Levels of explanation in cognitive science

Max Coltheart
max@maccs.mq.edu.au
Macquarie Centre for Cognitive Science
Macquarie University
Sydney NSW 2109
Australia

Abstract

Cognitive science is the interdisciplinary study of the mind. Two of its constituent disciplines are cognitive psychology and cognitive neuroscience. Each of these can provide partial explanations of how people are able to perform certain tasks. The explanations from cognitive psychology are hypotheses about mental-information-processing programs. The explanations from cognitive neuroscience are hypotheses about the nature of the neural hardware on which these programs run. At present it is quite unclear whether findings in either of these domains can constrain hypotheses in the other; I argue that there are so far no examples of this actually having happened.

Keywords: Brain imaging; Mind; Cognitive psychology; Cognitive science; Cognitive neuroscience; Levels of explanation.

Ned Block's "The mind as the software of the brain" (Block, 1995) begins thus (p. 377):

> "Cognitive scientists often say that the mind is the software of the brain. This chapter is about what this claim means."

That's also what the present paper is about.

Block goes on (1995, p. 391) to provide a diagram of a device which consists of a cat, three mice, three pieces of cheese and some carpentry, these disposed in such a way that Mouse 3 will get some cheese *iff* both Mouse 1 and Mouse 2 also get some cheese.

He also provides a diagram of a device which consists of three switches, a battery, an electromagnet and some circuitry, these disposed in such a way that Switch 3 will close *iff* both Switch 1 and Switch 2 are also closed.

Block's point is that the two devices, despite their being composed of very different materials, are computationally identical. Both are AND-gates. He uses these diagrams to dramatize a key point: "The irrelevance of hardware realization to computational description" and elaborates this point as follows: "We are beings who have a useful and interesting biological level of description but the computer model of the mind aims for a level of description that abstracts away from the biological realization of cognitive structures." (Block, 1995, p. 390).

All of this I agree with. But then Block (1995, p. 390) adds "How such gates work is no . . . part of the domain of cognitive science." I don't see how this conception of cognitive science can be defended. Everyone agrees that cognitive science is the interdisciplinary study of how the mind works. So the move from cognitive psychology to cognitive science involves considering what disciplines other than cognitive psychology can contribute to efforts at understanding how the mind works. What are these other disciplines? Philosophy is obviously one of them; that's why the philosopher Block can write about cognitive science. Other disciplines traditionally associated with cognitive science are linguistics, computer science – and, of course, neuroscience (more specifically, cognitive neuroscience). If cognitive neuroscience is a part of cognitive science, then understanding how the gates work in Block's examples *is* part of the domain of cognitive science: that is, investigating the neural basis of cognition counts as cognitive science. Why should neuroscience be excluded from cognitive science if philosophy, linguistics, computer science (and indeed anthropology, archaeology, history etc.) are included?

In my view, what Block *should* have said here is that "How such gates work is no . . . part of the domain of cognitive psychology." That is what follows from the conception of the mind as the software of the brain. On this conception, cognitive psychology (not cognitive science, which is more general) is the study of that software. Here the mind is conceived of as containing a collection of programs such as the Face Recognition Program, the Speech Production Program, the Episodic Memory Program, and so on. And the, or at least an, aim of cognitive psychology is to learn more about the architecture of those programs (whereas in contrast the, or at least an, aim of cognitive neuroscience is to learn more about "how the gates work", that is, about the neural hardware in which any such program runs).

If it is really possible that cognitive psychologists studying face recognition might develop a complete understanding of the sequence of information-processing stages which are used to accomplish this task, then this understanding could be written down as a computer program. That would be a program that runs on the brain, but it could also be run on a computer: thus multiple realizability automatically follows from the idea that the mind is the software of the brain.

A question such as "How do people do face recognition?" might be answered by a proposal re what the functional

architecture of the face recognition program is. That would be a cognitive psychologist speaking.

Or it might be answered with a description of the Fusiform Face Area of the brain, and how it works. That would be a cognitive neuroscientist speaking.

Either kind of answer to this question "How do people do face recognition?" tells us something important about how faces are recognized. That is, each kind of answer provides an explanation in response to the question, even though both explanations are incomplete ones - both are "explanation sketches" (Hempel, 1942). We would need answers to both questions before we could claim to have a complete explanation of face recognition; but having an incomplete explanation (of either kind) is surely an advance over having no explanation.

The two answers are at different levels of description: one at the software level, the other at the hardware level. The software-level explanation would describe, in information-processing terms, what computations are performed as we execute an act of face recognition. The hardware-level explanation would describe the neural systems that are brought into action as we execute an act of face recognition.

Constraints between the two levels of explanation

Does anything that might be known at one of these levels constrain anything that one might want to claim at the other level?

This question comes in two versions, one hard and one easier. The hard version is: Is it possible in principle for there ***ever*** to be such constraints? The easy version is: are there any examples of existing work where such constraints are clearly present? I will consider only the easier question here.

Such constraints can exist in either of two directions:

(a) Brain-to-mind constraints: Has cognitive neuroscience learned anything to date that would justify any claim about what the mind ***can't*** be like? If so, what are some examples of this?

(b) Mind-to-brain constraints: Has cognitive psychology learned anything to date that would justify any claim about what the brain ***can't*** be like? If so, what are some examples of this?

I have recently (Coltheart, 2004, 2006a, 2006b, in press) taken to asking for examples of (a) or (b). No examples that I have found persuasive have yet been forthcoming.

The claim that no such examples have yet been found, or even the claim that they never will be found, is not intended to denigrate either cognitive psychology or cognitive neuroscience. Cognitive psychology would remain a valuable endeavour even if we became convinced that this discipline could never in principle tell us anything interesting about how the brain works, and cognitive neuroscience would be remain a valuable endeavour even if we became convinced that this discipline could never in principle tell us anything interesting about how the mind works. But of course it would be highly desirable if work in either of these disciplines ***does*** turn out to be capable of informing our understanding of the other.

The view that no examples of work yet exist in which cognitive-neuroscientific data have constrained cognitive-psychological theorizing is perhaps surprising given the very large number of already-published papers in which conclusions about the nature of mental information-processing are reached on the basis of data from brain-imaging studies. However, careful scrutiny of this literature reveals a number of ways in which it has so far been less than adequate.

Testing cognitive models with brain-imaging data

A particularly common problem has been in the formulation of the theories of cognition which such brain imaging studies are meant to test.

Firstly, cognitive neuroimaging sometimes completely ignores cognitive psychology: "Unfortunately, task analyses are very rarely presented in imaging papers. Whereas formal theories from cognitive psychology could often provide substantial guidance as to the design of such tasks, it is uncommon for neuroimaging studies to take meaningful guidance from such theories" (Poldrack, in press).

Secondly, even when in cognitive-neuroscientific work formal theories from cognitive psychology are invoked rather than ignored, this invocation is surprisingly often incorrect or inadequate. Consider, as one example of this, a recent paper (Wilson et al., 2007) in which brain imaging data were intended to be brought to bear on cognitive models of reading. The avowed aim of this work was to use brain imaging data to adjudicate between two competing classes of models of reading, dual-route models (e.g. Coltheart et al., 2001) according to which there are two distinct procedures by which reading aloud (computation of phonology from print) is achieved, and single-route models. The authors cite Seidenberg and McClelland (1989) as proponents of the latter class of model, referring (p. 248) to their "single-mechanism formulation". But what Seidenberg and McClelland themselves said was: "Ours is a dual route model" (Seidenberg & McClelland, 1989, p. 559). Thus the dichotomy upon which this neuroimaging paper is based, the dichotomy between single-route and dual-route models of reading, is a false one, since scrutiny of the literature on the theoretical cognitive psychology of reading reveals that there are no single-route models of reading.

In order to be sure that any research at the cognitive-neuroscientific level (such as work on brain imaging) has the potential to inform theorizing at the cognitive level, one first must define, and define correctly, what the cognitive model or models to be tested are. Furthermore, since it would be pointless to investigate any such model if it were not currently regarded by the cognitive-modelling community as tenable, it is essential that the current tenability of any model being investigated by cognitive-neuroscientific methods (that is, by brain imaging) be demonstrated. The study by Wilson et al. (2007) is by no

means alone amongst brain imaging studies of cognition in failing to meet this initial test.

Attempts at testing cognitive models with brain imaging data can take either of two forms. The work might involve just a single model, with its aim being to confirm or refute the model; or instead the work might involve competing models, with its aim being to adjudicate between these.

Consider first the situation where there is only a single cognitive model being investigated. Here investigators must be careful to avoid what Mole and Klein (in press) call "the consistency fallacy". This is the argument from a demonstration that a set of data are consistent with a theory to the claim that these data provide evidence in support of the theory. "The fact that a body of data is *consistent with* a hypothesis is not enough to show that the data provide a reason to believe that the hypothesis is true . . . an informative body of data is a body of data that enables us to rule out certain possibilities . . . In order for a body of data to provide evidence for a hypothesis the data must not only be consistent with the hypothesis, they must also count against the contradictory of the hypothesis" (Mole and Klein, in press). This means that what we want of any brain-imaging study that is aiming to test some cognitive model is that we can see what outcome of the study

(a) might well be obtained – i.e. is a plausible possible outcome – and

(b) would be inconsistent with the model being tested.

Many cognitive neuroimaging studies fail to meet this criterion. Papers reporting such studies will conclude with such statements as "Our neuroimaging data are consistent with cognitive model X", having failed to say anything at all about whether any of the possible outcomes of the study would have been inconsistent with model X.

I am not of course suggesting here that there can be crucial experiments i.e. that a model could be permanently rejected on the basis of a single experiment. But just as the outcome of an experiment can provide us with reasons to believe certain hypotheses about cognition, so can other outcomes of the experiment provide us with reasons to doubt these hypotheses. An experiment for which none of its possible outcomes could provide us with reasons to doubt a particular hypothesis about cognition cannot be presented as evidence in favour of that hypothesis, no matter how consistent with that hypothesis its actual outcome is.

The same issue arises when an experiment is intended to adjudicate between competing cognitive models. For every model involved, it needs to be shown that there is a plausible outcome of the experiment which would count as evidence against the model. I don't know of any examples in the literature of cognitive neuroscience where this has been done. It is also critical here to show that the various models involved are in fact competing i.e. at most one can be true, and also to show that all of them are currently regarded by the cognitive-modelling community as tenable. Again, one rarely if ever sees this done in work on the functional neuroimaging of cognition.

In sum, then, I have sought to make two points in this paper. The first is that if the mind is the software of the brain, a complete account of cognition will require knowing both how the software works and how the hardware works: that is why cognitive psychology and cognitive neuroscience are two of the constituent disciplines of cognitive science. This is so regardless of what conclusion one reaches about the second point I have discussed, which is to do with whether knowledge about cognition in one of these domains could constrain theorizing about cognition in the other. I have expressed some skepticism about this possibility simply on the ground that, if such constraints are actually possible, surely we should have observed actual examples of this having already happened in cognitive science? I know of no such observations.

References

Block, N. (1995). The mind as the software of the brain. In D. Osherson, L. Gleitman, S. Kosslyn, E. Smith & S. Sternberg, (Eds.), *An Invitation to Cognitive Science*, *Volume 3*. Cambridge: MIT Press.

Coltheart, M. (2004). Brain imaging, connectionism, and cognitive neuropsychology. *Cognitive Neuropsychology*, *21,* 21-25.

Coltheart, M. (2006a). What has functional neuroimaging told us about the mind (so far)? *Cortex*, *42*, 323-331.

Coltheart, M. (2006b). Perhaps cognitive neuroimaging has not told us anything about the mind (so far). *Cortex, 42,* 422-427.

Coltheart, M. (in press). What is functional neuroimaging for? In Hanson, S.J. & Bunzl, M. (Eds), *Foundational Issues of Human Brain Mapping.* Cambridge: MIT Press.

Coltheart, M., Rastle. K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual-route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*, 204–256

Hempel, C. G. (1942). The function of general laws in history. *The Journal of Philosophy, 39,* 35-48.

Mole, C. & Klein, C. (In press). Confirmation, refutation and the evidence of fMRI. In Hanson, S.J. & Bunzl, M. (Eds), *Foundational Issues of Human Brain Mapping.* Cambridge: MIT Press.

Poldrack, R. (In press). Subtraction and beyond: The logic of experimental designs for neuroimaging. In Hanson, S.J. & Bunzl, M. (Eds), *Foundational Issues of Human Brain Mapping.* Cambridge: MIT Press.

Seidenberg, M.S. & McClelland, J.L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*, 523–568

Wilson, T.W., Leuthold, A.C., Moran, J.E., Pardo, P.J., Lewis, S.M and Georgopoulos, A.P. (2007). Reading in a deep orthography: neuromagnetic evidence for dual-mechanisms. *Experimental Brain Research, 180*, 247–262.

Citation details for this article:

Coltheart, M. (2010). Levels of explanation in cognitive science. In W. Christensen, E. Schier, and J. Sutton (Eds.), *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science* (pp. 57-60). Sydney: Macquarie Centre for Cognitive Science.
DOI: 10.5096/ASCS20099
URL:
http://www.maccs.mq.edu.au/news/conferences/2009/ASCS2009/html/coltheart.html