Integrating Mind and Brain Science: A Field Guide
David Michael Kaplan

1. Introduction

Long a topic of discussion among philosophers and scientists alike, there is growing appreciation that understanding the complex relationship between neuroscience and psychological science is of fundamental importance to achieving progress across these scientific domains. Is the relationship between them one of complete autonomy or independence – like two great ships passing in the night? Or is the relationship a reductive one of total dependence – where one is subordinate to the other? Or perhaps the correct picture is one of mutually beneficial interaction and integration – lying somewhere in between these two extremes? One primary strategy for addressing this issue, and one which occupies center stage in this volume, centers around understanding the nature of explanation in these different domains. Representative questions taken up by various chapters in this volume include: Are the explanatory patterns employed across these domains similar or different in kind? If their explanatory frameworks do in fact differ, to what extent do they inform and constrain each another? And finally, how should answers to these and other related questions shape our thinking about the prospects for integrating mind and brain science?

Several decades ago, during the heyday of computational cognitive psychology, the prevailing view was that the sciences of the mind and brain enjoy a considerable degree of independence or autonomy from one another – with respect to their theories, their research methods, and the phenomena they elect to investigate (e.g., Fodor 1974; Johnson-Laird 1983; Lachman et al. 1979; Newell and Simon 1972; Pylyshyn 1984; Simon 1979). In an expression of the mainstream perspective in the field at the time, the psychologist Philip Johnson-Laird proposes that “[t]he mind can be studied independently from the brain. Psychology (the study of the programs) can be pursued independently from neurophysiology (the study of the machine code)” (Johnson-Laird 1983, 9).

In the intervening decades, the doctrine of disciplinary autonomy has fallen on hard times. Today, it is far from being the universally held or even dominant view. In fact, given the emergence of cognitive neuroscience as a new scientific field formed precisely at the interface between these disciplines, one might reasonably wonder whether the consensus has now shifted in exactly the opposite direction – towards a view of complete disciplinary integration and interdependence rather than autonomy. In the inaugural issue of the Journal of Cognitive Neuroscience, then editor Michael Gazzaniga writes:

In the past 10 years, there have been many developments in sciences concerned with the study of mind. Perhaps the most noteworthy is the gradual realization that the sub-disciplines committed to the effort such as cognitive science, neuroscience, computer science and philosophy should not exist alone and that each has much to gain by interacting. Those cognitive scientists interested in a deeper understanding of how the human mind works now believe that it is maximally fruitful to propose models of cognitive processes that can be assessed in neurobiologic terms. Likewise, it is no longer useful for neuroscientists to propose brain
mechanisms underlying psychological processes without actually coming to grips with the complexities of psychological processes involved in any particular mental capacity being examined. (Gazzaniga 1989, 2)

From the outset, contributors to the cognitive neuroscience movement have explicitly recognized the interdisciplinary and integrative nature of the field, which is unified by the common goal of trying to decipher how the mind and brain work (Boone and Piccinini 2015; Churchland and Sejnowski 1988). Despite the rapidly growing influence of cognitive neuroscience and cognate fields such as computational neuroscience, some researchers continue to maintain that neuroscience is largely or completely irrelevant to understanding cognition (e.g., Fodor 1997; Gallistel and King 2009). Others maintain that psychology is (or ought to be) a tightly integrated part of the broader scientific enterprise to discover and elucidate the multi-level mechanisms underlying mind and cognition (e.g., Boone and Piccinini 2015; Piccinini and Craver 2011). Hence, the debate over an autonomous psychology remains incompletely settled.

The objective of this chapter is to provide a field guide to some of the key issues that have shaped and continue to influence the debate about explanation and integration across the mind and brain sciences. Along the way, many of the central proposals defended in the individual chapters will be introduced and important similarities and differences between them will be highlighted. Since questions on this topic have a long track record of philosophical and scientific engagement, providing some of the broader historical and theoretical context will facilitate a deeper appreciation of the contributions each individual chapter makes to these important and ongoing debates.

2. Autonomy and distinctness: some provisional definitions

It is frequently claimed that psychology is autonomous and distinct from neuroscience and other lower-level sciences. But what exactly do these terms mean? Before proceeding it will prove useful to have working definitions of these key concepts, which recur throughout this introductory chapter as well as the volume more generally.

First, consider the notion of autonomy. Generally speaking, autonomy implies independence from external influence, control, or constraint. The Tibet Autonomous Region is, at least according to the Chinese Government, appropriately so called because it is free of direct external control from Beijing. Autonomous robotic vehicles are appropriately so called because they are capable of sensing and navigating in their environments without reliance on direct human input or control. In a similar manner, scientific disciplines may also be autonomous from one another. Following Piccinini and Craver (2011), we might provisionally define disciplines as being autonomous from one another when at least one of the following conditions are satisfied:

(a) they can independently select which phenomenon to investigate,

(b) they can independently select which methods to use,
(c) they can independently select which theoretical vocabulary to apply,

(d) the laws/theories from one discipline are not reducible to the laws/theories of the other discipline, or

(e) evidence from one discipline does not exert any direct constraints on the explanations/theories of the other discipline.

Importantly, this characterization of autonomy is flexible and admits of degrees. Disciplines can in principle completely or partially satisfy one or more of these conditions (a-e), and consequently can be completely or partially autonomous with respect to one or more of these conditions (a-e). At one extreme, two disciplines may only incompletely or partially satisfy a single condition, comprising a minimal form of autonomy. At the other extreme, two disciplines may completely satisfy all conditions, instantiating a maximal form of autonomy from one another (at least with respect to identified conditions a-e).

The notion of distinctness is closely related, but logically weaker. Disciplines exhibit distinctness if they investigate different kinds of phenomena, use different kinds of methods, or construct different kinds of explanations. The last of these is most relevant in the context of the present volume. As we will see, the thesis of the explanatory distinctness of neuroscience and psychology – roughly, that they employ characteristically different kinds of explanation – is a key premise in a number of recent arguments for the autonomy of psychology.

It is important to distinguish between autonomy and distinctness because one can obtain without the other. Generally speaking, distinctness is a necessary but insufficient condition for autonomy (for additional discussion, see Piccinini and Craver 2011). Without distinctness there is clearly no scope for autonomy. If two disciplines investigate the same phenomena, in an important sense, they cannot independently select which phenomenon to investigate. They are instead constrained or bound to investigate the same phenomena. Similarly, if two disciplines employ the same methods or theoretical vocabularies, in an important sense, they cannot independently select which methods or theoretical vocabularies to use. They are bound to use the same across the disciplines. Although distinctness is required for autonomy, it does not entail it. Two or more things can be distinct yet be mutually dependent or interdependent. Consider a simple example. The Earth is distinct from the Sun, yet these systems influence one another in a multitude of ways (e.g., gravitationally and thermodynamically). They are distinct, but not autonomous in any interesting sense of the word. Similarly, a scientific field or discipline may have its own distinct laws, principles, and theories, yet these may turn out to be reducible to or evidentially constrained by those of another discipline. Even though distinctness does not entail autonomy, as will be discussed shortly, they are often endorsed as a package deal.

3. Reduction or autonomy? A debate oscillating between two extremes
Philosophers weighing in on this topic have tended to focus on the prospects of either (a) achieving integration or unification of psychology and neuroscience via theory reduction, or (b) securing the autonomy of psychology and establishing in principle its irreducibility to neuroscience via multiple realizability. Despite its historical prevalence, one obvious problem with this this way of carrying out the debate is that it assumes a binary opposition between two extreme positions – either psychological science reduces to or is completely autonomous from neuroscience. According to the traditional picture, the proposed relationship between psychology and neuroscience is either one of complete dependence (reduction) or complete independence (autonomy). There is no stable middle ground. Many recent contributors to the debate reject this strong binary assumption and instead recognize that there is a continuum of plausible positions lying in between these two extremes. These intermediate positions involve some kind of relationship of partial dependence or partial constraint. A major objective of this volume is to focus attention on some of these “middle ground” positions that have been staked out in the debate and highlight their associated prospects and problems. Before considering these intermediates, however, it will be useful to take a closer look at the extremes.

3.1 Theory reduction

Many of the dominant ideas concerning the relationship between the mind and brain sciences have emerged from traditional philosophical perspectives concerning explanation and reduction. No view is more influential in this regard than the covering law account of explanation. According to the covering law account, explaining some event or phenomenon involves showing how to derive it in a logical argument (Hempel and Oppenheim 1958). More specifically, a scientific explanation should be expressible as a logical argument in which the explanandum-phenomenon (that which is being explained) appears as the conclusion of the argument and the explanans (that which does the explaining) appears as the premise set, which includes statements characterizing the relevant empirical conditions under which the phenomenon obtains (initial conditions) and at least one general law required for the derivation of the explanandum. According to the view, good scientific explanations are those in which the explanans provides strong or conclusive evidential grounds for expecting the explanandum-phenomenon to occur (Hempel 1965).

In its most general formulation, the covering law account is intended to apply uniformly to the explanation of spatiotemporally restricted events such as the explosion of the Space Shuttle Challenger, as well as the explanation of general regularities or laws such as the explanation of Kepler’s laws of planetary motion in terms of more basic laws of Newtonian mechanics. A derivation of one or more sets of laws (comprising a theory) from another set of laws (comprising another theory) is known as an intertheoretic reduction. According to the covering law account, intertheoretic reduction comprises a special case of deductive–nomological explanation.

Nagel (1961) developed these ideas into an explicit model of theory reduction, proposing that a theory (or law) from a higher-level science such as psychology can be reduced to, and thereby explained
by, a theory (or law) from a lower-level science such as neuroscience or biology just in case (a suitably axiomatized version of) the higher-level theory can be logically derived from (a suitably axiomatized version of) the lower-level theory. Since the terminology employed in both the reduced and reducing theories will invariably differ in some way, so-called bridge principles or rules of correspondence are required to establish links between the terms of the two theories. For example, a bridge principle might connect terms from thermodynamics such as “heat” with those of statistical mechanics such as “mean molecular energy”. Finally, because the reduced theory will typically only apply over a restricted part of the domain of the reducing theory or at certain limits, boundary conditions that set the appropriate range for the reduction are often required in order for the derivation to be successful. The theory reduction model can be represented schematically as follows (Bechtel 2008, 131):

\[
\begin{array}{c}
\text{Lower-level laws (in the basic, reducing science)} \\
\text{Bridge principles} \\
\text{Boundary conditions} \\
\end{array} \\
\therefore \quad \text{Higher-level laws (in the secondary, reduced science).}
\]

Oppenheim and Putnam (1958) famously argue that the Logical Positivists’ grand vision of scientific unification can finally be achieved, at least in principle, by revealing the derivability relationships between the theories of the sciences. They start by assuming that each scientific discipline occupies a different level within a single global hierarchy. The Oppenheim-Putnam framework then involves an iterated sequence of reductions (so-called micro-reductions) starting with the reduction of some higher-level theory to the next lowest-level theory, which in turn is reduced to the next lowest-level theory, and so on, until the level of fundamental physical theory is eventually reached. As Fodor succinctly puts it: “all true theories in the special sciences should reduce to physical theories in the long run”. (1974, 97). Oppenheim and Putnam’s general framework entails a specific conception of how psychology will eventually reduce to neuroscience and beyond:

It is not absurd to suppose that psychological laws will eventually be explained in terms of the behavior or individual neurons in the brain; that the behavior of individual cells – including neurons – may eventually be explained in terms of their biochemical constitution; and that the behavior of molecules – including the macro-molecules that make up living cells – may eventually be explained in terms of atomic physics. If this is achieved, then psychological laws will have, in principle, been reduced to laws of atomic physics… (1958, 7).

Although many philosophers once held out high hopes for reductive successes of the kind, few are so optimistic today. The theory reduction account faces challenges along several fronts including those raised about its adequacy as a general account of the relations between the sciences and as a specific account of the relation between neuroscience and psychology. Its descriptive adequacy as a general account of reduction in science has been called into question as it has proved exceedingly difficult to locate real examples that satisfy the account even in domains thought to be paradigmatic such as physics (e.g., Sklar
Other general issues concern its oversimplified or otherwise inaccurate portrayal of the relationships between the various sciences including the relationships between the theories, concepts, and explanandum phenomena of those sciences (e.g., Bickle 1998; Churchland 1989; Churchland and Churchland 1998; Feyerabend 1962; Schaffner 1967, 1969; Wimsatt 2007). Yet, it is the specific challenges that stand out as most relevant for present purposes.

One primary reason for heightened skepticism about theory reduction as an adequate account of the specific relationship between neuroscience and psychology is the conspicuous absence of laws or lawlike generalizations in these sciences. This is what Rosenberg (2001), in the context of biology, aptly refers to as the “nomological vacuum”. Since unification is supposed to be achieved by deriving the laws of psychology from the laws of neuroscience (or some other lower-level science such as biophysics), clearly a precondition for such unification is the availability of laws at both the level of reduced and reducing theories. If the theoretical knowledge of a given discipline cannot be specified in terms of a set of laws (an assumption that mechanists and others reject), there is simply no scope for unification along these lines. Yet, despite decades of effort to identify genuine lawful generalizations in psychology or neuroscience of the sort one finds in other scientific disciplines such as physics, few if any candidate laws have been revealed.

In their chapter, Martin Roth and Robert Cummins echo similar criticisms about the “nomical conception of science” at the heart of the covering law framework. As Cummins puts it in his earlier and highly influential work on the nature of psychological explanation: “Forcing psychological explanation into the subsumptivist [covering law] mold made it continuous with the rest of science only at the price of making it appear trivial or senseless” (Cummins 1983, 27). In their chapter, Roth and Cummins identify one source of confusion underwriting views that attribute an explanatory role to laws in psychological science. Building on previous work by Cummins (2000), they indicate how the term ‘law’ in psychology is often (confusingly) used by researchers working in the field to refer to effects (i.e., robust patterns or regularities), which are the targets of explanation rather than explanations in themselves. For example, Fitts’ law describes but does not explain the widely observed tradeoff between speed and accuracy in skilled human motor behavior. The Weber-Fechner law describes but does not explain how the just-noticeable difference between two stimuli is proportional to the magnitude of the stimuli. Nevertheless, someone might be tempted to try to read off the nomological character of psychological science (and the explanatory role of psychological laws) from the mere appearance of the word ‘law’ in these instances. Yet these nominal laws, which simply describe effects or phenomena to be explained, do not satisfy any of the standardly accepted criteria for lawhood such as being exceptionless, having wide scope, etc., and are thus poorly suited to play the required role in covering law explanation and theory reduction. Roth and Cummins instead maintain that psychological laws are better understood as capturing the explananda for psychological science rather than the explanans, and argue that, appearances notwithstanding, psychological explanations do not involve subsumption under laws. Their efforts to expose how the nomical character of psychology is largely illusory places additional pressure on efforts to recruit the covering law
framework to shed light on the nature of psychological explanation and reduction.

Another reason many participants in this debate see intertheoretic reduction as a problematic way to achieve unification among the scientific disciplines is that successful reduction renders the laws and theories of the higher-level (reduced) science expendable in principle. Since all of the laws and all of observational consequences of the higher-level (reduced) theory can be derived directly from information contained in the lower-level theory, the resulting picture is one in which the higher-level sciences in principle provide no distinctive, non-redundant explanatory contribution over and above that made by the lower-level science. As Fodor puts it, reductionism has “the curious consequence that the more special sciences succeed, the more they ought to disappear” (1974, 97). In practice, however, higher-level sciences might retain their usefulness temporarily until the lower-level sciences become theoretically mature enough to support the reductions on their own, or they might play heuristic roles such as revealing the regularities or phenomena that the lower-level sciences seek to explain. Hence, even hard-core reductionists such as John Bickle can admit that “psychological causal explanations still play important heuristic roles in generating and testing neurobiological hypotheses” (author’s emphasis; Bickle 2003, 178). But this picture will nevertheless appear deeply unsatisfying to those who seek to secure a long-term explanatory role for psychological science. For these and other reasons, using theory reduction as the basis for an account of the relationship between psychology and neuroscience has appeared unpromising to many.

The traditional theory reduction framework thus offers one potential strategy for unifying or integrating psychological and brain science. However, it is not without problems. The well-known considerations rehearsed above indicate that the prospects for achieving unification by reduction are either extremely dim due to the lack of explanatory laws in psychology and neuroscience, or else reduction can succeed but in doing so would impose unbearably heavy costs by rendering psychology explanatorily inert and obsolete. Neither of these paths appears particularly promising. This has consequently inspired a search for alternative ways of characterizing the relationship between the sciences of the mind and the brain that do not bottom out in theory reduction but nevertheless succeed in securing some degree of autonomy for psychology.

3.2 Autonomy and multiple realizability

Another traditional response that philosophers have given is to argue that psychology exhibits a similar kind of autonomy with respect to “lower-level” sciences such as neuroscience in the sense that their theories or explanations are unconstrained by evidence about neural implementation.

Many early defenses of the autonomy of psychology and other higher-level sciences involved appeals to multiple realizability in order to deny the possibility of reducing the theories or laws of the higher-level science to those of the lower-level science (e.g. Fodor 1974, 1997; Putnam 1975). These views emerged as direct responses to the traditional theory reduction model and its negative implications for the independent status of psychology and the rest of the special sciences.
Recall that according to the classical theory reduction model, successful intertheoretic reduction requires a specification of appropriate bridge principles and boundary conditions (Nagel 1961). Bridge principles establish systematic mappings or identities between the terms of the two theories, and are essential for the reduction to go through. Anti-reductionists therefore naturally gravitate towards these bridge principles in their attacks, claiming that bridge principles will generally be unavailable given that the events picked out by special science predicates or terms (e.g., functionally-defined terms such as “money” or “pain”) will be “wildly disjunctive” from the perspective of lower-level sciences such as physics (Fodor 1974, 103). In other words, the enterprise to build bridge principles connecting the vocabularies or predicates of the higher- and lower-level sciences in an orderly, one-to-one manner breaks down because higher-level phenomena are often multiply realized by heterogeneous sets of lower-level realizers. Put somewhat differently, multiple realizability entails that the predicates of some higher-level science will cross-classify phenomena picked out by predicates from a lower-level science. The one-to-many mapping from the psychological to the neurobiological (or physical) implied by multiple realizability renders the bridge principle building enterprise at the heart of the theory reduction model a non-starter. Since the establishment of bridge principles is a necessary condition for classical intertheoretic reduction, multiple realizability directly implies the irreducibility and autonomy of psychology.

This line of argument has connections to functionalist and computationalist views in the philosophy of mind, which also depend on a notion of multiple realizability. According to one influential version of computationalism, cognition is identified with digital computation over symbolic representations (Newell and Simon 1976; Anderson 1983; Johnson-Laird 1983; Pylyshyn 1984). Proponents of computationalism have long maintained that psychology can accomplish its explanatory objectives without reliance on evidence from neuroscience about underlying neural mechanisms. Multiple realizability is taken to justify a theoretically principled neglect of neuroscientific data based on the alleged close analogy between psychological processes and running software (e.g., executing programs) on a digital computer, and the multiple realizability of the former on the latter. According to the analogy, the brain merely provides the particular hardware on which the cognitive programs (e.g., software) happen to run, but the same software could in principle be implemented in indefinitely many other hardware platforms. For this reason, the brain is deemed a mere implementation of the software. If the goal is to understand the functional organization of the software – the computations being performed – determining the hardware details is a relatively unimportant step.

If psychological capacities are akin to the functional capacities of computer software in that they can be implemented in diverse physical substrates or hardware, then, in an important sense, they are distinct from and irreducible to the neural mechanisms that happen to realize them. For parallel reasons, psychological explanations making reference to psychological properties are likewise thought to be autonomous and distinct from neurobiological explanations citing the neural properties that realize them.

Although this line of argument held sway in philosophy for some time, multiple realizability-based arguments for the explanatory autonomy of psychology have been vigorously challenged in recent
decades. For example, critics maintain that the evidence for multiple realization is substantially weaker than has been traditionally assumed (Bickle 2003, 2006; Bechtel and Mundale 1999; Churchland 2005; Polger 2004, 2009; Shapiro 2000, 2004) or that the thesis of multiple realizability is consistent with reductionism (Richardson 1979; Sober 1999), and so psychological explanations either reduce to or ought to be replaced by neurobiological explanations.

In his chapter, Kenneth Aizawa enters into this debate and contends that multiple realization is alive and well in the sciences of the mind and brain, albeit in a more restricted form than many proponents of autonomy have previously endorsed. Focusing on examples from vision science, he argues that when one attends to actual scientific practice it becomes clear how evidence for different underlying neural mechanisms (lower-level realizer properties) for a given psychological capacity (higher-level realized properties) are not always handled in identical ways. Sometimes this information is used to support multiple realizability claims. Other times it is not. More specifically, Aizawa makes the case that scientists do not always adopt an “eliminate-and-split” strategy according to which differences in the realizer properties result in the elimination of the putative multiply realized higher-level property in favor of two (or more) distinct higher-level psychological properties corresponding to the different realizers. The role of the “eliminate-and-split” strategy has been the subject of much philosophical discussion since Fodor (1974) first explicitly identified it as a theoretical possibility:

[W]e could, if we liked, require the taxonomies of the special sciences to correspond to the taxonomy of physics [or neuroscience] by insisting upon distinctions between the natural kinds postulated by the former wherever they turn out to correspond to distinct natural kinds in the latter. (1974, 112).

If neuroscientists always applied this taxonomic strategy, multiple realizability would be ruled out in principle since differences in how the realizer properties are taxonomized would always reflect differences in how the realized properties are taxonomized. Clearly, this would undercut the prospects for an autonomous psychology. Aizawa aims to show that this is not always the case; sometimes the higher-level taxonomy is resilient in the face of discovered differences in lower-level realizers. Aizawa defends the view that how discoveries about different lower-level realizers are treated depends on specific features of the higher-level theory. In particular, sometimes higher-level psychological kinds are theorized in such a way as to permit a degree of individual variation in underlying mechanisms; other times they are not. It is only in the latter case that higher-level psychological kinds are eliminated and split in light of evidence about different underlying mechanisms. In cases of the former, higher-level kinds are retained in spite of such evidence.

Aizawa thus offers a more nuanced account of the role of multiple realizability considerations in contemporary mind and brain science, and aims to show how a piecemeal or partial but nonetheless genuine form of autonomy of higher-level psychological kinds may be secured. This is not the sweeping autonomy that Fodor envisioned, where the structural taxonomy of neuroscience never interacts, informs, or otherwise constrains the functional taxonomy of psychology. Neither is it a wholesale form of reduction
where the higher-level kinds are slavishly dictated by the taxonomic divisions established by the lower-level science. Instead, sometimes (but not always) higher-level kinds are retained in spite of such divisions.

4. Functional and computational explanation

A somewhat different strategy for establishing the autonomy of psychology, which does not directly rely on appeals to multiple realizability, involves identifying the distinctive kind (or kinds) of explanation constructed and used across these different disciplines. The key idea here is that the discipline of psychology has its own explanatory patterns, which do not simply mimic those of another more fundamental discipline and are not reducible to them. According to the general line of argument, although the prevalent form of explanation in the neurosciences and other biological sciences is mechanistic explanation (Bechtel 2008; Bechtel and Richardson 1993/2010; Craver 2007; Machamer et al. 2000), the dominant form of explanation in psychology is functional or computational explanation. Critically, the latter are not to be assimilated to the former; they are distinct kinds of explanation.

4.1 Functional explanation

It is widely assumed that the primary (although not exclusive) explananda in psychology are sensory, motor, and cognitive capacities such as object recognition or working memory (e.g., Von Eckardt 1995); and that psychologists explain these capacities by providing a functional analysis (e.g., Cummins 1975, 1983; Fodor 1965, 1968). Cummins defines functional analysis as follows:

Functional analysis consists in analysing a disposition into a number of less problematic dispositions such that [the] programmed manifestation of these analyzing dispositions amounts to a manifestation of the analysed disposition. (Cummins 1983, 28)

The central idea is that functional analysis involves decomposing or breaking down a target capacity (or disposition) of a system into a set of simpler sub-capacities and specifying how these are organized to yield the capacity to be explained. Traditionally, functional analysis has been thought to provide a distinct form of explanation from mechanistic explanation, the dominant mode of explanation employed in many lower-level sciences including neuroscience (Cummins 1975, 1983; Fodor 1965, 1968). Call this the DISTINCTNESS thesis. As a reminder, mechanistic explanations describe the organized assemblies of component parts and activities responsible for maintaining, producing, or underlying the phenomenon of interest (Bechtel 2008; Bechtel and Richardson 1993/2010; Craver 2007; Machamer et al. 2000). Cummins expresses his commitment to DISTINCTNESS in the following passages:

Form-function correlation is certainly absent in many cases, however, and it is therefore important to keep functional analysis and componential [mechanistic] analysis conceptually distinct. (1983, 29)

Since we do this sort of analysis [functional analysis] without reference to an instantiating system, the analysis is evidently not an analysis of the instantiating system. (1983, 29)

Fodor similarly embraces DISTINCTNESS when he states:
[V]is-à-vis explanations of behavior, neurological theories specify mechanisms and psychological theories do not” (1965b, 177).

Although DISTINCTNESS does not entail the autonomy of psychology from neuroscience (call this the AUTONOMY thesis), often these are defended together. Thus, Cummins embraces AUTONOMY when he claims that:

[T]his analysis [functional analysis] seems to put no constraints at all on [a given system’s] componential analysis. (Cummins 1983, 30)

Taken together, these claims about DISTINCTNESS and AUTONOMY form what has been called the received view about psychological explanation (Barrett 2014; Piccinini and Craver 2011).

In their chapter in this volume, Roth and Cummins further refine the influential position first developed by Cummins (1983). They argue that a proper understanding of functional analysis permits us to see how it provides a distinct and autonomous kind of explanation that cannot be assimilated to that of mechanistic explanation, but which nevertheless bears an evidential or confirmational relation to the description of underlying mechanisms. As an illustrative example, they describe a functional analysis of the capacity to multiply numbers given in terms of the partial products algorithm. The step-by-step algorithm specification provided by the functional analysis reveals little to no information about the implementing mechanism, yet they argue that the analysis provided by the algorithm provides a complete explanation for the capacity in question. According to the view Roth and Cummins defend, the functional analysis permits an understanding of why any system possessing the capacity for computing the algorithm ipso facto exhibits the specific regularities or patterns that constitute the phenomenon to be explained. And this, they argue, is all that is being requested of the explanation.

Roth and Cummins acknowledge that information about lower-level implementation details can deepens our understanding of the systems whose capacities are targeted by functional analyses. But they nevertheless stress that, strictly speaking, this information should neither be interpreted as a proper part of the explanation nor as a requirement on adequate functional explanation. In their words, “having a fuller understanding of a system, in this sense, is not the same thing as having a more complete explanation of the [capacity] targeted for functional analysis”. Roth and Cummins instead suggest that there is a crucial distinction between explaining a capacity via functional analysis (what they call horizontal explanation) and explaining how a functional analysis is implemented (what they call vertical explanation) – a distinction which, in their opinion, has been repeatedly elided or conflated in the literature. In other words, functional explanation is not mechanistic explanation.

While evidence from neuroscience is relevant to determining which functional analysis is correct, they argue that the specific role that details about underlying neural mechanisms plays is one of confirmation not explanation. As they put it, “bringing such knowledge to bear in this instance would be an exercise in confirming a proposed analysis, not explaining a capacity”. Their discussion provides an important clarification of the original, highly influential position first articulated by Cummins (1983). The
chapter also raises the stakes in the current debate, since it stands diametrically opposed to recent attempts by proponents of the mechanistic perspective to identify functional analyses as elliptical or incomplete mechanistic explanations (Piccinini and Craver 2011). This view will be taken up in detail below.

Along similar lines, in his chapter, Daniel Weiskopf argues that psychological models can be explanatorily adequate in the sense that they satisfy standardly accepted norms of good explanation such as providing the ability to answer a range of counterfactual questions regarding the target phenomenon and the ability to manipulate and control the target phenomenon, without necessarily being mechanistic. A cognitive model (sometimes also referred to as a “box-and-arrow” model; see Datteri and Laudisa 2014) involves a set of functionally interacting components each of which is characterized on the basis of its functional profile (and typically couched in representational or information-processing terms). According to Weiskopf, cognitive models can provide perfectly legitimate causal explanations of psychological capacities by describing the way information is represented and processed. Although these models evidently describe real causal structure, they do not embody determinate commitments about the neural mechanisms or structures underlying these capacities. They do not provide a specifiable decomposition of the target system into spatially localizable physical parts, and critically, these mechanistic details do not need to be filled in for the model to be endowed with explanatory force. Consequently, on Weiskopf’s view, psychological explanation is fundamentally different in kind to mechanistic explanation.

4.2 Computational explanation

Along closely related lines, others have argued that computational explanations of psychological capacities are different in character from the mechanistic explanations found in neuroscience and other life sciences. Sejnowski, Koch, and Churchland (1988) express one primary motivation for this view:

Mechanical and causal explanations of chemical and electrical signals in the brain are different from computational explanations. The chief difference is that a computational explanation refers to the information content in the physical signals and how they are used to accomplish a task. (Sejnowski, Koch, and Churchland 1988, 1300)

According to this perspective, computational explanations differ because they appeal to information-processing or representational or semantic notions, and this somehow makes them incompatible with the mechanistic framework. Recent advocates of the mechanistic approach counter that computational explanation can be readily understood as a species of mechanistic explanation despite having distinctive features (Bechtel 2008; Kaplan 2011; Piccinini 2007; Boone and Piccinini 2015; Piccinini 2007).

Others have attempted to draw a stark boundary between computational and mechanistic explanations by arguing that computational explanations are abstract or mathematical in a way that prevents their integration into the mechanistic framework (e.g., Chirimuuta 2014). On Chirimuuta’s view, computational explanations – even those constructed in computational neuroscience – embody a “distinct explanatory style” which “cannot be assimilated into the mechanistic framework” because they “indicate a mathematical operation – a computation – not a biological mechanism” (2014, 124). Since these
explanations are claimed to be highly abstract – focusing on the high-level computations being performed – they are supposed to enjoy a considerable degree of autonomy from low-level details about underlying neural mechanisms. This is a version of the multiple realizability claim encountered above. In his contribution to this volume, David Kaplan argues that this kind of claim rests on persistent confusions about multiple realizability and its implications for mechanistic explanation. Specifically, he argues against the lessons that Chirimuuta and others wish to draw from recent modeling work involving so-called canonical neural computations — standard computational modules that apply the same fundamental operations across multiple brain areas. Because these neural computations can rely on diverse circuits and mechanisms, modeling the underlying mechanisms is argued to be of limited explanatory value. They take this work as evidence that computational neuroscientists sometimes employ a distinctive explanatory scheme from that of mechanistic explanation. Kaplan offers reasons for thinking this conclusion is unjustified, and addresses why multiple realization does not always limit the prospects for mechanistic explanation.

In her contribution to the volume, Frances Egan argues for a position on the same side of the debate as Chirimuuta and others who seek to defend the distinctness and autonomy of computational explanation. Egan argues that a common type of explanation in computational cognitive science is what she terms function-theoretic explanation. Building on ideas from her earlier work (Egan 1995, 2010), she contends that this type of explanation involves articulating how a given system computes some mathematically well-defined function and that in performing this computation contributes to the target capacity in question. For example, Marr famously proposed that the early visual system performs edge detection by computing a well-defined mathematical function – the zero-crossing of second derivative filtered versions of the retinal inputs (i.e., the Laplacian of a Gaussian; $\nabla^2 G \ast I$). This is a paradigmatic function-theoretic explanation – it provides a mathematically precise specification of what the early visual system does and an adequate explanation of how it does it. According to Egan, function-theoretic characterizations can possess explanatory import even in the absence of details about how such computations are implemented in neural systems. Consequently, they are not properly interpreted as mechanistic explanations.

Views about autonomous computational explanation are often backed up by appeals to David Marr’s influential tri-level computational framework. According to Marr (1982), there are three distinct levels of analysis that apply to all information-processing systems ranging from digital computers to the brain: the computational level (a specification of what function is being computed and why it is computed), the algorithmic level (a specification of the representations and computational transformations defined over those representations), and the implementation level (a specification of how the other levels are physically realized). Marr’s discussion of the relationships between these levels appears to reinforce the idea of an autonomous level of computational explanation. First, he repeatedly prioritizes the relative importance of the computational level:
[I]t is the top level, the level of computational theory, which is critically important from the information-processing point of view. The reason for this is that the nature of the computations [. . .] depends more upon the computational problems to be solved than upon the particular hardware in which their solutions are implemented. (Marr 1982, 27)

This privileging of the computational level, coupled with the fact that his preferred methodology is top-down, moving from the computational level to the algorithmic, and ultimately, to implementation, has fostered the idea of an autonomous level of computational explanation. Second, in some passages, Marr appears to claim that there are either no direct constraints between levels or that the operative constraints are relatively weak and only flow downward from the computational level – claims that are clearly at odds with the mechanistic view. For instance, he states that: “since the three levels are only rather loosely related, some phenomena may be explained at only one or two of them” (1982, 25). If computational explanations were unconstrained by one another in this manner, this could certainly be used to draw a conclusion about an explanatorily autonomous level.

Nevertheless, there are numerous places where Marr sounds much more mechanistic in his tone (for further discussion, see Kaplan 2011). Although his general computational framework clearly emphasizes how one and the same computation might in principle be performed by a wide array of distinct algorithms and implemented in a broad range of physical systems, when the focus is on explaining a particular cognitive capacity such as human vision Marr appears to strongly reject the claim that any computationally adequate algorithm (i.e., one that has the same input–output profile or computes the same function) can provide an equally appropriate explanation of how the computation is performed in that particular system. After outlining their computational hypothesis for the extraction of zero-crossings in early vision, Marr quickly shifts gears to determine “whether the human visual system implements these algorithms or something close to them” (Marr and Hildreth 1980, p. 205; see also Marr et al. 1979a,b; Marr 1982). The broader context for this passage suggests that Marr did not view this as a secondary task, to be undertaken after the complete and fully autonomous computational explanation is given. Instead, Marr appears to be sensitive to the critical explanatory role played by information about neural implementation. On this interpretation, Marr’s view is much more closely aligned with the mainstream of contemporary computational neuroscience. Interestingly, Tomaso Poggio, one of Marr’s principal collaborators and a highly accomplished computational neuroscientist in his own right, recently espouses a view that similarly emphasizes the importance of elaborating the various connections and constraints operative between levels of analysis. He argues that real progress in computational neuroscience will only be achieved if we attend to the connections between levels (Poggio 2010).

In their contributed chapter, Oron Shagrir and William Bechtel shed further light on the nature of computational explanation and its status vis-à-vis the mechanistic approach. Like many seeking to understand computational explanation, they too engage with Marr’s foundational work on the topic. They focus their attention on what they view as an underappreciated feature of Marr’s (1982) account of the computational level of analysis. Marr defines the computational level as the “level of what the device does and why” (1982, 22). The role of the what-aspect is relatively straightforward, involving a specification of
what computation is being performed (or what mathematical function is being computed). The role of why-aspect is different – it specifies how the specific computations being performed are adequate to the information-processing task.

According to Shagrir and Bechtel, many interpreters of Marr have provided an incomplete analysis of the computational level because they have neglected the what-aspect. Part of the reason for this neglect is that Marr never provides a detailed and systematic account of this aspect of the computational level. In their chapter, Shagrir and Bechtel offer a plausible reconstruction of Marr’s views concerning the computational level. They maintain that the why-aspect characterizes why a particular computation is the one the system in fact needs to perform, given the structure of the physical environment in which it is embedded (i.e., the target domain). Marr (1982) calls these constraints imposed by the physical environment “physical constraints”, and implies that any visual system worth its salt must be capable of preserving certain structural relations present in the target domain (i.e., must be “designed” to reflect these physical constraints). However, Marr’s original discussion raises more questions than it provides answers. It is here that Shagrir and Bechtel make real headway. They argue that the why-aspect of the computational analysis provides a characterization of the structure-preserving mapping relation between the computed function and the target domain. It thus serves to relate the physical constraints to the computed function – and in doing so, it demonstrates the appropriateness of the computed function for the information-processing task at hand. This, according to Shagrir and Bechtel, is why the early visual system computes the Laplacian of a Gaussian as opposed to performing multiplication or exponentiation or factorization.

Shagrir and Bechtel also make the case that the computational level of analysis provides indispensable information for the construction of mechanistic explanations in so far as it specifies the target phenomenon to be explained in precise quantitative or mathematical terms. They argue that delineating scientific phenomena in general is an essential and highly non-trivial scientific task, and it is a specific prerequisite for building mechanistic explanations. Hence, another one of Marr’s great insights was to highlight the importance of having a clear and precise specification of the computational phenomenon in order to develop an explanation.

5. Mechanistic explanation
Advocates of the mechanistic approach to explanation have articulated a vision of disciplinary integration that neither bottoms out in classical theory reduction nor attempts to undermine arguments for the autonomy of psychology by challenging multiple realizability claims (Bechtel 1997, 2008; Piccinini and Craver 2011). According to many defenders of the mechanistic perspective, the traditional framing of the debate imposes a false choice between reduction and autonomy because it implies that these are mutually exclusive options. Bechtel, for example, maintains that the key to resolving this debate is understanding how the mechanistic framework enables a “rapprochement between reductionism and the independence of investigations focused on higher levels of organization” (Bechtel 2008, 158).

5.1 Modest reductionism afforded by the mechanistic approach
According to Bechtel (1997, 2008), the kinds of reductions achieved through mechanistic explanations, in contrast to those posited by the traditional theory reduction model, are fully compatible with a robust notion of autonomy for psychology and other special sciences. He states:

Within the mechanistic framework one does not have to reject reduction in order to allow for the independence of the higher-level sciences. The decomposition required by mechanistic explanation is reductionist, but the recognition that parts and operations must be organized into an appropriate whole provides a robust sense of a higher level of organization. (Bechtel 2008, 130).

Mechanistic explanations are reductionist in the specific sense that they seek to explain the overall pattern of activity or phenomenon generated by the mechanism as a whole by appealing to lower-level component parts and their activities. Yet despite this reductionist character, it is claimed to be (more) palatable to anti-reductionists because mechanistic explanations involve a non-trivial form of autonomy in so far as the higher-level (spatial and temporal) organization of the components in a target mechanism is often essential to producing the phenomenon to be explained. According to Bechtel, “[m]odes of organization are not determined by the components but are imposed on them” (Bechtel 2007, 192). Furthermore, successful mechanistic explanations sometimes goes beyond describing the local mechanism and its underlying components because they appeal to conditions of the broader system or environment in which the mechanism is embedded and without which they could not perform their functions (Bechtel 2008). This too has been argued to secure the autonomy of higher-levels of organization and explanation that does not directly depend on multiple realizability.

Relatedly, the mechanistic framework embodies a distinctive account of levels of organization in mechanisms, which in turn affords a more modest view of reduction than the traditional theory reduction model. Whereas the traditional approach assumes that higher-level theories can be reduced in succession to increasingly lower levels until some fundamental level is reached, which in turn grounds all the higher levels, the mechanistic approach rejects this global account of reduction. Although mechanistic explanations are reductionist in the sense that they appeal to lower-level parts and their operations to explain some higher-level behavior of the mechanism, the reductions supported have a local character since there is no single fundamental level that globally grounds all higher levels of mechanisms. In stark contrast to traditional approaches that construe levels as global strata spanning the natural world (Oppenheim and Putnam 1958), levels of organization in mechanisms are local in the sense that they are defined relative to a given mechanism (Bechtel 2008; Craver 2007). In a particular mechanistic context, two arbitrary elements are deemed to reside at the same mechanistic level only if they are components in the same mechanism, and they occupy a higher or lower level depending on how they figure into a componential or part-whole relation within a mechanism. Critically, questions concerning whether components of a given mechanism (or the mechanism as a whole) reside at a higher, lower, or the same level as entities outside the mechanism are not well defined (Bechtel 2008; Craver 2007).

5.2 Functional analysis as elliptical mechanistic explanation
Along somewhat different lines, Piccinini and Craver (2011) maintain that the mechanistic perspective encourages a rethinking of the received view of psychological explanation as a kind of functional analysis or functional explanation (e.g., Cummins, 1975, 1983, 2000; Fodor 1967), which eliminates all commitments to autonomy. Piccinini and Craver (2011) reject the received view and instead argue that functional and mechanistic explanations are neither distinct nor autonomous from one another precisely because functional analysis, when properly constrained, provides a kind of mechanistic explanation – partial or elliptical mechanistic explanation. Mechanistic explanations, which are prevalent throughout the biological sciences including neuroscience, involve the identification of the mechanism responsible for maintaining, producing, or underlying the phenomenon of interest (Bechtel 2008; Bechtel and Richardson 1993/2010; Craver 2007; Craver and Darden 2013; Machamer et al. 2000). They maintain that this shift in perspective will open up a pathway for “building a unified science of cognition” (2011, 284). Piccinini and Craver’s (2011) main claim is as follows:

The core idea is that functional analyses are sketches of mechanisms, in which some structural aspects of a mechanistic explanation are omitted. Once the missing aspects are filled in, a functional analysis turns into a full-blown mechanistic explanation. By this process, functional analyses are seamlessly integrated with multilevel mechanistic explanations. (Piccinini and Craver 2011, 284)

According to Piccinini and Craver (2011), a functional analysis is a mechanism sketch in which the capacity to be explained is decomposed into sub-capacities, yet most if not all of the information about the underlying structural components or parts is omitted. According to the mechanistic perspective they endorse, structural information provides an essential source of constraints on functional analyses. It must be incorporated if a given analysis is to count as revealing the causal organization of the system and in turn explanatory. As they put it:

If the connection between analyzing tasks and components is severed completely, then there is no clear sense in which the analyzing sub-capacities are aspects of the actual causal structure of the system as opposed to arbitrary partitions of the system’s capacities or merely possible causal structures. (Piccinini and Craver 2011, 293)

Once the missing structural information about the components underlying each identified sub-capacity is filled in, the mechanism sketch is transformed into a (more complete) mechanistic explanation.

---

1 Arguably, a version or precursor of this view was articulated and defended much earlier by Bechtel and Richardson (1993/2010). In that work, they repeatedly emphasize how both functional and structural decompositions of a target system (decomposition and localization, respectively) must be incorporated to produce adequate mechanistic explanations. Decomposition “allows the subdivision of the explanatory task so that the task becomes manageable and the system intelligible” and “assumes that one activity of a whole system is the product of a set of subordinate functions performed in the system. (Bechtel and Richardson 1993/2010, 23). In addition, the decomposed sub-capacities must also be assigned to structural components of the underlying mechanism. In other words, they must be localized. Localization involves the “identification of the different activities proposes in a task decomposition with the behavior or capacities of specific components. (Bechtel and Richardson 1993/2010, 24). Therefore, according to their view, identifying either the functional or structural or properties of a system alone will fail to yield an adequate mechanistic explanation. Instead, mechanistic explanation requires both a functional and structural analysis of the target system. These are complementary, not independent or competing endeavors.
The proposed picture involves a rejection of both DISTINCTNESS and AUTONOMY. Since functional analysis is conceived as a kind of mechanistic explanation – elliptical mechanistic explanation – it cannot be distinct from mechanistic explanation. Because distinctness is a necessary condition for autonomy, the view also entails a rejection of AUTONOMY. Beyond this, the view also embodies a positive account of the interaction between the explanatory frameworks of psychology and neuroscience. The identification of sub-capacities in a functional analysis is argued to place very real and direct constraints on which components can engage in those capacities. In particular, the analysis generates, at a minimum, the expectation that for each identified sub-capacity there will be a corresponding structure or set of structures that implements the capacity. This is what is supposed to help to ensure that the proposed functional decomposition goes beyond providing a merely arbitrary partitioning of the system, and succeeds in revealing its real causal structure. On this mechanistic view, the explanatory projects of psychology and neuroscience coincide and are deeply intertwined because both provide complementary and mutually constraining descriptions of different aspects of the same multilevel mechanisms. One describes function. The other describes underlying structure.

In their contribution to the volume, Corey Maley and Gualtiero Piccinini aim to provide a suitable foundation for functional ascriptions at the heart of the mechanistic enterprise. Mechanistic explanations involve the identification of underlying component parts and attributions of specific functions performed by those components. Yet surprisingly little work has been done to investigate/explore what underwrites/grounds these functional ascriptions (for a notable exception, see Craver (2001, 2013). Having an account of functions in hand would, for example, allow one to distinguish cases that justify the ascription of particular functions from those that do not. Maley and Piccinini contend that understanding how functions are ascribed to neural and cognitive mechanisms and their parts is critical for a fully adequate account of multi-level mechanistic explanation.

They reject standard etiological accounts of function, which face many well-known criticisms including that the selective or evolutionary histories proposed to ground functional attributions are often exceedingly difficult if not impossible to discover and so routinely remain unknown. Relatedly, it is frequently objected that functions are often plausibly attributed in the absence of historical information about a system. They also reject causal role accounts, which successfully avoid the discovery problem by grounding function in a system’s current causal powers, but nevertheless face a different set of challenges. It is widely argued that causal accounts involve an overly permissive concept of function, which makes it difficult to define a counterpart notion of malfunction and relatedly distinguish between how things ought to work (their proper functions) from how they in fact work. For these reasons, Maley and Piccinini instead develop a teleological account of function according to which functions are defined in terms of their stable contribution to a current objective goal of a biological organism (e.g., survival or inclusive fitness). They maintain that a primary advantage of their account is that, like standard causal accounts, functions are grounded in current causal powers. However, unlike standard accounts, theirs is claimed to be more
restrictive such that a distinction between function/malfunction can be drawn.

The mechanistic perspective thus appears to offer several a number of promising routes to achieving explanatory integration or unification of mind and brain science, while at the same time, undermining the historically influential view of autonomous psychological explanation. But, like the philosophical views canvassed above, it too faces obstacles. One primary objection is that in treating functional analyses in psychology as elliptical mechanistic explanations to be filled in by neuroscience, the prospects for a sufficiently weighty or substantive form of autonomy for higher-level psychological explanation becomes rather bleak. A number of challenges along these lines are raised in contributions to this volume.

6. High-level causal explanation
Recent philosophical work on mechanistic explanation is often interpreted as having undesirable imperialistic tendencies. In his contribution, James Woodward argues against the claim recently attributed to some proponents of mechanism that only mechanistic models in neuroscience and psychology explain. In particular, he seeks to combat the view that models which include more mechanistic detail will always be explanatorily superior to those that include less detail. This more-details-the-better view has been attributed to Kaplan and Craver (2011), among others. Woodward instead maintains that many successful explanatory models across both neuroscience and psychology often purposefully abstract away from all manner of lower-level implementation (e.g., neurobiological or molecular) details in order to highlight just those factors that make a difference to whether or not the target phenomenon occurs (so-called difference-makers). Woodward claims that such models can and often do provide perfectly legitimate explanations, and that resources from the interventionist account of causal explanation can illuminate their explanatory status.

According to the interventionist approach, explanatory models permit the answering of what Woodward (2003) calls what-if-things-had-been-different questions. They identify conditions that, had they been otherwise, would “make a difference” to the target phenomenon to be explained. This includes conditions under which the target phenomenon would not have occurred, would have occurred at a different rate, etc. This requirement is important because it implies that successful explanations will pick out just those conditions or factors that are explanatorily or causally relevant to the phenomenon to be explained (i.e., the difference-makers). The notion of causal or explanatory relevance (or difference-making) is in turn cashed out in terms of interventions. Roughly, X causes (or is causally relevant to) Y just in case, given some set of background circumstances, it is possible to change Y (or the probability distribution of Y) by intervening on X. The notion of intervention is here understood in a technical sense common in the philosophical and statistical literature (e.g., Spirtes et al. 2000; Woodward 2003). The idea is that a causal relationship can be inferred between X and Y when the intervention is “surgical”, i.e., when the intervention on X changes the value of Y “directly” and does not involve changing the values of other
possibly confounding variables that could in turn change the value of Y (for further discussion, see Woodward 2003). Interventions are therefore naturally thought of as idealized (perfectly controlled and non-confounded) versions of real experimental manipulations routinely performed in the lab.

The interventionist approach stands to legitimize higher-level explanations in two ways. First, it opens up the possibility that sometimes relatively abstract, higher-level explanations can provide better explanations than more detailed, lower-level ones because the lower-level ones might include irrelevant or inessential details that obscure the difference-making factors. This is because the former, in focusing on the difference-makers, may omit details whose variations or changes make no difference to the target phenomenon. By contrast, the latter might include or cite irrelevant details that serve to mask those difference-making factors. Second, this particular way of thinking about causal relationships opens the door to higher-level causal explanations since higher-level factors such as attentional load, memory capacity, or general psychological state can in principle serve equally well as the targets of such interventions as lower-level neurobiological or molecular factors. Hence, the interventionist framework holds promise to illuminate the causal and explanatory relevance of high-level factors, and in doing so legitimize high-level, relatively abstract explanations found throughout psychology and neuroscience.²

In his chapter, Woodward focuses on relatively abstract neurobiological models such as conductance models and even the Hodgkin-Huxley model of the action potential, whose explanatory credentials has been subject to intense debate in the recent philosophical literature (e.g., Bogen 2005, 2008; Craver 2006, 2007, 2008; Kaplan 2011; Levy 2014; Levy and Bechtel 2013; Schaffner 2008; Weber 2008). Woodward’s general conclusion here is that the interventionist framework can be used to illuminate how models in neurobiology and psychology that abstract away from certain lower-level implementational details can nonetheless be explanatory. If successful, this secures a kind of partial autonomy of higher-level explanations and models from lower-level mechanistic details.

Woodward argues that higher-level psychological models need not be seen as automatically competing with lower-level neurobiological models. Whether the higher- or lower-level model is most appropriate, or provides a superior explanation, depends on the phenomenon one is trying to explain. Sometimes lower-level details about neural implementation will be causally relevant and so must be incorporated into the model if it is to be explanatorily adequate. Other times such details will be irrelevant to (make no difference for) the phenomenon of interest, and so can be safely ignored in one’s model without affecting its explanatory power. Woodward finds that modeling practices in psychology and neuroscience are often exquisitely sensitive to the goal of trying to include just enough detail to account for what one is trying to explain but no more. This message dovetails nicely with views commonly expressed by computational modelers who are continually trying to find the appropriate balance of detail and abstraction in their models so as to best account for the phenomenon of interest. For example, the computational neuroscientist Trappeenberg (2010) asserts that “[m]odels are intended to simplify experimental data, and thereby to identify which details of the biology are essential to explain particular

aspects of a system. (2010, 6). He is triangulating on the idea that simpler, relatively abstract models can often provide superior explanations in so far as they include only the “essential” details: Woodward provides justification for this idea. He aims to provide principled reasons for the neglect of lower-level mechanistic details when attempting to build explanatory models in mind and brain science. In particular, he argues that such details make be safely ignored precisely when they are causally and explanatorily irrelevant – they make no difference – to the phenomena under investigation. In these cases, higher-level explanations are not subject to constraint from facts about these lower-level details. The explanatory autonomy of psychology, according to this view, can be seen as stemming from the causal irrelevance of lower-level details about neural implementation. Variation in neural details sometimes makes no difference for the (occurrence of state/condition of the) phenomenon under consideration, and so they can be abstracted away from without explanatory repercussions.

In his contribution, Michael Strevens takes up similar themes. Like Woodward, he too seeks to shed light on higher-level causal explanations in sciences like biology, economics, and psychology, which seem to possess explanatory force despite the fact that they abstract away from – place “black boxes” around – many of lower-level mechanistic or implementational details. Although Strevens recognizes the intuitive pull of the idea that these models provide adequate explanations, he is cautious about embracing it.

Strevens carefully considers the challenges posed by convergent evolution for detail-oriented modeling approaches including the mechanistic approach. Because convergent evolution generates functional kinds that are instantiated by radically different physical realizations, modeling the underlying mechanisms is supposed to be of limited explanatory value. In such cases, more abstract or less detailed models appear to provide better (e.g., more unifying) explanations that those bogged down in the mechanistic details. Even worse, mechanistic explanation may seem entirely out of reach in such cases. For example, while there may well be some interesting high-level or abstract explanatory models or generalizations about wings, which are thought to have evolved independently approximately forty times in throughout history, a demand that their explanation satisfy the strictures of the mechanistic approach may go entirely unfulfilled since the mechanistic details vary considerably across these instances. (There are important parallels between these issues and those discussed in the chapter by Kaplan in this volume.)

Strevens recognizes the intuitive force behind this type of (multiple realizability-based) argument for the autonomy and explanatory superiority of abstract, higher-level explanations. He nonetheless maintains that sometimes models in which lower-level details are omitted or black-boxed can mistakenly be deemed explanatorily adequate and complete because of a subtle and unrecognized shift in the target phenomenon to be explained. Specifically, Strevens argues that there is a tendency to conflate the difference between explaining the common instantiation of the same functional kind (e.g., wing) by several different (types of) entities versus the instantiation of the functional kind by a single entity (e.g., the avian wing). According to Strevens, not only are these fundamentally different explananda, but they also require different explanations with varying amounts of mechanistic detail. Explanations of the former may be highly abstract (suppressing or “black-boxing” many or most of the underlying mechanistic details) in
order to highlight the common factor (or set of factors) causally relevant to the outcome across the different instances. But critically, the set of factors cited in such explanations is argued to fall well short of exhausting the complete set of factors relevant to any individual instantiation (e.g., the avian wing or the insect wing), and so these types of explanations will typically require considerably more mechanistic detail. Strevens suggests that some multiple realizability-based arguments for the explanatory autonomy of higher-level sciences including psychology similarly exploit this slippage in order to conclude that abstract explanations are superior to detailed ones. And while he agrees that models with more detail are not always better, he disagrees that models with less detail are always better. Instead, Strevens, like Woodward, maintains that the appropriate level of detail depends sensitively on the phenomena one wants to explain.

In the final contribution to the volume, Dominic Murphy addresses the role of folk psychology and its relation to the sciences of the mind and the brain. Is folk psychological explanation sui generis and therefore distinct and autonomous from scientific psychology and neuroscience? Or is it continuous with scientific approaches to the mind and brain, and therefore a potential candidate for integration? Folk psychology refers to the commonsense conceptual framework that all normally socialized humans use to understand, predict, explain, and control the behavior of other humans and higher non-human animals. Murphy identifies and explores three broad perspectives on folk psychology – integration, autonomy, and elimination. According to the integrationist perspective, folk psychology defines the phenomena that the cognitive and brain sciences seek to investigate and explain, and thus plays a permanent albeit limited role in scientific inquiry. According to the autonomist perspective, folk psychology comprises a perfectly legitimate explanatory framework but one that is fundamentally different and character and therefore incompatible or incommensurable with the explanatory frameworks of cognitive science and neuroscience. Whereas the explanatory framework of folk psychology operates at the level of people and their sensations, beliefs, desires, and intentions (often referred to as the “personal level”), the explanatory frameworks of cognitive science and neuroscience operate at the “sub-personal” level of the information-processing and brain mechanisms underlying these personal-level activities. According to the autonomist, folk psychology comprises a fully autonomous and self-contained domain of personal-level explanation that is neither confirmed nor refuted by empirical evidence from mind and brain science. According to the eliminativist perspective, folk psychology is a massively false theory that should be replaced in favor of another more predictively and empirically adequate scientific theory, presumably from neuroscience.

After quickly dismissing the autonomist perspective, Murphy focuses on exposing the limitations of the integrationist perspective. According to the integrationist, the job description for folk psychology is to specify the explananda for scientific psychology and cognitive neuroscience, and critically, that it has done reasonably well at completing this job. Murphy rejects the latter claim and argues that since the taxonomic divisions of folk psychology have been laid down independently of constraints from evidence about neural implementation they fail to limn the contours of the mind. Consequently, folk psychology cannot play the role integrationists envision for it. Instead, the explananda for the cognitive and brain sciences have themselves been subject to rather heavy revision in the light of information about the
workings and structure of the brain. Hence, Murphy argues, we are left in the position of endorsing eliminatativism as the only viable scientific option.

7. Conclusion

Understanding the multifaceted relationship between neuroscience and psychological science is vital to achieving progress across these scientific domains. Understanding the nature of explanation across these different domains provides one highly fruitful avenue for exploring these issues. Are the explanatory patterns employed across these domains similar or different in kind? To what extent do they inform and constrain each another? Or, are they autonomous? Questions of this sort concerning explanation and how this shapes our thinking about the prospects for integrating mind and brain science occupies center stage in this volume.

On the one hand, the emergence of cognitive neuroscience suggests that the integration of mind and brain science is already largely upon us or at very least an inevitable future outcome. Moreover, the growing dominance of the mechanistic approach to explanation further reinforces a picture unity and integration between explanatory frameworks. And yet, on the other hand, there nevertheless appears to be strong reasons for thinking that a psychological science will, over the long-term, retain some partial degree of explanatory autonomy. Although a final resolution continues to elude us, the chapters contained in this volume succeed in pushing this important debate forward.

References


http://doi.org/10.3389/fpsyg.2014.00464


http://doi.org/10.1162/jocn.1989.1.1.2


