that I have spelled out, this difference in the nature of causal processes in different domains leads to a difference in the nature of kinds. Such domains are dubbed "causal thickets" by Wimsatt (2007, 237), and he characterizes them as follows: "With increases in the complexity of objects, and in their number and variety of degrees of freedom, they can interact with one another in more varied and complex ways …"[16] This means both that the causal connections between properties is more messy (due to possible interventions by other causes) and that there are more potential candidates for kinds depending on which combinations of properties are chosen for singling out and which causal processes we investigate. (For instance, should we make a rough psychological generalization over all humans, or a narrower generalization over a smaller reference class of humans with a certain cultural background? Should we regard, say, *male humans* or *Western humans* as subordinate kinds for certain psychological purposes?) It also means that there are often a number of different kinds in the same vicinity and we can choose which ones to focus on in our scientific theorizing. That does not mean that the boundaries between kinds are arbitrary, but it does suggest that we are sometimes at liberty to shift them somewhat while continuing to capture real kinds, that is, clusters of properties causally linked to other properties. Does this provide some room for bringing non-epistemic interests to bear on the demarcation of kinds? I would argue that these taxonomic choices will generally serve different *epistemic*

[16] Wimsatt (2007, 238) writes explicitly that "[t]he neurophysiological, psychological, and social realms are mostly [causal] thickets …"

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

interests and that the demarcation of taxonomic boundaries is not driven by non-epistemic interests. That is because different causal processes are captured by different demarcations, which will deliver different explanations.[17] To provide further support for this claim, I will try to illustrate it in examining some of the cognitive kinds discussed in later chapters.

Here, something more needs to be said about the causal system that coincides with the cognitive domain. As I conceive it, the domain of cognition lies broadly within what Marr (1982) famously identified as the computational level, as opposed to the algorithmic and implementational levels. On Marr's view, the computational level of explanation "specifies what and why," whereas the algorithmic level "specifies *how*" (1982, 23; original emphasis). Meanwhile, the implementational level investigates "the particular mechanisms and structures … in our heads" (1982, 19). Marr was explicit in holding that the computational level can be multiply realized in the algorithmic, as can the algorithmic in the implementational. As he put it: "the same algorithm may be implemented in quite different technologies" (1982, 24). A theoretical explanation at the computational level characterizes a cognitive system in informational terms, taking one kind of information as input and transforming it into another. Computational explanations also demonstrate why that informational mapping is appropriate and adequate for solving the problem or performing the task at hand (Marr 1982, 24). To use an example that will be briefly encountered in Chapter 4, vervet monkeys have an alarm call system to alert conspecifics to the presence of three different types of predators: leopards, eagles, and snakes (Cheney & Seyfarth 1990). The vervets give a different response to each type of predator and conspecifics react by taking the appropriate evasive action. A computational or cognitive theory of the system would describe the different inputs that the vervets rely on (e.g. visual, auditory) and what kinds of outputs they make to each input ("what"). It would also explain why these specific predators are the ones that elicit these calls, types and rates of errors, the degree of innateness or learning involved, and the kinds of selection pressures that might have given rise to this system ("why"). This explanation stops short of detailing the precise algorithm that is used to encode the input in a mental representation and how that representation is then translated into appropriate motor activity ("how"), let alone the neural circuitry recruited in the task. But I would add an ontological dimension to Marr's methodological and epistemic account

[17] On this point, I disagree with Ludwig (2016), who argues that non-epistemic factors play a role in choosing a scientific ontology.

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

of the computational level. Computational explanations work precisely because there are relatively self-contained causal processes in the cognitive domain that can be understood somewhat independently of the algorithms that enable them to perform their functions and the physical structures that implement them. Furthermore, computational or cognitive systems are functionally individuated, since what a system does in solving a problem or performing a task is naturally understood in terms of its function. Computational systems are a species of functional systems and computational explanations are a type of functional explanation. In this context, functions can be understood both etiologically (selected-effect functions) and synchronically (causal-role functions). Indeed, in some cases, they can be understood as combining synchronic and etiological elements.[18] The function of a cognitive system is understood both as a response to a problem in the organism's current environment, as well as, (sometimes) an adaptive response that has been naturally selected. This means that cognitive entities are individuated both environmentally and etiologically, in keeping with the argument of previous sections.

The cognitive domain is an aspect of the world that lends itself to scientific inquiry, although it cannot (obviously) be considered "mind-independent." In much recent philosophical work, "mind-independence" has been considered criterial for realism about a certain domain, but this is ill-advised, at least if we assume (as I have) that mental entities can be real. To be sure, there have been some attempts to distinguish different kinds of mind-dependence, which argue that only certain types of mind-dependence preclude realism about an entity. But I have argued elsewhere (Khalidi 2016b) that mind-independence is not an appropriate criterion by which to distinguish real from non-real kinds. Whether or not a kind depends on the mind, whether causally, constitutively, metaphysically, definitionally, or otherwise, does not have anything to do with whether it is a real kind. If we think minds themselves are real, then dependence on them should not be taken as a sign that something should not be admitted into our ontology. Rather, I would suggest that realism about kinds be understood in terms of the account of kinds that I have already outlined in this chapter (especially Section 1.2), namely in terms of being

[18] Griffiths (1993, 410) has proposed that "the proper functions of a biological trait are the functions it is assigned in a [causal-role] functional explanation of the fitness of the ancestral bearers of that trait." Garson (2019) has proposed a variation on the selected effect account of function that involves differential retention of traits or features in addition to differential reproduction, which may serve to capture the notion of function needed in cognitive science, whether or not evolutionary considerations are relevant. In particular, it handles cases of trial-and-error learning.

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

part of the causal structure of the universe. If minds, their states, capacities, processes, and so on, are causally efficacious then that is what makes them features of reality.[19] Mind-independence has nothing to do with it. To be sure, there are interesting differences between mind-dependent and mind-independent kinds, as well as between different kinds of mind-dependence, but none of these distinctions demarcate the distinction between real and non-real kinds.

It is obvious that all mental kinds, including cognitive kinds, are unavoidably mind-dependent (causally, constitutively, and otherwise), but it is debatable whether they are mind-dependent in another sense. One way to capture this additional dimension of mind-dependence is by dubbing it "response-dependence." Now there is a trivial sense in which many cognitive kinds are response-dependent. For example, whether or not I possess the concept KUMQUAT seems to depend at least in part on the responses I give to certain stimuli in my environment. Among other things, it depends on whether I can sort or categorize kumquats in the grocery store, whether I can discriminate kumquats from loquats, answer questions about their color and taste, and so on. But is possession of the concept just a fact about *my own responses* and other causal powers (as well as perhaps my causal history), or does it also depend on *others' responses* toward me? It is controversial whether some cognitive kinds are response-dependent in this second sense, as is held, for example, by an interpretivist or ascriptionist view of concepts. On such a view, whether or not a thinker possesses a concept and which concepts are possessed by a thinker are matters that depend ultimately on how they are interpreted by others; they are facts about the way in which others respond to them. In Chapter 2, I will tackle the question as to whether concepts are response-dependent in this sense, but for now I want to make the point that even if they were, this should not undermine realism about concepts (or other cognitive kinds). Many social kinds exhibit response-dependence, yet they are robust causal kinds nevertheless. As various philosophers of social science have argued, this type of mind-dependence is compatible with an entity having causal efficacy in the social domain (Hacking 1995; Mallon 2003). For example,

[19] Some philosophers (e.g. Kim 1992) would question the claim that minds are causally efficacious. I will not try to justify that claim here, though I have put forward some arguments against the view that causal efficacy pertains exclusively to the most fundamental entities in the universe (see Khalidi 2011). One of those arguments can be summarized succinctly as follows. Consider a possible world just like ours except that there is no fundamental level, a possibility that some scientists and philosophers take seriously and consider to be coherent (see e.g. Block 2003; Schaffer 2003). Would this be a world in which there is *no* causation? If not, it seems to be a mistake to confine causation to the most fundamental level.

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

social kinds like *money*, *ritual*, *gender*, and *race* are often thought to be mind-dependent in this sense. Thus, what it is to be a *woman* or *man* in many contemporary societies is thought to depend ultimately on how they are perceived and represented by others in their society (e.g. Haslanger 2000; Àsta 2013), even though these perceptions are based on how people themselves behave, are socialized, and present themselves. Yet *gender* is a real kind with robust causes and effects in the social domain. If some kinds in the psychological or cognitive domain are likewise response-dependent, then this would not seem to undermine their causal nature. Interpretivism about the mental is sometimes thought to dictate an anti-realist or instrumentalist position about mental properties or kinds, but even though this may be the position of some interpretivists and their critics, it seems possible for someone to hold both that mental entities are mind-dependent in this sense (i.e. response-dependent) yet real nevertheless.

There is a certain self-reflexivity involved in using our mental capacities to study those mental capacities themselves. Cognitive science is self-referential in a way that most other sciences are not: we are trying to understand the human mind (as well as the minds of other creatures) using the resources of the human mind. Perhaps this reflexivity is sometimes overblown, but it does lead to some tricky situations when thinking about categorization and kinds. This reflexivity comes to the fore in the very next chapter, which focuses on concepts, a topic intimately tied to the whole issue of categorization, since I take it that categories are best understood as classificatory concepts (as mentioned in Section 1.2). Thus, we will be trying to determine the nature of concepts while at the same time presupposing that there are such things as classificatory concepts that are used to pick them out. To put it more succinctly, we are attempting to articulate the concept of *concept*. But though some care needs to be taken in using the mind to study the mind itself, this inquiry does not seem to lead inevitably to a kind of paradox, or to result in hopeless subjectivism. Another way to bring out the self-reflexivity involved in the scientific study of the mind is to observe that real kinds crop up in two guises in cognitive science. The first is the one that I have emphasized so far: the identification of the real divisions or joints in the mind–brain. The second is subtly different and interacts in a complex way with the first. That is the capacity of the human mind to identify such divisions or joints in the world (including the mind–brain). Many cognitive scientists are interested in the psychological processes of categorization and in the nature of the categories that we use to understand the world and ourselves. They investigate the structure of these categories and their manner of implementation in the brain.

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

They are also interested in how these categories are acquired and the extent to which they are innate. This dual role for categories in cognitive science, as aspects of the theory and as elements of the mind itself, might give rise to a worry. Could it be that the categories that we use in our scientific accounts of the world are just an outgrowth of the categories inherent to our minds, which are either innate or acquired early in development on the basis of insufficient or faulty evidence?[20] This is a definite risk, but the enterprise of science aims partly to ensure that such categories, if faulty, do not persist in higher-order cognition. We have ways of weeding out categories in science if they do not play an epistemic role, and I have argued that the epistemic role of scientific categories is based on the causal role of kinds. This gives us some grounds for thinking that faulty categories can be detected and discarded. Many such mental categories have been abandoned in intellectual history and the history of science (e.g. many of the categories associated with humorism, phrenology, and psychoanalysis). But I would argue that widescale conceptual removal and replacement is rarer than philosophers sometimes suppose because successive scientific theories are generally formulated over largely the same conceptual base. Part of the justification for this assertion lies in the theory of concepts that I will defend in the next chapter. This is one respect in which there is a reflexive relationship between the theory of categorization presented in this chapter and the account of concepts put forward in the next chapter. But though this means that the theory of concepts to be proposed in the next chapter both corroborates and is supported by the account of categorization that I have defended in this chapter, I take this to be an unavoidable feature of any theory that considers concepts as a taxonomic category in cognitive science, and not as a case of vicious circularity.

1.6 Conclusion

In this chapter, I have tried to motivate a naturalist and realist account of cognitive kinds, and have argued against the reduction of cognitive kinds to neural kinds. But if, as I have argued, there are principled obstacles to reducing our mental categories to neuroscientific ones, does that mean

[20] For example, Leslie (2013, 109) argues that essentialist positions in philosophy rest on intuitions that are "due to a deep-seated cognitive bias, rather than to any special insight into the nature of reality." But the fact that such positions are not universally accepted and have been widely criticized in both philosophy and cognitive science is evidence that we are not fated to adopt our intuitive and innate categories.

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press

that neuroscience cannot shed light on our mental lives? Even though there is not likely to be a one-to-one correspondence between the mental and neural in all cases, there will be many ways in which a better understanding of neural mechanisms may help shed light on cognitive processes. Discoveries in neuroscience can certainly inform our psychological theories and explanations even though they are not likely to replace or preempt them. The relationship may be compared to that between genotype and phenotype, which is also not a reductive one, at least according to many philosophers of biology (see e.g. Kitcher 1984; Schaffner 1998; Wimsatt 2007). Once it was thought that genes would correspond directly to phenotypic features and that we would be able to read off the phenotype from the genotype. But as we find out more about the complex relationship between the two, it has become clear that this is just not the case, even when it comes to fairly simple traits like eye color and height in humans.[21] A vanishingly small number of phenotypic traits can be traced directly to single allelic variants or even a limited number of them. It may be objected here that when it comes to genotype and phenotype, the former causes the latter and does so in conjunction with other causes, but neurophysiological processes do not cause mental ones, since they are identical with them. Even though some instances of neurophysiological tokens and types may be identical with cognitive tokens or types, or bear a compositional or mereological relation to them, I do not think this holds generally, as I argued in Section 1.4. In most cases, when we individuate the items under consideration carefully, the relationship between the neural and cognitive turns out not to be one of identity, either when it comes to types or tokens. The relationship may not always be straightforwardly causal either, but it combines elements of causation, composition, constitution, and other relations, as I will try to indicate in later chapters.

[21] According to current estimates, there are around 20,000 genetic variations (i.e. single nucleotide polymorphisms (SNPs), out of a total 4–5 million, i.e. 0.5 percent) in the human genome that are thought to influence a person's height (Lello, Avery, Tellier, et al. 2018).

https://doi.org/10.1017/9781009223645.002 Published online by Cambridge University Press