Studies in History and Philosophy of Science xxx (2015) 1–10

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa

Giving up on convergence and autonomy: Why the theories of psychology and neuroscience are codependent as well as irreconcilable

Eric Hochstein

Washington University, St. Louis, United States

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Psychology;
Neuroscience;
Theory reduction;
Idealization;
Autonomy;
Convergence

ABSTRACT

There is a long-standing debate in the philosophy of mind and philosophy of science regarding how best to interpret the relationship between neuroscience and psychology. It has traditionally been argued that either the two domains will evolve and change over time until they converge on a single unified account of human behaviour, or else that they will continue to work in isolation given that they identify properties and states that exist autonomously from one another (due to the multiple-realizability of psychological states). In this paper, I argue that progress in psychology and neuroscience is contingent on the fact that both of these positions are false. Contra the convergence position, I argue that the theories of psychology and the theories of neuroscience are scientifically valuable as representational tools precisely because they cannot be integrated into a single account. However, contra the autonomy position, I propose that the theories of psychology and neuroscience are deeply dependent on one another for further refinement and improvement. In this respect, there is an irreconcilable codependence between psychology and neuroscience that is necessary for both domains to improve and progress. The two domains are forever linked while simultaneously being unable to integrate.

There is a long-standing debate in the philosophy of mind and the philosophy of science regarding how best to interpret the relationship between the theories and models of neuroscience, and those of psychology. Both domains are dedicated to the scientific study and explanation of cognitive behaviour, yet each domain appears to explain and predict this behaviour by appealing to distinct sets of theories and models, and by employing different kinds of concepts and categories. How then can we make sense of the relationship between these different domains, and move forward in our scientific understanding of such behaviour? Traditionally, philosophers of science have proposed one of two possible options for how this relationship might ultimately be understood, and their seemingly conflicting accounts resolved:

1. As neuroscience and psychology improve and change over time, the theories and models of both domains will slowly co-evolve together, each undergoing alterations and changes until they converge on a single unified theory of cognitive behaviour (e.g. Bickle, 1998, 2003, 2006; Boone & Piccinini, 2015; Churchland, 1989; Craver, 2007; Piccinini & Craver, 2011).
2. Neuroscience and psychology will not converge because the two domains characterize systems at different levels of organization. The theories of psychology characterize functional states of systems that can be realized in different ways by different mechanisms, while the theories of neuroscience only characterize the physical implementation of neurological mechanisms. Given that the functional properties and regularities of psychology exist autonomously from any one system that realizes them, the theories of psychology are therefore irreducible to those of neuroscience. As a result, psychology and neuroscience will proceed largely in isolation from one

E-mail addresses: ehochstein@hotmail.com, eghochst@gmail.com.

http://dx.doi.org/10.1016/j.shpsa.2015.10.001
0039-3681/

another (e.g. Aizawa & Gillett, 2011; Burge, 2010; Crane, 2001, pp. 62-66; Cummins, 1983; Fodor, 1974, 1998; Johnson-Laird, 1983; Menzies & List, 2010).

In this paper, I argue that progress in neuroscience and psychology has been made possible by the very fact that both of these positions are false. More specifically, I propose that scientific progress in both domains is contingent on their theories being irreconcilable with one another in various respects (making convergence impossible), but also on the fact that the different theories and models do *not* identify states and properties that exist autonomously from one another. The theories and models of psychology and neuroscience are deeply dependent on one another for further refinement and improvement, yet this dependence does not imply the eventual convergence of the two disciplines.

Human behaviour is the product of a vast number of causal influences, from historical, to biological, to environmental. The sheer complexity of the causal influences at work means that we often must employ distinct scientific theories with radically different idealizing and simplifying assumptions depending on which of these influences we wish to study, and which we do not. Different idealizations will be used as the foundation for different types of scientific theories depending on which representational goals we seek to satisfy.

When it comes to understanding the relationship between psychology and neuroscience, the relevant question therefore becomes: what are the representational goals of neuroscience, and how do they differ from those of psychology? I propose that in virtue of representing different aspects of cognitive systems, the two domains must adopt different idealizing assumptions about the target system, resulting in vastly different and incompatible sets of theories useful for their own representational purposes, but not the other's. Convergence between these domains would therefore require that they give up the very idealizing assumptions that allow them to effectively represent the different aspects of the cognitive system we use them to study.

The fact that the models and theories employed by psychology are not useful for the same representational tasks as those employed in neuroscience (and vice versa) has led a number of philosophers to mistakenly infer that the two domains operate autonomously from one another, with the theories and findings of one domain being largely unhelpful to the theories and findings of the other. I will demonstrate that such a view is false, and is not supported by empirical research. There is strong empirical evidence that as we develop more detailed psychological theories and models, it puts essential constraints on what the neural mechanisms of the system are, and how they operate. Likewise, the more we know about the underlying neurological architecture of a system, the more it constrains the sorts of psychological generalizations we can make about it. As such, while psychology and neuroscience will not converge towards a single unifying account, neither can they stand apart from each other. This is not a problem that must be overcome, however, but is in fact a virtue that makes scientific understanding possible. It is the very tension between the irreconcilability of these different theories, and their required codependence, that drives scientific practice forward.

In order to make this argument, I begin in Section 1 by discussing how the relationship between psychology and neuroscience has been traditionally conceived. In Section 2, I demonstrate why these options are inappropriate for understanding the relationship that exists between the two domains. Lastly, in Section 3, I argue for an alternative account that justifies both the irreconcilability of psychology with neuroscience, as well as their necessary codependence. I end by demonstrating why this irreconcilable codependence is essential for scientific progress.

1. Traditional characterizations of the psychological/neuroscientific divide

1.1. Convergence

The motivating assumption that underlies the argument for the convergence of neuroscience and psychology is that both domains share the same general goal of developing an ideally correct theory of cognitive behaviour, but differ in their approaches for achieving it. Psychology is an attempt to understand cognitive behaviour by employing a largely "top down" approach, while neuroscience is an attempt to understand this same behaviour from a "bottom up" perspective. In other words, psychology attempts to understand cognitive behaviour by identifying and characterizing the high-level cognitive capabilities and deficits of the system, the behavioural patterns displayed by the system, and the environmental contexts in which certain behaviours appear. They then use this information to draw conclusions about what the underlying neurological mechanisms of the system must be like. Neuroscience, meanwhile, starts by studying the neurological mechanisms themselves, and then uses this information to draw conclusions about what the overall cognitive behaviour of the system is likely to be in various situations. Both domains therefore directly inform and constrain one another. Knowing more about the underlying mechanisms of the system informs our understanding of how the system will behave. This allows us to change and improve our psychological models to better account for this information. Likewise, the more detailed our psychological theories become regarding the overall behaviour of the system, the more it informs our understanding of what the neurological mechanisms are doing, and thus puts constraints on what their underlying architecture is. As Patricia Churchland notes:

> Crudely, neuroscience needs psychology because it needs to know what the system does; that is, it needs high-level specifications of the input-output properties of the system. Psychology needs neuroscience for the same reason: it needs to know what the system does. That is, it needs to know whether lower-level specifications bear out the initial input-output theory, where and how to revise the input-output theory, and how to characterize processes at levels below the top. (Churchland, 1989, p. 373)

A similar claim is made by Boone and Piccinini (2015), who argue that:

> The upshot is that cognition cannot be explained without accounting for the ways in which structures constrain functions and vice versa. In the long run, the mutual constraints between structures and functions lead cognitive psychologists and neuroscientists to look to each other's work to inform their analyses. […] The best strategy is to investigate both structures and functions simultaneously. […] This is the main driving force between the merging of neuroscience and cognitive psychology into cognitive neuroscience. (pp. 14-15)

Under these accounts, the concepts, categories, and theories of both domains will be constantly changing as they are continuously altered to better fit with the emerging findings of the other domain. This process of mutual refinement continues until a single unified account of the system is developed.

Please cite this article in press as: Hochstein, E., Giving up on convergence and autonomy: Why the theories of psychology and neuroscience are codependent as well as irreconcilable, Studies in History and Philosophy of Science (2015), http://dx.doi.org/10.1016/j.shpsa.2015.10.001

It is also worth noting that many philosophical arguments for the reduction of psychology to neuroscience, or the elimination of psychology in favour of neuroscience, can likewise be folded into this general project of convergence. Reduction is often thought to be the endpoint of the convergence process. Churchland, for instance, claims that:

> From the reductionist viewpoint, this possibility [that psychological categories will not be able to map onto neurobiological categories] does not look like an obstacle to reduction so much as it predicts a fragmentation and reconfiguration of the psychological categories. (1989, p. 365)

As evidence for such a claim, Churchland offers the psychological concept of "memory" as a case study. While the original folk notion of memory assumed that there was a single unified memory system operating within the brain, neurological research has revealed a number of distinct memory systems. Thus the concept of "memory" in psychology has fragmented from a single category into multiple categories in order to better fit with our discovery of the underlying neurological kinds. Through this continual fragmentation and reconfiguration of our psychological categories, she proposes that the eventual reduction of psychology to neuroscience "is more or less inevitable" (Churchland, 1989, p. 374).

Likewise, arguments for the elimination of certain psychological concepts from science are often defended on the grounds that such concepts are extremely unlikely to survive this process of fragmentation and refinement that drives convergence (see, for example, Bickle, 1998, 2003, 2006; Churchland, 1981; Hooker, 1981; Stich, 1983). Paul Churchland, for instance, in his seminal defence of Eliminative Materialism argues that certain mental categories, specifically those characterized as propositional attitudes (which he labels "folk psychology"), are so different from the sorts of categories that will survive the convergence process that we ought to abandon them. In his words,

> …the eliminative materialist is […] pessimistic about the prospects of reduction, but his reason is that folk psychology is a radically inadequate account of our internal activities, too confused and too defective to win survival through intertheoretic reduction. On his view it will simply be displaced by a better theory of those activities. (1981, p. 72)

The underlying assumption is that some psychological categories are so different from the sorts of categories that will emerge from a convergence of psychology and neuroscience that they are extremely unlikely to make it through the process in any recognizable form. As a result, deliberately keeping such concepts and categories as part of our scientific theories will only cause problems for future scientific endeavours and stands in the way of the development of a unified account of cognitive behaviour.

This general account is also compatible with more recent theories of convergence that base their accounts on the mechanistic nature of explanation in the life sciences (see: Boone & Piccinini, 2015; Craver, 2007; Piccinini & Craver, 2011). These positions emphasize the fact that cognitive behaviour is the product of complex neurophysiological mechanisms that can be described at different grains of abstraction. Under these accounts, psychological models and theories provide abstract descriptions of these mechanisms, often in the form of functional analyses. In this respect, they act as *sketches* of the neurophysiological mechanisms being studied, with many structural details omitted. This sketch can then be used to guide neuroscientific research, which in turn fills in this sketch with more explicit details of the structures and causes at work in the system. Through this process, psychology and neuroscience converge on a single unified neuroscientific account. As Piccinini & Craver put it:

> Functional analysis of a system's capacities provides a sketch of a mechanistic explanation. […] Thus, if psychological explanation is functional, as so many people assume, and psychological explanation is worthy of its name, then psychological explanation is mechanistic. Once the structural aspects that are missing from a functional analysis are filled in, functional analysis turns into a more complete mechanistic explanation. By this process, functional analyses can be seamlessly integrated with mechanistic explanations, and psychology can be seamlessly integrated with neuroscience. (2011, pp. 307-308)

All of these views implicitly assume that psychology and neuroscience are both striving towards the same goal of a single unified account of cognitive behaviour, and so will inevitably converge along the way towards this goal. This position can be contrasted with another view that has gained a great deal of prominence in the philosophy of mind. It is to this contrasting view we turn to next.

1.2. Autonomy

What is likely the most commonly held view among contemporary philosophers of mind regarding the relationship between psychology and neuroscience is that the explanations of the two domains are largely autonomous from one another. This idea is primarily based on appeals to computationalism and multiple-realizability. More specifically, it is argued that psychological theories and models are best construed as identifying *computational* states of complex cognitive systems, but not the way in which these states are realized in any one system. Given that computational states are defined functionally, systems that are physically implemented in numerous different ways can all instantiate the same computational states just so long as they have the appropriate functional organization. In contrast, neuroscientific theories and models only describe the implementation of neurological systems, and not the multiply realizable functional states characterized by psychology. In this respect, the two domains cannot converge on a single unified theory given that they cross-classify cognitive systems in order to identify autonomous sets of properties and relations (for an elaboration on this position, see: Aizawa & Gillett, 2011; Burge, 2010; Crane, 2001, pp. 62-66; Cummins, 1983; Fodor, 1974, 1998; Johnson-Laird, 1983; Menzies & List, 2010).

Put another way, the reason that we can never use a neuroscientific model to characterize the regularities identified by a psychological model is because the computational states identified by the psychological model, and the causal powers they have, are autonomous from any particular implementation that might realize them. As Menzies & List argue:

> Given that a mental state is typically realized in many different ways, we can expect that whatever causal powers it has, it has them independently of the particular way it is realized. In other words, we can expect that a mental state's causal powers do not depend on which of its possible realizers happens to be the actual one. (2010, p. 121)

This particular account of the relationship between neuroscience and psychology has gained a great deal of traction in the philosophy of mind over the years, to the point of being described as having "such overwhelming *prima facie* plausibility that the burden of

Please cite this article in press as: Hochstein, E., Giving up on convergence and autonomy: Why the theories of psychology and neuroscience are codependent as well as irreconcilable, Studies in History and Philosophy of Science (2015), http://dx.doi.org/10.1016/j.shpsa.2015.10.001

proof is on the critic to come up with reasons for thinking otherwise" (Block, 1980, p. 178). According to such accounts, psychology and neuroscience will proceed largely in isolation from one another, with convergence between the two domains being impossible.

So of these two possible accounts (convergence and autonomy), which is the best for understanding the relationship between psychology and neuroscience? I propose that neither are appropriate, and that a new alternative is required. Before offering such an alternative, it is important to understand why both existing positions run into problems.

2. Where the traditional accounts go wrong

2.1. Problems with convergence

At the heart of the convergence account is the idea that through sufficient refinements to the theories of psychology and neuroscience, their convergence towards a single unified account is inevitable. Yet there are good reasons to deny this idea. In order to see why, let us explore the sorts of complications and complexities that would have to emerge from the convergence project.

Recall that according to the convergence account, the theories and models of neuroscience provide a "bottom up" approach to studying cognitive behaviour by showing how interventions in the neurological mechanisms responsible for cognition affect their output. Meanwhile models and theories from domains like social psychology, developmental psychology, and community psychology, provide a "top down" study of cognitive behaviour by characterizing how these same neurological mechanisms behave when embedded in various social and environmental contexts. The underlying assumption is that both neuroscience and psychology are studying the same neurological mechanisms, but from different angles. Consider Piccinini & Craver's claim that:

> Insofar as psychologists pursue constitutive explanations, they ought to acknowledge that psychological explanations describe aspects of the same multilevel neural mechanism that neuroscientists study. (2011, p. 285)

But is this true? While it is certainly the case that psychological theories and models describe cognitive behaviour produced by systems that are *partially* constituted by multilevel neural mechanisms, it does not follow from this that psychological theories and models are describing *only* neural mechanisms. After all, cognitive behaviour is not *solely* the product of neural mechanisms. Additional causal influences include genetic, epigenetic, historical, environmental, dynamic, developmental, socio-economic, cultural, and the embodied characteristics of the system, to list only a few.[1] In virtue of characterizing and predicting the behaviours and capacities of entire cognitive systems, the models and theories of psychology are intended to describe far more than the neurological mechanisms identified by neuroscientists. The theories and model of neuroscience characterize and predict only a small subset of the causes of cognitive behaviours studied by psychologists. Therefore forcing the concepts and theories of psychology to fragment and change in order to map directly onto the neurological mechanisms of the brain is counterproductive. By doing so, psychological theories will become better at describing neurological mechanisms in particular, but worse at representing the behaviour of entire cognitive systems.[2]

Another problem facing the convergence account is that the hierarchical nature of biological mechanisms further complicates any account of convergence between the two domains. As is often the case with biological mechanisms, the parts of a given mechanism can satisfy the criteria for being mechanisms themselves, creating a hierarchy of levels whereby mechanisms are embedded within larger and more complex mechanisms (for a detailed account of this sense of compositional levels within biological mechanisms, see: Bechtel, 2007, 2008; Boone & Piccinini, 2015; Craver, 2007, 2009; Machamer, Darden, & Craver, 2000; Piccinini & Craver, 2011).

One important feature of such multi-level mechanisms is that the behaviour of higher level mechanisms depends not only on the behaviour of the lower level mechanisms, but also on the way in which these lower level mechanisms are organized together so as to instantiate the higher level system. The behaviour of the lower level mechanisms in isolation will be very different from the behaviour of those mechanisms when embedded within the context of the larger system. This means that the theories most useful for characterizing the behaviour of the lower level mechanisms will not be sufficient to account for the way in which the integrated system as a whole behaves. This is further exacerbated

[1] To complicate matters further, many of these causes interact together in complex ways. In doing so, they directly alter the behaviour of neural mechanisms, and change the sorts of contributions they make to cognitive behaviour.

[2] It should be noted that this does not mean that we *never* encounter instances of local reductions between psychological theories and neuroscientific ones. In some cases, psychological capacities have been shown to map directly onto certain neural mechanisms. Domains like cognitive psychology and cognitive neuroscience have developed explicitly as an attempt to identify such local reductions. Yet it is a mistake to assume that the success of particular local reductions implies that all, or even most, theories of psychology are likely to reduce to neuroscientific theories in this way, since many complex cognitive behaviours characterized by psychologists are known to be dependent on causal features beyond the mechanisms studied by neuroscientists.

Ironically some advocates of convergence have raised this very objection against the theories of convergence proposed by others, while implicitly making the very same mistake themselves. John Bickle (2003), for example, points to specific cases of local reductions to argue that the theories of psychology and neuroscience will inevitably converge on a unified account of cognitive behaviour in terms of the theories of molecular neuroscience. Meanwhile, Boone & Piccinini have countered by arguing that…

> The main problem with this form of reductionism is that specific molecular events are at best only partial explanations of cognitive phenomena. […] Molecular events are only relevant to the extent that they occur within specific neural structures, and locating the relevant neural structures requires more than purely molecular neuroscience. (2015, p. 6)

In other words, even though there are some specific instances where a cognitive capacity can be locally reduced to particular molecular events, this does not imply that all cognitive behaviour can be understood strictly in terms of such events, as the causal mechanisms and influences required to produce many cognitive behaviours require more than the molecular mechanisms characterized by molecular neuroscience.

Yet despite making this observation, Boone & Piccinini go on to claim that psychology and neuroscience are converging on a unified account of cognitive behaviour in terms of the theories of *cognitive* neuroscience instead of *molecular* neuroscience. They claim to avoid Bickle's mistake by noting that cognitive neuroscience does not characterize neurological mechanisms only at the molecular level, but that they characterize such mechanisms at multiple levels simultaneously. Meanwhile, they still insist that "every level of a neurocognitive mechanism is neural—or more precisely, every level is either (at least partially) composed of neurons or is a component of a neuron" (2015, p.15). Yet this falls into the same trap that snared Bickle. The theories of cognitive neuroscience used to characterize such multi-level mechanisms similarly provide only a *partial* explanation of cognitive phenomena. Historical, developmental, and environmental influences play an essential role in the production of cognitive behaviours, and the study of these influences go well beyond the boundaries of cognitive neuroscience. Thus even if there are instance of local reductions between certain psychological theories and certain cognitive neuroscientific ones, this does not mean that all cognitive capabilities of the system studied by psychologists can be understood strictly in terms of theories of cognitive neuroscience.

Please cite this article in press as: Hochstein, E., Giving up on convergence and autonomy: Why the theories of psychology and neuroscience are codependent as well as irreconcilable, Studies in History and Philosophy of Science (2015), http://dx.doi.org/10.1016/j.shpsa.2015.10.001

by the fact that the way in which the entire system interacts with structures and causes external to it can also change its behaviours in ways that cannot be predicted from an account of its lower level mechanisms alone. As William Bechtel notes:

> In virtue of being organized systems, mechanisms do things beyond what their components do. But beyond this, the organization of the components typically integrates them into an entity that has an identity of its own. (2007, p. 186)

Moreover, as we move up mechanistic levels (from simple neural mechanisms, to orchestrated higher level neural mechanisms, to larger physiological systems more generally), the mechanistic system under investigation increases in scale and complexity. To compensate for this, scientific theories based on distinct idealizing assumptions will need to be created in order to represent the behaviour and capacities of the higher level system in a manner that remains manageable and applicable.

This means that the scientific theories best used to characterize the unique behaviour of entire integrated cognitive systems will not be the same as those best used to characterize its lower level mechanisms, and will have as their underlying foundation idealizations and simplifying assumptions not found in the theories used to characterize the behaviours of the sub-mechanisms that compose them.

In recent years, a growing number of philosophers have acknowledged this necessity in adopting multiple models with conflicting idealizations in the study of complex systems (e.g. Hochstein, 2015; Kellert, Longino, & Waters, 2006; Parker, 2006; Wimsatt, 2007; Weisberg, 2013, pp. 103-105). Yet, in many cases scientists can still use the same underlying *theories* as the basis for the creation of conflicting *models*. One might, for instance, use the same underlying theories of neuroscience to construct two conflicting models of the action potential. One might idealize the morphology of the neuron, while the other might idealize electrical properties of the sodium and potassium channels. In these cases, we introduce idealizations for pragmatic purposes, but still adopt the same underlying theories as our foundation for the different models.

With the models of psychology and neuroscience, on the other hand, the idealizations involved go far deeper than this. Instead of a single theory used as the basis for the construction of both psychological and neurological models, the sorts of idealizations that the two domains adopt in their practices are used as the foundation for distinct types of scientific theories altogether. These theories are largely incompatible with one another given that the theories of each domain often cannot work with the necessary idealizations that the other domain needs to function.

For instance, given that the behaviour of an entire cognitive system is the product of both neurological and non-neurological causes, psychologists cannot simply take the underlying theories of neuroscience to create useful psychological models of the behaviour of the whole cognitive system; too many other causal variables would be missing.

Instead, a useful psychological theory will allow for the creation of elegant and testable hypotheses about the distinctive behavioural capacities and patterns of the cognitive system as a whole, without getting bogged down by the sheer volume of causal influences which contribute to those behaviours. In this respect, instead of appealing to neuroscientific theories as a basis for characterizing these cognitive behaviours, psychological theories posit idealized states and causes that often have no place in neuroscientific theories, but which mitigate complexity to help represent higher level behavioural patterns.

For example, theories in psychology which attribute "beliefs", "intentions", and other propositional attitudes to cognitive systems in order to understand and predict their behaviours have as their foundation the idealized assumption that such systems will behave rationally. In other words, understanding a cognitive system in terms of what it believes, intends, and desires, can only be informative if we assume that the system will act in accordance with the rational connection between those mental states (i.e. that it will do what it intends, pursue what it wants, and make decisions based on its beliefs). If the system acts irrationally, then its behaviour becomes indecipherable in terms of the ascription of propositional attitudes (for discussion, see: Davidson, 1974; Dennett, 1987; Føllesdal, 1982; Sehon, 1997). This assumption of rationality is a deliberate idealization (as most people violate the norms of rationality fairly regularly), but one that has helped psychologists to construct theories that are predictive of a range of complex cognitive behaviours.[3]

Yet this idealization puts these psychological theories at odds with many theories employed by neuroscientists. Assumptions of rationality often have no place in our neuroscientific understanding of the structural and causal properties of neural mechanisms. Instead of using rationality as the foundation for constructing neuroscientific theories, neuroscientists instead treat neural mechanisms as neither rational nor irrational. The idealization of rationality is one that has no role to play in such neuroscientific theories (for explicit arguments to this effect, see: Churchland, 1981; Dennett, 1987, p. 342; Quine, 1960, p. 218; Sehon, 1997; Stich, 1983).[4]

In fact, many have argued that propositional attitudes themselves should be treated as idealizations: mental categories that do not correspond to any neurological categories or types of mechanisms (see: Bickle, 1998, 2003, 2006; Churchland, 1981; Stich, 1983). While some have used this to argue for the elimination of propositional attitudes from scientific psychology, others have noted that such mental categories have allowed for the creation of more manageable theories which can be used to form predictions of high level behaviours that neuroscientific theories are simply ill-equipped to provide (for more, see: Hochstein, 2012, 2013).

Another example is the use of psychological theories that make reference to "emotions", or specific emotional categories such as "anger". It has been argued that these emotional categories subsume disparate and unrelated neurological mechanisms and processes under the same ill-fitting categories, and thus they have no place in a correct theory of the mechanisms responsible for behaviour (see: Griffiths 1997, 2004). However, if we think of these emotion categories as idealizations, as simplified categories used to mitigate the complexity of the system in order to better characterize its overall behavioural capabilities, then there is evidence that psychological theories which employ such categories have

[3] Examples of psychological models that interpret cognitive systems in terms of propositional attitudes, and which have been shown empirically to be predictive of a range of cognitive behaviours, include: the Belief-Desire-Intention Model of Agency (Bratman, 1987; Georgeff, Pell, Pollack, Tambe, & Woodridge, 2003), the Theory of Reasoned Action (Blue, 1995; Fishbein & Ajzen, 1981; Hausenblas, Carron, & Mack, 1997), and the Theory of Planned Behaviour (Ajzen, 1985, 1988, 1991; Armitage & Connor, 2001; Connor & Sparks, 1996).

[4] It is worth noting that there are some instances where neuroscientists do appear to describe the behaviour of complex neural systems as if they are behaving rationally, but even in these cases the assumption of rationality is not part of the neuroscientific theories used to characterize and explain such behaviours. When developing these theories, this assumption of rationality is abandoned in order to characterize the workings of the mechanisms responsible for them. In contrast, this idealized assumption of rationality is required in order for many psychological theories to function at all. To abandon this assumption is to cause the theories which rely on them to become incoherent and inapplicable.

Please cite this article in press as: Hochstein, E., Giving up on convergence and autonomy: Why the theories of psychology and neuroscience are codependent as well as irreconcilable, Studies in History and Philosophy of Science (2015), http://dx.doi.org/10.1016/j.shpsa.2015.10.001

played an essential role in our study of cognitive behaviours (for details, see: Barrett, 2006, p. 46). In all these cases, psychologists are not taking established neuroscientific theories as the basis for creating psychological models. Instead, their models are based on distinct types of theories which have as their foundation idealizations which do not apply in the context of neuroscience.

In a similar vein, neuroscientists cannot simply take the underlying theories of psychology and use them to create useful neuroscientific models, since psychological models only characterize how neural systems behave when they are organized into a complex cognitive system, and are engaging in particular environmental contexts. Moreover, these psychological theories are built upon idealizations that are unhelpful for characterizing the particular neurological structures and causes working within the system. Instead, neuroscientists develop their own sets of theories which adopt different sets of idealizations as their foundation.

For example, in order to study the specific contributions that neurological mechanisms make to cognitive behaviour, neuroscientists often need to hold environmental factors constant to allow for reliable and replicable manipulations of these mechanisms in order to learn about their functions. Since environmental factors can interfere with the functioning of the mechanism, by not holding the environment constant, it becomes impossible to tell whether aspects of the system's behaviours are due to structural or environmental factors. This means that an understanding of how the parts and operations of the system contribute to its overall behaviour require isolating it from such factors. This information is then used by neuroscientists to construct neuroscientific theories, and to form generalizations about the overall behaviour of neurological systems.

The problem is that this leaves us with neuroscientific theories that only provide an account of the behaviour of neurological mechanisms in isolation from other causal influences and organizational features that alter the behaviour of these mechanisms when embedded within larger cognitive systems (see: Bechtel, 2015; Datteri & Laudisa, 2012; Longino, 2006; 2013). As Datteri & Laudisa note, "these [neurological] generalizations are highly idealized, as they omit reference to the myriads of conditions that could perturb the behaviour of the modelled system in real-world settings" (2012, p. 602).[5]

Other kinds of neurological idealizations involve downplaying or ignoring other biological causes that contribute to cognitive behaviour in order to focus on the contributions that neural mechanisms make in particular. Helen Longino, for example, notes that behavioural genetics and neurobiology characterize the causes of phenomena like aggression and sexual orientation in incompatible ways. She notes that:

> Each approach employs methodologies that require particular ways of understanding the causal space. Some phenomena regarded as causally active in one approach are simply not included in another. These differential selections result in incongruous causal spaces. (2006, p. 118)

While these sorts of idealizations are foundational to many neuroscientific theories, it puts these theories into direct conflict with the sorts of scientific theories needed to characterize the behaviour of entire cognitive systems. If one is interested in studying how entire cognitive systems behave under varying environmental conditions, then it is unfruitful to use as our foundation scientific theories which idealize away many of the other underlying causes of such behaviours, and the environmental contexts in which they appear. Psychological theories in social psychology, community psychology, and developmental psychology, are intended to characterize the capacities and behaviours of entire cognitive systems under different conditions and in different environments. Thus the idealizations which are used to form neuroscientific theories must be abandoned in order to effectively construct and test psychological theories.

So far, I have provided a largely descriptive account of the practices of psychology and neuroscience in order to demonstrate the problems that stand in the way of any sort of convergence between the two domains into a single unified account. There is an important normative point that is worth making as well however. Namely, that even if convergence between psychology and neuroscience *were* possible, there are good reasons for thinking that such a convergence would actually be *undesirable* and *unhelpful* to the practices of both neuroscience and psychology.

We must not forget that the practice of science is performed by scientists with finite resources and time. Thus a single all-encompassing theory of cognitive behaviour may not prove useful or beneficial for scientists to actually apply given practical and pragmatic limitations. Recall that cognitive behaviour is not solely the product of neurological mechanisms, but also the product of many other biological and non-biological factors. Thus if we take the convergence position seriously, then many more scientific domains must be included in the convergence process. For a convergent account of cognitive behaviour to work, it will require a convergence of scientific domains such as psychology, neuroscience, evolutionary biology, genetics, developmental systems theory, economics, and sociology to name only a few (since each identifies a different dimension of causal influences responsible for the production of cognitive behaviour in various social and environmental contexts). Even if all these theories could converge on a single unified account of cognitive behaviour, the sheer volume of causal information that would need to be gathered to understand even simple systems would make this unified theory impractical and inapplicable in most real world contexts. We will almost never have all such information available to us in any one situation, even if we knew how to generate such a unified theory in principle. Moreover, time constraints may make it impossible to apply in practice even if we could gather this information, and limited funds may make it impractical as a scientific tool if it is particularly costly to generate.

Most importantly, the development of such an all-encompassing theory or model of the system is *not necessary* in order for scientists to carry out the representational tasks that they need accomplished. Thus instead of altering all our theories so as to converge on a single account, what is far more advantageous is to have a great number of idealized theories which will conflict with one another, but which will have different pragmatic value as representational tools in different contexts. Theories which are highly idealized in different ways will be simpler to develop, easier to apply, and more useful as scientific tools in the study of cognitive systems.

The particular sorts of idealizing and simplifying assumptions that psychology and neuroscience employ is what allows them to be pragmatically useful at representing the particular aspects of cognitive systems we use them to study while still working within

[5] It is worth noting that domains like ecological neuroscience do try to understand how neural mechanisms behave under different environmental conditions. But even in these cases, many environmental conditions must be held constant to see how particular environmental conditions will affect particular neural mechanisms. The behaviour of the system will change radically when all these environmental influences are in play, and causally interacting with the system simultaneously. Thus the theories of ecological neuroscientists still require forming idealized theories of the behaviour of these mechanisms given the methodological necessity of studying these systems under laboratory conditions that remove many environmental factors to study the influence of others.

Please cite this article in press as: Hochstein, E., Giving up on convergence and autonomy: Why the theories of psychology and neuroscience are codependent as well as irreconcilable, Studies in History and Philosophy of Science (2015), http://dx.doi.org/10.1016/j.shpsa.2015.10.001

the practical and pragmatic limitations we must adhere to. The different representational goals of psychology and neuroscience necessitate employing different types of theories and models which are pragmatically useful for their particular representational tasks, but not for the other.

This also explains why instead of witnessing psychology and neuroscience merging into a single unified theory, what we find instead is a further fragmentation of both psychology and neuroscience into numerous sub-fields, each of which focuses on studying a different restricted aspect of cognitive behaviour. Within psychology, we have the development of cognitive psychology, developmental psychology, clinical psychology, and social psychology to name only a few. Meanwhile in neuroscience we have the emergence of behavioural neuroscience, ecological neuroscience, computational neuroscience, and molecular neuroscience, among others. Each sub-domain focuses on representing different aspects of cognitive systems, and in doing so employs its own sets of models and theories which simplify and idealize systems in different ways essential for pragmatically satisfying the goals of those particular sub-domains. Some idealize environmental causes, some idealize neurological details, and some idealize other known biological or historical causes of behaviour.

All this gives us compelling reasons for denying that neuroscience and psychology are inevitably converging on a single unified account of cognitive behaviour. Instead, what we have are different domains that work with different representational goals and interests. In virtue of this, they embrace theories which cannot be amalgamated with those of their counterparts without losing what make them useful for the representational tasks to which they are put. It is their inability to integrate that makes the theories from the different domains useful for different representational tasks.

2.2. The problem with autonomy

Recall that the primary argument used to defend the autonomy of psychology from neuroscience depends on the notion of multiple-realizability. This argument suggests that psychological models identify computational states that can be instantiated in different ways by different systems, just so long as they have the appropriate functional organization. As such, psychological and neuroscientific models describe properties and states which are autonomous from one another.

While this argument has strong intuitive appeal, it runs into problems for engineering reasons. To better understand why, it is important to note that the way in which a computational system is physically structured always affects the sorts of functions it is optimized to perform. Thus, in order to get different systems that are implemented in different ways to carry out the same function, one must always do so at the cost of increased complexity, resources, and time (see: Eliasmith, 2002; Le Cun & Denker, 1992; Syropoulos, 2008). When dealing with Turing machine proofs, these sorts of issues can be safely ignored since we assume the system has infinite time and infinite resources. Systems in the real world, however, are always constrained by such limitations, and this will always translate into real world behavioural differences. As Syropoulos notes,

> …machine equivalence provides little information regarding the way a machine actually computes something, and it is this way that is cognitively relevant. For instance, although a modern CISC [Complex Instruction Set Computer] machine is equivalent a RISC [Reduced Instruction Set Computer] machine, in the sense that one can compile and execute exactly the same programs on both machines, a RISC machine is faster. […] Clearly, if we compile the same program under the same operating system running on two different architectures the resulting binary files will be completely different. Obviously both binaries will produce the same results, but one will be executed much faster than the other. The reason for this difference in performance is due to the simplicity of the RISC architecture or the complexity of the CISC architecture. (Syropoulos, 2008, p. 111)

Put simply, physical properties of the underlying mechanisms always affect the way in which given functions are carried out, and thus have different effects on the behavioural outputs than alternative implementations of the same functions. This means that the more behavioural and environmental constraints we can identify, the more it narrows the list of implementations that can generate those sort of functions under those sorts of conditions within those limitations. Ultimately, only a single physical implementation will be capable of producing exactly those behaviours, under those conditions, and in that space of time (for further discussion and elaboration, see: Bechtel, 2008; Bechtel & Mundale, 1999; Boone & Piccinini, 2015; Craver, 2007, 2009; Eliasmith, 2002, 2013; Keeley, 2000; Piccinini & Craver, 2011; Shapiro, 2004; Syropoulos, 2008).

Given this, knowing more about the behavioural outputs of the system necessarily informs us as to what the neurological mechanisms working within the system must be like. For a concrete example of this, consider SPAUN, a recent large-scale model of the functioning brain developed by Eliasmith et al. (2012) used to simulate various kinds of cognitive tasks. When discussing the implementation of the model, Eliasmith notes the following:

> Before leaving consideration of this model I want to highlight what I think is perhaps its most theoretically interesting feature —namely, that this model only works if it is implemented in neurons. […] If we directly simulate the equations that describe this model, then it is unable to accurately reproduce the recency and primacy effects observed in the human data. […] Consequently, we realized that one of the main reasons that this model is able to capture the human data as it does is that the individual neurons themselves saturate when participating in the representation of large vectors. This saturation serves as a kind of “soft” normalization, which is neither ideal mathematical nor a complete lack of normalization. Rather, it is more of a subtle kind of constraint placed on the representation of vectors in virtue of neuron response properties. And, crucially, this constraint is directly evident in the behavioral data (i.e., it enables reconstructing the correct U-shaped curve).
>
> This observation is theoretically interesting because it provides an unambiguous example of the importance of constructing a neural implementation for explaining high-level psychological behaviour. All too often researchers consider psychological and neural-level explanations to be independent. […] But in this case, the dependence is clear. Without constructing the neural model, we would have considered the mathematical characterization a failure and moved on to the other, likely more complex model. However, it is now obvious that we would have done so unnecessarily. (Eliasmith, 2013, pp. 218-219).

Here we can see how the regularities and features identified by psychological models are not, in fact, autonomous from those identified by neuroscientific ones. The more detailed our psychological generalizations about the system become, the more they tell us about what the underlying neurological architecture of the system must be like. Likewise, the more we know about the

Please cite this article in press as: Hochstein, E., Giving up on convergence and autonomy: Why the theories of psychology and neuroscience are codependent as well as irreconcilable, Studies in History and Philosophy of Science (2015), http://dx.doi.org/10.1016/j.shpsa.2015.10.001

neurological architecture, the more it constrains the sorts of psychological generalizations we can make about the system.

With this in mind, the assumption that neuroscience and psychology can proceed in complete isolation from one another is unjustified. The sorts of behaviours and capacities identified by psychological models put essential constraints on the construction and testing of our neuroscientific models. Likewise, knowing more about the neurological architecture of the system informs what sorts of behaviours are possible in different environmental contexts. This allows us to better refine the sorts of psychological generalizations we can make about the system. As such, arguments for a strict autonomy of psychology from neuroscience are undermined. Instead, neuroscientific models and psychological models depend on one another for better understanding and refinement.

It is important to understand how this insight connects to the conclusion of Section 2.1. Given that cognitive behaviour is not produced solely by neurological mechanisms, we cannot use neuroscientific theories alone to accurately characterize many of the cognitive behaviours displayed by entire cognitive systems that psychological theories are used to represent. However, given that the system is *partially* constituted by these mechanisms, the way in which these mechanisms are implemented *does* place direct constrains on the possible cognitive behaviours that the entire system is capable of engaging in under different conditions. This information regarding the implementation of the neural mechanisms allows us to refine and improve our psychological models in light of the fact that these mechanisms are *part of* the cognitive system. However, this does not mean that the two domains will converge on a single unified theory given that the theories and models of the different domains still must idealize cognitive systems in incompatible ways to effectively represent different aspects of the system.

3. An irreconcilable codependence

The relationship between neuroscience and psychology now appears to be neither one of convergence, nor one of autonomy. On the one hand, in order for the theories of psychology to effectively represent the behavioural capacities and patterns of entire cognitive systems, they must take as their foundation sets of idealizing assumptions that put them at odds with the sorts of idealizations that neuroscientific theories must adopt. In this sense, to force neuroscientific theories and psychological theories to converge and unify would be to strip them of the very conflicting idealizations that make them representationally useful for the different purposes to which they are put. On the other hand, if we conclude that neuroscience and psychology should therefore operate in isolation from one another, then we ignore the essential constraints that the two domains place on one another. In this respect, the theories from one domain are needed to help construct and test theories in the other domain.

We now appear to be in a position where the more we try to bring psychology and neuroscience together, the worse both become at achieving their different representational goals, but similarly the more we pull them apart, the worse both become at being able to refine and improve their models and theories. We are therefore left with an irreconcilable codependence between the two domains; a constant and unavoidable back and forth between domains that allows each domain to improve and provide better theories without moving towards unification or convergence.

More importantly, I propose that this inability for the two domains to converge or stand apart from each other is not a problem that must be overcome, but is in fact *the very thing that allows both domains to progress and improve*. The fact that the theories and models from the two domains are irreconcilable is what allows them to generate different kinds of information about the target system by representing them in distinct ways. The idealizations employed by the different domains is precisely what allows scientists to form testable theories and gather distinct information about different facets of the cognitive system that the other domain needs in order to refine its own theories, but cannot gather using the limitations of their own idealized theories and models. In this respect, attempts to pull psychology and neuroscience apart from one another, or to merge them together into a single account, are both inappropriate for the same reason: they ignore what makes each domain informative to the other. It is the dynamic interaction between the two domains that allows them to be useful for distinct representational goals, but also to learn and improve from each other.

When it comes to the arguments for autonomy and convergence, both get part of the story right, but draw the wrong conclusions from it. Those who argue for autonomy are correct that we can generate psychological generalizations about cognitive systems autonomously from studying their neurological implementation. The models from psychology heavily idealize many of the mechanistic details of the system so as to form more elegant and easily applicable theories which can be used to identify general behavioural patterns and capacities without having to gather all the underlying causal information about the system. Likewise, neuroscientists often study the behaviour of neurological mechanisms in isolation of the other mechanisms and environmental factors which influence cognitive behaviour.

But while this gives the appearance of autonomy between the two domains, it would be wrong to conclude from this that the way in which the neurological mechanisms of the system are implemented does not matter to, or inform our understanding of, the overall cognitive capabilities and regularities of the system. Knowing these details allows us to better determine what the system can do in different situations, and thus to refine our psychological models. Likewise, the more we know about the overall behaviours of the system gained through the application of our psychological models, the more it informs what the neural mechanisms must be like. Thus autonomy between neuroscience and psychology would be the wrong conclusion to draw.

Meanwhile, those who advocate for convergence make the opposite mistake. They rightly note that the models of psychology and the models of neuroscience directly inform one another, and allow us to refine and improve our different accounts. Their mistake is to conclude from this that the two domains must therefore be moving towards *integration*. Yet this does not follow either. Improvements to both domains do not lead towards integration or isolation, but merely to the fact that different theories and models get better at the particular representational tasks we use them for. Refining and improving a psychological model based on neurological data, for instance, does not require that we thereby turn the psychological model into a neurological one (since this would defeat the purpose of using it for a distinct representational purpose), only that we use the neurological data to improve the psychological model's ability to represent the behaviours of the entire cognitive system by including new boundary conditions, more detailed generalizations, and recognizing previously unknown constraints.

Before bringing things to a close, it is worth noting that there have been other theorists who have similarly defended a view that is neither one of converge, nor autonomy. William Bechtel and Lindley Darden, for instance, both advocate views that differ from the traditional accounts. While in many ways the position I defend

Please cite this article in press as: Hochstein, E., Giving up on convergence and autonomy: Why the theories of psychology and neuroscience are codependent as well as irreconcilable, Studies in History and Philosophy of Science (2015), http://dx.doi.org/10.1016/j.shpsa.2015.10.001

builds off the ideas of both Bechtel and Darden, their accounts differ in important respects from the one defended here.[6]

Darden, for instance, defends a view of interfield integration (Darden, 1986, 1993; Darden & Craver, 2002; Darden & Maull, 1977), whereby the theories from different domains do not converge on a single unifying theory, but instead are all integrated into a larger unifying framework. In these cases, the theories from the different domains remain distinct, but are used to characterize different compositional levels of the same multi-level mechanistic system. As Darden puts it:

> In contrast, in interfield analyses no derivation is postulated and no elimination occurs; instead, bridges are built between two different bodies of knowledge. The focus is on the bridging relations, which sometimes constitute an actual theory, such as the chromosome theory of Mendelian heredity. (Thus, what for the reductionist are reduction functions constructed by the philosopher are, for the interfield case, the actual bridging theory constructed by the scientist.) In cases in which the separate bodies of knowledge can be ordered hierarchically, the bridging theories are interlevel theories. The bridges serve to unify, but not to eliminate, fields. (1993, p.143)

The problem with Darden's account is the assumption that the theories from the different domains will seamlessly fit together (once the appropriate bridging theories, or hierarchical relations, are discovered) in the context of the unifying framework. In contrast, I have proposed that the very reason why the theories from the different domains can be informative to one another is precisely because they *cannot* be integrated within the same framework in this way. It is their very *inability* to integrate together that allows the two domains to relevantly inform one another, as well as to improve and grow.

A similar account of interfield integration has also been defended by Bechtel (1984, 1986). In this respect, Bechtel's account runs into the same problems that Darden's account faces. There have been some instances, however, where Bechtel offers a somewhat different story. In some cases, Bechtel suggests that what differentiates psychological theories from neuroscientific theories is in fact the grain of abstraction adopted in the description of the neurological mechanisms under investigation (it is unclear exactly how this fits into his integrative framework however). He claims that psychological descriptions are coarse-grained descriptions of the same neurological mechanisms that neuroscientific theories describe in fine-grained detail (see: Bechtel, 2008, p.138-142; Bechtel & Mundale, 1999). More specifically:

> Thus, one diagnosis of what has made the multiple realizability claim as plausible as it is has been is that researchers have employed different grains of analysis in identifying psychological states and brain states, using a coarse grain to identify psychological states and a fine grain to differentiate brain states (Mundale & Bechtel, 1999, p. 202)

This idea is problematic in two important ways. First, I have demonstrated that psychological descriptions should not be interpreted merely as coarse grained descriptions of neurological mechanisms. They describe behaviours caused by influences beyond the neural. Second, I have argued that what differentiates psychological theories from neuroscientific theories is not simply the grain of abstraction employed, but more importantly the sorts of *idealizations* employed by the different domains.

Darden and Bechtel both overlook the fact that the very idealizations which stand in the way of the integrative projects they espouse plays not only an essential role in differentiating psychological theories from neuroscientific ones, but also in being the driving force that allows the two domains to learn from each other and improve over time. In this sense, the project of integration proposed by Bechtel and Darden run into the same problems that plagued the convergence account.

4. Conclusion

The relationship between neuroscience and psychology has been traditionally conceived as either moving towards unification, or moving towards autonomy. In this paper, I have argued that neither is correct. Instead, progress in both domains is dependent on the fact that they represent systems in ways that make integration impossible, but likewise do not work in isolation from one another. There is a back and forth between the two domains that allows both to be improved and refined over time despite never resulting in a unified account, or working autonomously from one another. Instead, there is an irreconcilable codependence between the two fields.

References

Aizawa, K., & Gillett, C. (2011). The autonomy of psychology in the age of neuroscience. In P. M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences*. Oxford: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199574131.003.0010.

Ajzen, I. (1985). From intention to actions: A theory of planned behavior. In J. Kuhi, & J. Beckmann (Eds.), *Action-control: From cognition to behaviour* (pp. 11-39). Heidelberg: Springer.

Ajzen, I. (1988). *Attitudes, personality and behavior*. Milton Keynes: Open University Press.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179-211. http://dx.doi.org/10.1016/0749-5978(91)90020-T.

Armitage, C., & Connor, M. (2001). Efficacy of the theory of planned behaviour: A meta-analytic review. *British Journal of Social Psychology, 40*, 471-499. http://dx.doi.org/10.1348/014466601164939.

Barrett, L. (2006). Are emotions natural kinds? *Perspectives on Psychological Science, 1*(1), 28-58. http://dx.doi.org/10.1111/j.1745-6916.2006.00003.x.

Bechtel, W. (1984). Reconceptualizations and interfield connections: The discovery of the link between vitamins and coenzymes. *Philosophy of Science, 51*, 265-292. http://dx.doi.org/10.1086/289180.

Bechtel, W. (1986). Introduction: The nature of scientific integration. In W. Bechtel (Ed.), *Integrating scientific disciplines* (pp. 3-52). Dordrecht: Nijhoff. http://dx.doi.org/10.1007/978-94-010-9435-1_1.

Bechtel, W. (2007). Reducing psychology while maintaining its autonomy via mechanistic explanations. In M. Schouten, & H. L. De Joong (Eds.), *The matter of the mind: Philosophical essays on psychology, neuroscience and reduction*. Blackwell Publishing.

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. New York: Lawrence Erlbaum Associates.

Bechtel, W. (2015). Can mechanistic explanation be reconciled with scale-free constitution and dynamics? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 53*, 84-93. http://dx.doi.org/10.1016/j.shpsc.2015.03.006.

Bechtel, W., & Mundale, J. (1999). Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science, 66*(2), 175-207. http://dx.doi.org/10.1086/392683.

Bickle, J. (1998). *Psychoneural reduction: The new wave*. Cambridge, MA: MIT Press.

Bickle, J. (2003). *Philosophy and neuroscience: A ruthlessly reductive account*. Dordrecht, NL: Kluwer.

Bickle, J. (2006). Reducing mind to molecular pathways: Explicating the reductionism implicit in current cellular and molecular neuroscience. *Synthese, 151*, 411-434. http://dx.doi.org/10.1007/s11229-006-9015-2.

[6] Given the prolific nature of both Bechtel and Darden's work, one can find passages in their writings that both support, and argue against, points made in this paper. As such, I will focus on the sorts of commitments that Bechtel and Darden have defended which run counter to the view defended here in order to highlight where such commitments go wrong. Likewise, given the sheer volume of literature on the relationship between neuroscience and psychology, there are likely many more theorists who can be mentioned here in addition to Bechtel and Darden. Many of these theorists propose views very similar in kind to the sorts proposed by Darden and Bechtel, and so using them as paradigm examples should hopefully demonstrate how the account on offer here differs from many of these alternatives as well.

Please cite this article in press as: Hochstein, E., Giving up on convergence and autonomy: Why the theories of psychology and neuroscience are codependent as well as irreconcilable, Studies in History and Philosophy of Science (2015), http://dx.doi.org/10.1016/j.shpsa.2015.10.001

Block, N. (1980). *Readings in philosophy of psychology*. Cambridge, MA: Harvard University Press.

Blue, C. L. (1995). The predictive capacity of the theory of reasoned action and the theory of planned behavior in exercise research: An integrated literature review. *Research in Nursing and Health, 18*, 105-121. http://dx.doi.org/10.1002/nur.4770180205.

Boone, W., & Piccinini, G. (2015). The cognitive neuroscience revolution. *Synthese*, 1-26. http://dx.doi.org/10.1007/s11229-015-0783-4.

Bratman, M. (1987). *Intention, plans, and practical reason*. Harvard University Press.

Burge, T. (2010). *Origins of objectivity*. Oxford: Oxford University Press.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy, 78*, 67-90. http://dx.doi.org/10.2307/2025900.

Churchland, P. S. (1989). *Neurophilosophy: Toward a unified science of the mind-brain*. Cambridge, Massachusetts: MIT Press.

Connor, M., & Sparks, P. (1996). The theory of planned behaviour and health behaviours. In M. Conner, & P. Norman (Eds.), *Predicting health behaviour* (pp. 121-162). Buckingham: Open University Press.

Crane, T. (2001). *Elements of mind*. Oxford: Oxford University Press.

Craver, C. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford: Oxford University Press.

Craver, C. (2009). Mechanisms and natural kinds. *Philosophical Psychology, 22*(5), 575-594. http://dx.doi.org/10.1080/09515080903238930.

Cummins, R. (1983). *The nature of psychological explanation*. Cambridge, MA: MIT Press.

Darden, L. (1986). Relations among fields in the evolutionary synthesis. In W. Bechtel (Ed.), *Integrating scientific disciplines* (pp. 113-123). Dordrecht: Nijhofl.

Darden, L. (1993). Interfield theories and strategies for theory change. In H. V. Rappard, P. J. Van Strien, L. P. Mos, & W. J. Baker (Eds.), *Annals of theoretical psychology* (Vol. 9, pp. 141-144)New York: Plenum Press. http://dx.doi.org/10.1007/978-1-4615-2986-6_8

Darden, L., & Craver, C. (2002). Strategies in the interfield discovery of the mechanism of protein synthesis. *Studies in History and Philosophy of Biological and Biomedical Sciences, 33*, 1-28. http://dx.doi.org/10.1016/S1369-8486(01)00021-8.

Darden, L., & Maull, N. (1977). Interfield theories. *Philosophy of Science, 44*, 43-64. http://dx.doi.org/10.1086/288723.

Datteri, E., & Laudisa, F. (2012). Model testing, prediction, and experimental protocols in neuroscience: A case study. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 43*(3), 602-610. http://dx.doi.org/10.1016/j.shpsc.2012.04.001.

Davidson, D. (1974). Psychology as philosophy. In S. Brown (Ed.), *Philosophy of psychology* (pp. 41-52). London: Macmillan Press.

Dennett, D. (1987). *The intentional stance*. Cambridge, Massachusetts: The MIT Press.

Eliasmith, C. (2002). The myth of the turing machine: The failing of functionalism and related theses. *Journal of Experimental & Theoretical Artificial Intelligence, 14*(1), 1-8. http://dx.doi.org/10.1080/09528130210153514.

Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. Oxford University Press.

Eliasmith, C., Stewart, T., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science, 338*(6111), 1202-1205. http://dx.doi.org/10.1126/science.1225266.

Fishbein, M., & Ajzen, I. (1981). Attitudes and voting behavior: An application of the theory of reasoned action. In G. M. Stephenson, & J. M. Davis (Eds.), *Progress in applied social psychology* (Vol. I, pp. 253-313)London: Wiley.

Fodor, J. (1974). Special sciences and the disunity of science as a working hypothesis. *Synthese, 28*, 77-115. http://dx.doi.org/10.1007/BF00485230.

Fodor, J. (1998). *Concepts*. Oxford: Clarendon Press.

Føllesdal, D. (1982). The status of rationality assumptions in interpretation and in the explanation of action. *Dialectica, 36*(4), 301-316. http://dx.doi.org/10.1111/j.1746-8361.1982.tb01545.x.

Georgeff, M., Pell, B., Pollack, M., Tambe, M., & Woodridge, M. (2003). The belief-desire-intention model of agency. In J. Muller, M. Singh, & S. Rao (Eds.), *Intelligent Agents V: Agents theories, architectures, and languages lecture notes in computer science*. Springer. http://dx.doi.org/10.1007/3-540-49057-4_1.

Griffith, P. (1997). *What emotions really are: The problem of psychological categories*. Chicago: University of Chicago Press.

Griffith, P. (2004). Emotions as natural and normative kinds. *Philosophy of Science, 71*(5), 901-911. http://dx.doi.org/10.1086/425944.

Hausenblas, H. A., Carron, A. V., & Mack, D. E. (1997). Application of the theories of reasoned action and planned behavior to exercise behavior: A meta-analysis. *Journal of Sport and Exercise Psychology, 19*, 36-51.

Hochstein, E. (2012). Minds, models, and mechanisms: A new perspective on intentional psychology. *Journal of Experimental & Theoretical Artificial Intelligence, 24*(4), 547-557. http://dx.doi.org/10.1080/0952813X.2012.693688.

Hochstein, E. (2013). Intentional models as essential scientific tools. *International Studies in the Philosophy of Science, 27*(2), 199-217. http://dx.doi.org/10.1080/02698595.2013.813251.

Hochstein, E. (2015). One mechanism, many models: A distributed theory of mechanistic explanation. *Synthese*, 1-21. http://dx.doi.org/10.1007/s11229-015-0844-8.

Hooker, C. (1981). Towards a general theory of reduction. Part I: Historical and scientific setting. Part II: Identity in reduction. Part III: Cross-categorial reduction. *Dialogue, 20*, 38-59, 201-236, 496-529 http://dx.doi.org/10.1017/S0012217300023593

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness*. New York: Cambridge University Press.

Keeley, B. (2000). Shocking lessons from electric fish: The theory and practice of multiple realizability. *Philosophy of Science, 67*, 444-465. http://dx.doi.org/10.1086/392790.

Kellert, H., Longino, H., & Waters, C. K. (2006). Introduction: The pluralist stance. In S. Kellert, Longino, & Waters (Eds.), *Scientific pluralism* (pp. vii-xxix). Minneapolis: University of Minnesota Press.

Le Cun, Y., & Denker, J. S. (1992). Natural versus universal probability, complexity, and entropy. In *IEEE workshop on the physics of computation*.

Longino, H. (2006). Theoretical pluralism and the scientific study of behavior. In S. Kellert, H. Longino, & C. K. Waters (Eds.), *Scientific pluralism* (pp. 102-132). Minneapolis: University of Minnesota Press.

Longino, H. (2013). *Studying human behaviour*. Chicago: The University of Chicago Press.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1-25. http://dx.doi.org/10.1086/392759.

Menzies, P., & List, C. (2010). The causal autonomy of the special sciences. In C. MacDonald, & H. MacDonald (Eds.), *Emergence in mind*. Oxford: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199583621.003.0008.

Parker, W. S. (2006). Understanding pluralism in climate modeling. *Foundations of Science, 11*, 349-368. http://dx.doi.org/10.1007/s10699-005-3196-x.

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese, 183*(3), 283-311. http://dx.doi.org/10.1007/s11229-011-9898-4.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: The MIT Press.

Sehon, S. (1997). Natural-kind terms and the status of folk psychology. *American Philosophical Quaterly, 34*, 333-344.

Shapiro, L. A. (2004). *The mind incarnate*. Cambridge, MA: MIT Press.

Stich, S. (1983). *From folk psychology to cognitive science: The case against belief*. Cambridge, MA: The MIT Press.

Syropoulos, A. (2008). *Hypercomputation: Computing beyond the Church-Turing Barrier*. Springer.

Weisberg, M. (2013). *Simulation and similarity: Using models to understand the world*. Oxford: Oxford University Press.

Wimsatt, W. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations of reality*. Cambridge: Harvard University Press.

Please cite this article in press as: Hochstein, E., Giving up on convergence and autonomy: Why the theories of psychology and neuroscience are codependent as well as irreconcilable, Studies in History and Philosophy of Science (2015), http://dx.doi.org/10.1016/j.shpsa.2015.10.001