Australian Journal of Psychology

Australian Journal of Psychology 2012; **64**: 52–59
doi:10.1111/j.1742-9536.2011.00044.x

APS Australian Psychological Society

Levels of explanation and the workings of science

Amy Perfors

School of Psychology, University of Adelaide, Adelaide, South Australia, Australia

Abstract

I address two questions that underlie most of the articles in this special issue: 1) What do different levels of explanation in psychology reveal? And 2) how do the dynamics of science affect what can be learned? I suggest that understanding hypothesis testing and generation in the abstract can provide a useful framework for understanding how cognitive modelling and neuroscience may interact. I further suggest that the preference for simple explanations and the dynamics of hypothesis testing may play out in different ways within the two fields, and that their overlap may prove most useful in the realm of hypothesis generation.

Key words: cognitive science and intelligent systems, neuroscience, psychology as a discipline

INTRODUCTION

What can neuroscience tell us about questions in cognitive science? How do—and how *should*—neuroscience and cognitive modelling interact with each other? These questions, which are the focus of this special issue, can be distilled to more precise queries. First, what do *particular* studies, findings, or approaches in neuroscience have to say about topics in cognitive science? Second, what does (or can) neuroscience *as a field* say to cognitive science? Third, more broadly, what is the relationship between explanations that occur at different levels or that use different methodologies? What can they say to each other, and in what situations? Finally, for all of these questions, how are they affected by the dynamics of the process of scientific inquiry?

Most articles in this issue focus, as they should, on the first two questions—on specific experiments or situations in which neuroscience and cognitive modelling give explanations that are either contrasting (Kalish & Dunn, 2012; Lewandowsky, Ecker, Farrell, & Brown, 2012) or complementary (Brown, Forstmann, & Wagenmakers, submitted; de Zubicaray, 2012). I will be focusing mainly on the third and most abstract question, though I will also touch on the second when warranted. My goal with this article is to offer a commentary about the forest rather than the trees, and to present some (hopefully thought-provoking) ideas, not to communicate polished conclusions. As such, there is a lot of speculation and few ironclad results.

The plan for this article is as follows. I begin by sketching a basic framework for understanding the process of scientific inquiry as a process of hypothesis generation and testing. This framework forms the backdrop for consideration of three questions that arise, whether explicitly or implicitly, upon consideration of the relationship between neuroscience and cognitive science. The first two centre on hypothesis testing. First, how should a rational reasoner evaluate theories, and how does this relate to how people *actually* evaluate them? What sort of explanations feel 'good' to us, and why? I posit that people's differential willingness to accept cognitive and neuroscientific explanations may have to do with our intuitive notions of simplicity and elegance. Second, what kinds of tests are the most effective and efficient ways to identify correct theories or hypotheses? I present preliminary research that suggests that the answer may depend on the 'newness' of the field or question being asked. Since cognitive modelling and neuroscience differ in this respect, it may be one reason for apparent differences between the ways each field operates. Finally, I will consider hypothesis generation, arguably one of the most difficult aspects of science. I suggest that, even if two fields to scientific inquiry are otherwise incommensurate (which may or may not be the case with cognitive science and neuroscience), their interaction can still be beneficial, because it may assist in hypothesis generation.

A FRAMEWORK FOR SCIENTIFIC INQUIRY

The goal of any scientific pursuit is to identify—out of all possible hypotheses or theories purporting to explain the world—which hypothesis is most correct. We can conceptualise this pursuit with a slightly bizarre but illuminating

Correspondence: Amy Perfors, PhD, School of Psychology, University of Adelaide, Level 4, Hughes Building, Adelaide, SA 5005, Australia. Email: amy.perfors@adelaide.edu.au

Received 9 April 2011. Accepted for publication 20 September 2011.

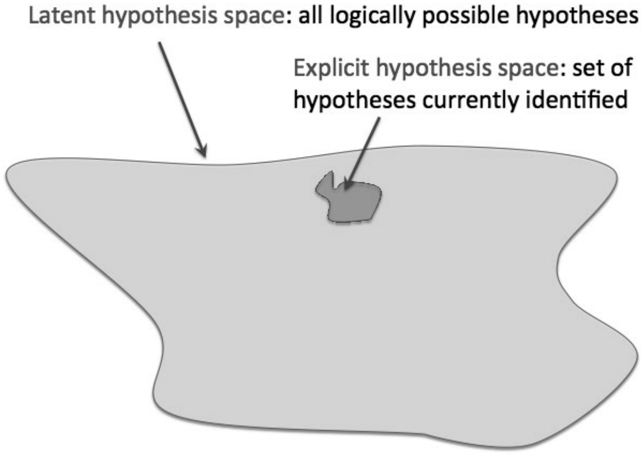
Latent hypothesis space: all logically possible hypotheses
Explicit hypothesis space: set of hypotheses currently identified

Figure 1 Schematic illustration of latent and explicit hypothesis spaces. A latent hypothesis space consists of the set of all logically possible hypotheses that could be explanations for a given question. The explicit hypothesis space is the set of explicitly identified hypotheses, which are the ones available for hypothesis testing.

example involving number rules, loosely inspired by similar tasks like the 'number game' in Tenenbaum and Griffiths (2001) or the 2–4–6 task of Wason (1960). Imagine that we are radio astronomers who start receiving a series of signals from a distant galaxy. Every 20 hours, we receive a new signal consisting of a single number. After 60 hours, we have received a 10, a 50, and a 30. Our goal as scientists is to determine what generative process has resulted in this series of numbers: Is there an underlying rule of some sort? Are the numbers completely random? Are they being sent by aliens bent on world domination, or are they the result of some natural phenomenon, like a pulsar?

The process of answering these questions involves two components: hypothesis testing (evaluating our hypotheses about the answers to these questions against each other) as well as hypothesis generation (identifying hypotheses to test in the first place). We can call the set of hypotheses the *hypothesis space*, denoted $\mathcal{H}$ and shown schematically in Fig. 1. Since most of the time we cannot explicitly list (nor do we even know) all of the hypotheses we could *possibly* test, it is important to make a conceptual distinction between two types of hypothesis spaces. The *latent hypothesis space* corresponds to all logically possible hypotheses—in this case, all possible number rules, from the trivial (ALL NUMBERS, EVEN NUMBERS) to the totally bizarre (NUMBERS CONTAINING AT LEAST ONE DIGIT TOPOLOGICALLY IDENTICAL TO A DONUT). This differs from the set of hypotheses we have explicitly identified for testing, which we might call the *explicit hypothesis space*. In our example, it includes the hypotheses we the scientists have identified to test (like EVEN NUMBERS) but not the ones that are so bizarre they never occur to us (like the one about digits with certain topologies).[1]

Under this conceptualisation, hypothesis generation consists of moving hypotheses from the latent to the explicit hypothesis space. This is a hard problem about which relatively little is known with much certainty. Much more is known (or at least theorised) about hypothesis testing; indeed, it is common within the philosophy of science to frame the process of hypothesis testing as a type of Bayesian reasoning (e.g., Howson, 2001; Jaynes, 2003). Under this view, the probability of some hypothesis h_i is a function of its prior probability (denoted $P(h_i)$), and the probability that one would see the observed data if the hypothesis were true (denoted $P(d|h_i)$). These terms combine to form Bayes' Rule:

$$P(h_i \mid d) = \frac{P(d \mid h_i)P(h_i)}{\sum_{h_j \in \mathcal{H}} P(d \mid h_j)P(h_j)} \qquad (1)$$

Altogether, this simple Bayesian framework provides a useful tool for formalising and conceptualising the questions that arise in this special issue, as we will see in the next sections.[2] A Bayesian perspective is advantageous both because it provides a rigorous and clear explanatory framework in which to conceptualise otherwise somewhat fuzzy issues, but also because Bayesian reasoning is 'optimal' in a specific mathematical sense (see Perfors, Tenenbaum, Griffiths, & Xu, 2011 for further discussion of this issue). However, it does have its shortcomings, most notably that the calculations underlying Bayes' Rule are only possible over a set of explicitly enumerated hypotheses. This has two implications of interest here. First, it does not address one of the most difficult problems in science, which is defining the relevant hypothesis spaces and/or reasoning within ill-defined hypothesis spaces. We can still talk about these problems in the abstract, which I do towards the end of the article, but it does put a severe limitation on our ability to apply formal solutions to them. Second, although data can rule out hypotheses that are not being explicitly considered (presuming the learner can remember that data once the hypotheses move into explicit consideration), the probability of different hypotheses can be calculated using Bayes' Rule only relative to the others under consideration. This is an important issue in hypothesis testing, which I consider next in light of the issue of what makes a good explanation.

FAVOURING SIMPLER EXPLANATIONS

Consider the three numbers we radio astronomers have received so far: 10, 50, and 30. These numbers are consistent with several hypotheses, including MULTIPLES OF 10 and MULTIPLES OF 5, among others. Since both hypotheses can explain the same data, why might we—or any Bayesian reasoner—prefer one of these hypotheses over another? There is some evidence that simplicity considerations play a role: all other things being equal, people tend to prefer simpler hypotheses (Lombrozo, 2007).

But what does 'simpler' mean? In practice, there are two kinds of simplicity: making more precise predictions, or having fewer parameters. Both of these matters to people, and both emerge naturally from a Bayesian reasoning framework (though it is an open question the degree to which people's assumptions map directly onto Bayesian reasoning). I'll talk here about both kinds of simplicity, starting with the former. The goal is to explain why it is sensible to evaluate hypotheses based on this sense of simplicity. Then I'll discuss how this might relate to evaluating theories in neuroscience and cognitive science.

Simplicity as precision

One idea is that simpler hypotheses are those that make more precise predictions. 'Precision' here should be taken to mean how closely a prediction matches an outcome. If, as in the numbers example, we assume that all allowable outcomes are equally likely, then precision is inversely proportional to the number of outcomes the theory predicts. This implies that a hypothesis consistent with a small number of possible data points is more precise than one that is consistent with many. For instance, MULTIPLES OF 10 is more precise than MULTIPLES OF 5: there are half as many numbers consistent with MULTIPLES OF 10 than MULTIPLES OF 5. If we know that numbers are constrained to be within a certain range, say from 0 to 100, then we can calculate this difference in precision: the probability of randomly drawing, say, a 50 from the rule MULTIPLES OF 10 is 1/10 (because there are 10 multiples of 10 between 0 and 100), but the probability of drawing it from MULTIPLES OF 5 is only 1/20.[3] In this case, the explicit probabilities support the intuition most people have: MULTIPLES OF 10 seems like a better explanation of 10, 50, and 30 than does MULTIPLES OF 5.

This notion of precision[4] is discussed in some detail by Roberts and Pashler (2000), although not by name. They make the distinction between goodness of fit (which measures the extent to which a theory captures the observed data) and the likelihood of other outcomes under the theory. Their sense of goodness of fit is not precision: for example, both MULTIPLES OF 10 and MULTIPLES OF 5 fit all of the data points. Rather, precision is analogous to the likelihood of other outcomes under the theory—or, equivalently, the extent to which the theory constrains the data one should expect. A theory like MULTIPLES OF 10 is more constraining and therefore more precise than MULTIPLES OF 5; for this reason, Roberts and Pashler (2000) argue it should be preferred over one that happens to fit the data, but only because it can fit *any* data.

A preference for more precise hypotheses is built into the likelihood and emerges naturally from the laws of probability (Jaynes, 2003; Jeffreys, 1939; MacKay, 2003). Sometimes called the 'Bayesian Ockham's Razor', it reflects the fact that probabilities must sum to one. Thus, if there are more data points to divide the total probability mass over, any individual data point will have less probability for itself. This provides a formal justification for why theories that are consistent with any (or many) outcomes should be dispreferred: if a theory is consistent with many things, then being consistent with any *specific* outcome is not very strong evidence for that theory.

What does this have to do with neuroscience? Several of the other articles in this special issue (e.g., Kalish & Dunn, 2012; Lewandowsky et al., 2012) identify instances where the neuroscientific theory is difficult to falsify, because it can be made consistent with any observed pattern of data. This is indeed a problem with the theory in question, but it is not a problem (as far as I can tell) specific to neuroscience: it is a mark of poor science in general. If neuroscientific hypotheses *by their nature* are more difficult to make precise, then these case studies would indeed be an indictment of the entire field—or at least a suggestion that neuroscientists should take more care when formulating their theories. If they are not, then it is an indication that it is possible that poorly specified, imprecise theories are bad science wherever they are found.

Simplicity as fewer parameters

There is, of course, another notion of simplicity, as a moment's thought will make clear. If hypotheses were preferred only on the basis of their precision, then we should always favour the hypothesis that included *only* the observed data; in the example above, we should think that the correct hypothesis is just THE NUMBERS 10, 30, AND 50. Most people do not do this, probably because hypotheses like this have low prior probability—we don't automatically and intuitively think that this hypothesis is *a priori* reasonable. In Bayesian terms, prior probability reflects, at least in part, the ease of describing it in the representation language used: hypotheses that have longer descriptions, or require more 'choices' in the description language, have lower prior probability.[5] In the number rule example, both MULTIPLES OF 10 and MULTIPLES OF 5 might have higher probability, because the concept of 'multiple' is an elementary concept in most people's representation of number. By contrast, the hypothesis THE NUMBERS 10, 30, AND 50 is effectively a concatenation of *three* elementary concepts (the three specific numbers involved), and is thus more complex. Although this can be captured within the Bayesian framework (Perfors et al., 2011), the important thing for our purposes is simply to note that both kinds of simplicity exist and have a formal justification.

Although there are some subtleties, it is relatively straightforward to incorporate this sort of intuitive simplicity into a prior probability in a context like the number rules example. However, in the real world—where real humans have to make judgments about much more complex hypotheses—doing so is a highly non-trivial problem. It becomes especially problematic if people must evaluate hypotheses in a field in which they are not an expert. For instance, Weisberg, Keil, Goodstein, Rawson, and Gray (2008) found that non-experts tended to confuse bad explanations for good ones when the explanations made reference to neuroscience but that experts were not similarly led astray. These results have many possible explanations, but one is that experts are well calibrated about how simple neuroscientific explanations really are (which is to say, generally not very simple). This stands in contrast to folk psychological theories of the mind, which are often of the form 'X part of the brain is responsible for Y behaviour'. Given this folk theory, an 'explanation' consisting of a picture of a brain with one part lit up, or a naming of the location where activity was found, does in fact feel satisfying, even though it *actually* explains very little.

If it is indeed the case that part of the allure of neuroscience relates to how our folk theories evaluate the simplicity of neuroscientific hypotheses, then it does suggest a distinction between neuroscience and other scientific fields—not in how it is done but in how it is explained to and understood by non-expert lay people. It also suggests an interesting test case for comparison: mathematical models of cognition. My intuition is that lay people have the opposite intuitions about math than they do about pictures of brains: Adding math to an explanation feels like over complicating it or perhaps even obscuring tricks done by the experimenters to make their results look good. Thus, I would predict that non-experts would show the opposite pattern of responses they do in Weisberg et al. (2008): they should interpret even *good* explanations as bad ones if mathematics is added. To my knowledge, nobody has done this study, but it would be revealing about the underlying mechanism(s).

The larger point, though, is the suggestion that people may be relying on their intuitive notions of simplicity when evaluating scientific theories. In some cases, like when we favour hypotheses that are more precise over those that are more unfalsifiable, there is no in-principle difference between neuroscience and other fields. In other cases, our preference for simplicity might play out differently in neuroscience and other fields because we—especially non-experts—have different intuitions about what makes an explanation simple.

POSITIVE AND NEGATIVE TESTS

A common theme in this special issue is the notion of falsification—the extent to which theories (in particular, neuroscientific theories) can be eliminated from consideration. This issue has been widely discussed in both philosophy of science and psychology, along with a closely related issue: what kinds of tests are most *effective* at falsifying theories. Positive tests evaluate a hypothesis by investigating whether the events that it predicts are eventually observed. In the number rules example, let's imagine that an alien arrives and offers to answer queries about what numbers might be received. A positive test of the hypothesis MULTIPLES OF 10 would be to ask the alien whether numbers like 20 or 80 could be observed. If so, this is a partial confirmation of the hypothesis; if they are not, this is a falsification. Conversely, a negative test of the same hypothesis would be to ask whether numbers that *aren't* multiples of 10, like 23 or 48, could be observed. If they are, this is a falsification of the hypothesis; if not, it is a partial confirmation.

It has been known for a few decades that positive tests are more efficient than negative tests if the goal is to eliminate hypotheses, at least when most hypotheses in the space are sparse (e.g., Austerweil & Griffiths, 2008; Klayman & Ha, 1987; Navarro & Perfors, 2011). Sparse hypotheses are those in which fewer data points are consistent than inconsistent with the hypothesis. Thus, more precise hypotheses are sparser.[6] For instance, a hypothesis like MULTIPLES OF 50 is sparser than MULTIPLES OF 10, which is sparser than EVEN NUMBERS (which is not sparse at all).

The intuitive reason that positive tests are more efficient for sparse hypotheses relates to the fact that fewer data points are consistent with those hypotheses. As a result, any single data point can rule out many of them. Conversely, if most hypotheses are non-sparse, then a single *negative* test will rule out more hypotheses for a similar reason. This is illustrated in Fig. 2.

Since we tend to favour hypotheses that are precise (i.e., sparse), this would seem to imply that positive tests will always be more effective in science as well. Is it indeed always better to evaluate a theory by testing its positive predictions rather than investigating whether the events it predicts will not occur actually do not? Not necessarily. Previous work explores only what would be true of the first test in a hypothesis space (Austerweil & Griffiths, 2008; Klayman & Ha, 1987; Navarro & Perfors, 2011). It does not investigate how such testing might change the overall nature of the hypothesis space through the selective elimination of different kinds of hypotheses. In particular, since sparse hypotheses are eliminated more often than non-sparse hypotheses by positive tests, one would expect that after a series of positive tests most hypotheses remaining would not be sparse. And since positive tests are only more effective when most hypotheses are sparse, one might expect negative tests to eventually become more efficient.

We can put these intuitions to the test by simulating a learner trying to acquire number rules. I created a Bayesian model capable of learning number rules on the basis of positive and negative evidence. The model was equipped with an explicit hypothesis space of rules determined by

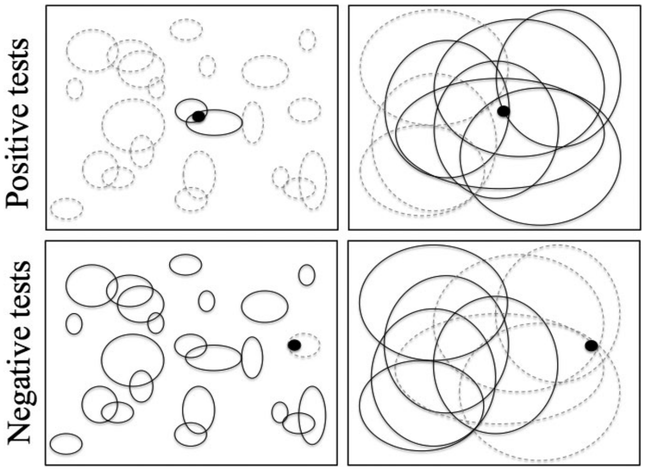

Figure 2 Schematic illustration of two hypothesis spaces in which individual hypotheses correspond to circles and boxes indicate the entire hypothesis space or set of outcomes possible. The boxes on the left show a hypothesis space with all sparse hypotheses, while the boxes on the right show one with non-sparse hypotheses. Hypotheses that are ruled out by a test are shown as dashed grey circles; hypotheses consistent with the test are solid black lines. As the top row demonstrates, a positive test can rule out many more hypotheses in a sparse space than a non-sparse one. Conversely, it is apparent from the bottom row that the opposite is true for negative tests, which rule out more hypotheses in a non-sparse space than a sparse one.

Table 1 Rules making up hypothesis space

Rule
Prime numbers
Perfect numbers
x^N, for $2 < N < 7$ (perfect squares, cubes, etc)
Powers of N, for $2 < N < 10$
Fibonacci numbers
One-digit numbers
Two-digit numbers
All numbers
Even / odd numbers
Numbers between $10N$ and $10(N + 1)$, for $1 < N < 10$
Numbers greater than N, for $1 < N < 100$
Numbers less than N, for $3 < N < 99$
The number N, for $1 < N < 101$
Numbers containing the numeral N, for $0 < N < 10$
Numbers ending with the numeral N, for $0 < N < 10$
Numbers beginning with the numeral N, for $0 < N < 10$
Multiples of N, for $2 < N < 12$ Digits are non-decreasing / increasing
Digits are non-increasing / decreasing

Table 2 Optimal data points at each step for each rule. Positive tests are labelled P and negative tests are labelled N

Rule	Step 1	Step 2	Step 3	Step 4
Digits non-decreasing	7 (P)	99 (P)	71 (N)	
Odd numbers	53 (P)	52 (N)	45 (P)	
Greater than 50	97 (P)	50 (N)	52 (P)	51 (P)
Prime numbers	7 (P)	6 (N)	2 (P)	
Multiples of 5	10 (P)	11 (N)	15 (P)	
Contains a 3	83 (P)	82 (N)	38 (P)	
Ends in a 7	7 (P)	5 (N)	70 (N)	67 (P)
Multiples of 11	22 (P)	24 (N)	55 (P)	70 (N)
Perfect squares	4 (P)	100 (P)	72 (N)	

presenting 16 participants with a paper-and-pencil task in which they were asked to list all of the possible rules for numbers between 1 and 100 they could think of, as in Perfors and Navarro (2009). After eliminating idiosyncratic rules like ADDRESSES, 408 rules for the range (1–100) remained. They are shown in Table 1.

The prior probability of each hypothesis h was set to be approximately proportional to the number of participants who suggested each rule.[7] As such, rules like EVEN NUMBERS had much higher prior probability than DIGITS ARE INCREASING or FIBONACCI NUMBERS. The likelihood of a positive test d (where d is a single number) is equal to 0 if d is not predicted by the hypothesis and $1/|h|$ if it is, where $|h|$ is the number of data points predicted by the hypothesis. The likelihood of a negative test is 0 if d is predicted by the hypothesis and $1/\neg|h|$, where $\neg|h|$ is the number of data points *not* predicted by the hypothesis.

It is possible to determine which evidence is most effective for identifying the correct rule out of the entire explicit hypothesis space of rules. This can be done through the following algorithm. First, identify the rule to be learned: call this the 'target' rule. Then, at each step, calculate, for each of the possible n data points between 1 and 100, how many incorrect hypotheses it will eliminate.[8] Then choose the data point that eliminated the most hypotheses. In the next step, repeat the process for the hypothesis space consisting of all of the hypotheses that have not been eliminated, continuing until the only hypothesis that remains is the correct one.

Note that the purpose here is not to simulate the steps taken by a learner: no learner would know what the correct rule is, so they could not calculate which data point is most efficient. Rather, the purpose is to perform an analysis of what data points—and, therefore, what kinds of tests—are most efficient at different stages in the process. How does this depend on the nature of the target hypothesis? Does its sparsity or degree of overlap with other hypotheses affect this pattern?

Table 2 shows the results of performing the above algorithm for nine different target rules that vary in sparsity and overlap. At each step, the number reported is the one that eliminates the most hypotheses (of the ones that remain in the space at that point). Positive tests are labelled P, and negative tests are labelled N. In all cases, as predicted by previous research, the optimal query at the beginning is a positive test. However, the preference for a positive test flips to a preference for a negative test after one or two positive tests. This is because, as hypothesised, the positive test(s) eliminate the sparse, non-overlapping hypotheses, paving the way for tests that will eliminate the non-sparse ones that remain. The precise pattern varies for each rule, but in each case the qualitative picture is the same: Positive tests are superior at the beginning, followed by negative tests. By the end, there is often (but not always) a return to a preference for positive tests, generally because the negative tests have eliminated most of the non-sparse hypotheses, and all that remain are a tiny number of sparse hypotheses.

Although this example is a radically simplified version of the problem confronting scientists, who must investigate a hypothesis space that is far more complex and overlapping than the one here, the analysis is abstract enough that some of its implications may yet apply. It raises the possibility that the optimal strategy of inquiry might depend how well the hypothesis space has been already explored. When a new explicit hypothesis space is identified—perhaps because new tools or a new framework opens up new questions—positive tests of the theories or methods may be more appropriate. As hypotheses get eliminated, negative tests may matter more. This implies that to the extent that neuroscience and cognitive modelling have been around for different amounts of time, a reliance on different kinds of tests in the different fields might actually be sensible. This is completely speculative, but it does suggest that it may not actually always be appropriate to evaluate different fields according to the same metric.

MULTIPLE LEVELS OF INVESTIGATION: A TOOL FOR HYPOTHESIS GENERATION?

In the previous section, I discussed how optimal hypothesis testing may depend at least in part on how long it has been since the explicit hypothesis space was first identified. But how is that done? This is the problem of hypothesis generation—the problem of opening up new areas of the latent hypothesis space so that we are aware of those possibilities and can test them. Hypothesis generation is arguably much more difficult than hypothesis testing; it is also less studied, in part because it is harder to know where to begin (though see, Farris & Revlin, 1989; Gettys & Fisher, 1979; Gettys, Mehle, & Fisher, 1986; Weber, Böckenholt, Hilton, & Wallace, 1993; M. Cherubini, Castelvecchio, & Cherubini, 2005; Dougherty & Hunter, 2003; Thomas, Dougherty, Sprenger, & Harbison, 2008; M. R. Dougherty, Thomas, & Lange, 2010, for some attempts). However, several considerations lead me to think that it is in the realm of hypothesis generation that we get the most benefit from attempting to maintain a dialogue between cognitive science and neuroscience. Indeed, I suggest that there is a tremendous advantage that arises anytime a field has multiple methodologies that simultaneously pursue answers on multiple levels of explanation.

What is the advantage? Simply put, the interaction between multiple different approaches can be a powerful driver of hypothesis generation. This occurs for a variety of reasons. One is that new methodologies allow us to explore hypotheses that we may have been previously aware of but unable to test. Just as the development of calculus allowed for the formulation of the laws of motion and the development of microscopes made it possible to test and elaborate the germ theory, so too have tools like fMRI permitted scientists to begin to investigate questions about the brain basis of behaviour to an unparalleled degree. It is possible to combine neuroscience and cognitive modelling to address questions that neither could as well on its own, as in Brown et al. (submitted). Of course, as with every new frontier—*because* it is so new—there are not yet established ways of combining two different areas, and it may therefore require more care to avoid bad science. This is not a reason to throw the baby out with the bathwater; hypothesis generation is hard enough that it is probably worth working through these growing pains.

This is a relatively weak sense of hypothesis generation, because the advent of the new methodology makes hypotheses that we were *previously aware of* now testable. A stronger sense is what happens when the crosstalk between fields results in novel insights in each, either by importing solutions from one field to another, or by identifying novel questions that aren't obvious from *either* field taken individually. This kind of crosstalk has occurred many times throughout history—for example, between thermodynamics and information theory or population biology and epidemiology. Closer to home, we can see this in the overlap between machine learning or AI and computational cognitive science. Because humans have often solved the very problems that AI struggles with, understanding how humans do so can help AI researchers design machines that do as well, as has arguably started happening in the case of vision. Conversely, AI researchers have often developed tools that greatly enrich cognitive science, like non-parametric Bayesian models. It is possible that, as the fields mature, the cross-talk between neuroscience and cognitive science will greatly enrich both areas in a similar way. One might imagine that the benefits of this sort of 'cross talk' are a function of the extent to which the explicit hypothesis spaces of the two fields involved are different; the more they overlap, the less insight one might be able to inject in the other (although, conversely, there might need to be some overlap to ensure that they can communicate at all).

CONCLUSION

This article has offered a speculative look at some of the 'big picture' issues that arise when exploring the merits of neuroscience, cognitive science, and the interaction between the two. Many of the issues arise when comparing any two methodologies that seek answers to similar sets of questions. My goal here was not to present any polished results but to suggest different ways of thinking about this issue.

One of my main suggestions is that it is in the realm of hypothesis generation—one of the most difficult elements of science—that the existence of neuroscience and the overlap between the two fields, may have its richest payoff. I also considered several issues that arise in the context of hypothesis testing. I reviewed preliminary work suggesting that whether positive or negative tests are most efficient is to some extent a function of how old a field is (or how well explored its explicit hypothesis space is already). This may create an asymmetry between relatively new fields like neuroscience and older fields like cognitive modelling. I also considered the role played by our intuitive notions of simplicity in evaluating which hypotheses to favour. I suggested that perhaps part of the 'allure of neuroscience' may arise, especially for lay people, because our folk theories of psychology view these types of explanations as particularly simple or elegant.

ACKNOWLEDGEMENTS

I would like to thank Daniel Navarro for sharing in thoughtful discussions about some of the ideas in this article. In addition, Stephan Lewandowsky and two anonymous reviewers provided many useful suggestions for improvement. This work was funded in part by ARC Grant DP110104949.

NOTES

1. The distinction between explicit and latent hypothesis spaces is very reminiscent of a similar distinction proposed by Thomas et al. (2008) and M. R. Dougherty et al. (2010), although their notion of explicit hypotheses is limited to only the 'leading contenders'. They suggest an additional distinction of those hypotheses the learner has knowledge of, which is perhaps more in line with my notion of explicit hypotheses.

2. It will surprise nobody who knows any philosophers to learn that there is some debate within the philosophical literature about the interpretation and applicability of the Bayesian framework to the process of scientific inquiry. I lack the space to go into these issues in detail, but interested readers can turn to Strevens (2006) for an introductory overview and Perfors (in press) for a discussion of the philosophical implications of explicit and latent hypothesis spaces.

3. Note that if we don't assume that all allowable outcomes are equally likely, then the qualitative point remains, but the details of how precision is calculated would differ.

4. Another notion of precision refers to precision in *measurement*, which is something quite distinct from precision in a theory. Precision in measurement focuses on the degree to which the data can be measured exactly. As Roberts and Pashler (2000) point out, if the measurement precision is low, this has implications for evaluating theories: if the error in measurement is large enough then the data may not actually constrain the theory very well at all. This problem might or might not arise as an issue for neuroscientific theories, but is not what I address here.

5. This idea is deeply intertwined with similar concepts in information theory and Kolmogorov complexity; see, e.g., Solomonoff (1964); Kolmogorov (1965); Vitànyi and Li (2000).

6. I use the word 'sparse' rather than 'precise' here for two reasons. First, the literature on positive and negative hypothesis testing uses the word sparse. Second, and more importantly, there is a subtle distinction between the words: 'sparse' refers to a particular kind of precise hypothesis—namely one which predicts less than 50% of the outcomes. That is, 'sparse' refers to a particular degree of precision. Within the space of sparse hypotheses, of course, some are sparser and more precise than others.

7. This was only approximate because some participants listed rules like MULTIPLES OF 3 while others listed MULTIPLES OF X, and it was not straightforward how to apply this to the probability of each individual 'multiple' hypothesis. Where necessary, prior probabilities were adjusted by hand to be in line with our intuitions. Assignment of the priors was done before any subsequent evaluation. Moreover, the qualitative pattern of behaviour we are interested in here depends very little on the nature of the priors: our question is how the efficacy of different kinds of evidence changes over time, but the priors do not change over time.

8. It is also possible to compare data points on the basis of which ones result in the largest change in entropy. Results are qualitatively similar.

REFERENCES

Austerweil, J., & Griffiths, T. (2008). A rational analysis of confirmation with deterministic hypotheses. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.

Brown, S., Forstmann, B., & Wagenmakers, E. (submitted). Unifying decision models: Behavioural and neural. *Australian Journal of Psychology*.

Cherubini, P., Castelvecchio, E., & Cherubini, A. M. (2005). Generation of hypotheses in Wason's 2–4–6 task: An information theory approach. *The Quarterly Journal of Experimental Psychology*, *58A*(2), 309–332.

de Zubicaray, G. (2012). Strong inference in functional neuroimaging. *Australian Journal of Psychology*, *64*, 19–28.

Dougherty, M., & Hunter, J. (2003). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica*, *113*(3), 263–282.

Dougherty, M. R., Thomas, R. P., & Lange, N. (2010). Toward an integrative theory of hypothesis generation, probability judgment, and hypothesis testing. *Psychology of Learning and Motivation*, *52*, 299–342.

Farris, H., & Revlin, R. (1989). Sensible reasoning in two tasks: Rule discovery and hypothesis evaluation. *Memory and Cognition*, *17*(2), 221–232.

Gettys, C. F., & Fisher, S. D. (1979). Hypothesis plausibility and hypothesis generation. *Organizational Behavior and Human Performance*, *24*, 93–110.

Gettys, C. F., Mehle, T., & Fisher, S. (1986). Plausibility assessments in hypothesis generation. *Organizational Behavior and Human Decision Processes*, *37*(1), 14–33.

Howson, C. (2001). *Hume's problem: Induction and the justification of belief*. Oxford: Oxford University Press.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge: Cambridge University Press.

Jeffreys, H. (1939). *Theory of probability*. Oxford: Clarendon Press.

Kalish, M., & Dunn, J. C. (2012). What could cognitive neuroscience tell us about recognition memory? *Australian Journal of Psychology*, *64*, 29–36.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*, 211–228.

Kolmogorov, A. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, *1*(1), 1–7.

Lewandowsky, S., Ecker, U. K. H., Farrell, S., & Brown, G. D. A. (2012). Models of cognition and constraints from neuroscience: A case study involving consolidation. *Australian Journal of Psychology*, *64*, 37–45.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*(3), 232–257.

MacKay, D. (2003). *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.

Navarro, D., & Perfors, A. (2011). Hypothesis generation, sparse categories and the positive test strategy. *Psychological Review*, *118*, 120–134.

Perfors, A., & Navarro, D. (2009). Confirmation bias is rational when hypotheses are sparse. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*, 302–321.

Perfors, A. (in press). Bayesian models of cognition: What's built in after all? *Philosophy Compass*.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367.

Solomonoff, R. (1964). A formal theory of inductive inference, parts 1 and 2. *Information and Control*, *7*(1–22), 224–254.

Strevens, M. (2006). The Bayesian approach to the philosophy of science. In D. M. Borchert (Ed.), *Encyclopedia of Philosophy* (2nd ed.). Detroit: Macmillan.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–641.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, *115*(1), 155–185.

Vitànyi, P., & Li, M. (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory*, *46*(2), 446–464.

Wason, P. (1960). On the failure to eliminate hypotheses in a conceptual task. *The Quarterly Journal of Experimental Psychology*, *12*(3), 129–140.

Weber, E., Böckenholt, U., Hilton, D., & Wallace, B. (1993). Determinants of diagnostic hypothesis generation: Effects of information, base rates, and experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(5), 1151–1164.

Weisberg, D. S., Keil, F., Goodstein, J., Rawson, E., & Gray, J. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, *20*(3), 407–477.