Forthcoming in *European Journal for Philosophy of Science*

Modular architectures and informational encapsulation: A dilemma

Dustin R. Stokes and Vincent Bergeron[1]

Abstract

Amongst philosophers and cognitive scientists, modularity remains a popular choice for an architecture of the human mind, primarily because of the supposed explanatory value of this approach. Modular architectures can vary both with respect to the strength of the notion of modularity and the scope of the modularity of mind. We propose a dilemma for these modularity approaches, no matter how they vary along these two dimensions. First, if a modular architecture commits to the *informational encapsulation* of modules, as it is the case for modularity theories of perception, then modules are on this account impenetrable. However, we argue that there are genuine cases of the cognitive penetrability of perception and that these cases challenge any strong, encapsulated modular architecture of perception. Second, many recent massive modularity theories weaken the strength of the notion of module, while broadening the scope of modularity. These theories do not require any robust informational encapsulation, and thus avoid the incompatibility with cognitive penetrability. However, the weakened commitment to informational encapsulation greatly weakens the explanatory force of the approach and, ultimately, is conceptually at odds with the core of modularity.

Amongst philosophers and cognitive scientists, modularity remains a popular choice for theorizing the human mind at some broad level of architecture. Jerry Fodor (1983), who was influential in establishing the concept of a module, writes: "One day . . . Merrill Garrett made what seems to me to be the deepest remark that I have yet heard about the psychological mechanisms that mediate the perception of speech. 'What you have to

[1] This work was thoroughly collaborative and the paper thoroughly co-authored—the order of authors was chosen randomly.

remember about parsing is that basically it's a reflex.'" (Dedication). Reflexes are quick, inflexible, involuntary responses to stimuli, and Fodorian modules are like reflexes. In its most general form, the modularity hypothesis consists in viewing the human mind, or at least part of it, as a configuration of quick specialized mental mechanisms, or subsystems, that are dissociable, and that typically operate over a distinct domain of information.

There are compelling theoretical and empirical motivations for this approach. Theoretically, modularity nicely accommodates adaptationist and other evolutionary explanations of mental phenomena. It also provides materials for a simple explanation of important empirical data, including a wide range of behavioural dissociations, as well as the speed and robustness of processing enjoyed by the human mind. Most broadly, modularity provides an intuitive framework for characterizing the relations between brain structures and particular perceptual and cognitive functions.

Although it is sometimes misrepresented as doing precisely this, Fodor's pioneering discussion of the concept did not involve a definition of 'module'. (Fodor 1983; see also Coltheart 1999). Fodor did, however, provide a list of properties symptomatic of modules. Fodorian modules are *typically* domain specific, hardwired, computationally autonomous, informationally encapsulated, fast, and their operation is mandatory. It is noteworthy how much of this characterization follows the *reflex* metaphor. Domain-specificity parallels the singularity of the stimulus that sets off a reflex; autonomy, mandatoriness, hardwiring, and encapsulation mirror the standard reflex-arc model. Fodor maintains that "The notion of modularity ought to admit of degrees" (Fodor 1983: 37), *and* that "if a psychological system has most of the modularity properties, then it is very likely to have all of them" (Fodor 1983: 137). Importantly, Fodor claimed only that *input systems* are modular. His primary subject matter was perceptual systems, but he

also made the case for systems devoted to low-level linguistic decoding. Higher-level conceptual or cognitive systems, then, are not modular on Fodor's general architecture.

Commitments with respect to Fodor's original analysis of modularity vary. Several modularity theorists take domain-specificity to be definitive of modularity (Coltheart 1999). Fewer require innate specificity, even if related explanations and arguments often invoke evolutionary considerations. Others maintain that modules are informationally encapsulated and computationally autonomous (e.g. Farah 1994.)[2] Recent theorists have extended the modularity thesis beyond Fodor's input systems. It is common among evolutionary psychologists to endorse some version of what Dan Sperber (1994) has called the *massive modularity thesis*. The general hypothesis states that all, or nearly all, of the mind is modular, and modules have been postulated to account for cognitive capacities as diverse as theory of mind, face recognition, cheating detection, reading, and a variety of social understanding abilities.

Our suggestion is that *informational encapsulation* is essential to a distinctive, non-trivial modularity theory. As it will be understood here, if a module *m* is informationally encapsulated then *m* cannot, during the course of its processing, access or compute over information found in other components of the overall system. As such, an encapsulated module *m* is *impenetrable* with respect to the other components of the system, since the processing of *m* is insensitive to (and so does not compute over) the information available elsewhere in the system (Pylyshyn 1980). This basic analysis of modularity is important to any substantive modular account of the mind because it constitutes the foundation of modularity in so far as modules are, in a sense to be explained later, *dissociable systems*.

[2] In at least two places, Fodor himself explicitly states that "informational encapsulation is an essential property of modular systems" (Fodor 1985: 3; see also 1983: 71). Elsewhere, however, he is less clear on his commitment regarding the same claim.

In this respect, modularity—understood as a general approach to theorizing the architecture of the human mind—faces a dilemma that hinges on the commitment to informational encapsulation. On the one hand, a commitment to informational encapsulation, as made by modularity theories of perception, is inconsistent with the cognitive penetration of perceptual experience. And, we argue, there are genuine cases of the cognitive penetrability of perception. On the other hand, as recent modularity theorists have done, one might weaken the notion of module so as not to require informational encapsulation. (Relatedly, one might interpret literally Fodor's comment that the "notion of modularity ought to admit of degrees" (1983: 37) and thereby, focusing just on the property of encapsulation, take any system to be modular *to the degree that* it is partly but not wholly encapsulated.) The result, however, is an account that undermines one of the central motivations for modular strategies and, more fundamentally, that may not be consistent with the conceptual core of the very notion of modularity.

The first horn challenges strong, *encapsulated modularity*: any modularity approach that includes a commitment to informational encapsulation. The second horn challenges *massive modularity*, which broadens the scope of modularity but weakens the notion of modularity so as not to require informational encapsulation. Either way, modularity—as an approach to the study of mental architecture—is significantly challenged.

I. Informationally encapsulated modules: Cognitive penetrability and the challenge for encapsulated modularity

Both encapsulated and unencapsulated modularity theorists take perceptual systems to be modular. What exactly this claim amounts to varies from theorist to theorist, and indeed sometimes within the writings of the same theorist. As discussed a few

paragraphs below, Pylyshyn (1999) shifts between talk of "early visual processing" and "perception". And Fodor's initial 1983 analysis focuses on Input Systems, while later he speaks consistently of modular "perceptual systems" (Fodor 1985). Perhaps one way to bridge this apparent gap is to note how Fodor himself characterizes input systems. For vision, he identifies "mechanisms for color perception, for the analysis of shape, and for the analysis of three-dimensional spatial relations" plus "task-specific 'higher-level' systems concerned with the visual guidance of bodily motions and the recognition of faces of conspecifics" (1983:47). If we accept at face value this characterization of visual input systems, and these systems are supposed to be modular, then this is to identify much of visual perceptual systems as modular. Indeed, many theories of vision would take Fodor's list to capture all of vision proper. Accordingly, the target for this horn of the dilemma is any approach that identifies a large portion of the architecture of perception as modular. Pylyshyn and Fodor very plausibly fit this description.

If perceptual modules are informationally encapsulated, then at the very least, they are not penetrable by the information or processing of higher-level cognitive systems. Most theorists seem to take the concepts 'informational encapsulation' and 'cognitive impenetrability' to be co-extensive, if not equivalent—Fodor in fact originally argued for the encapsulation of modular systems by arguing against claims about the cognitive penetrability of those systems (Fodor 1983: 73-86). The following discussion requires only the assumption that informational encapsulation *of perceptual modules* entails cognitive impenetrability (with no commitment to the opposite entailment).

On this account, then, perceptual processing is not influenced by cognitive states like belief or desire. Evidence of this influence—that is, of the cognitive penetration of perception—thus threatens any modularity theory that includes a commitment to informational encapsulation of perceptual modules.

It will be useful here to offer some clarifications. First, distinguish perceptual experience from higher-level cognitive and affective states and processes like belief, judgement, desire, emotion, and so on. Perceptual experience, whatever else one says about it, is characterized by phenomenal character or content and depends non-trivially on one or more sensory organ. Philosophers debate how to draw the line between perception and cognition. The only point that need be granted here is that there are clear cases of perceptual states and clear cases of cognitive states. So there *are* visual experiences, auditory experiences, olfactory experiences, and so on; and these can be distinguished from states like belief and processes like decision making.[3]

Second, distinguish the cognitive penetration of perceptual experience from the cognitive penetration of perceptual processing. The former concerns some difference in the phenomenal content or character of a perceptual experience, where this difference depends non-trivially upon some cognitive state or processing in the system. The latter only concerns some cognitive effect on perception at the level of processing. The fact that perceptual processing at *some* stage is cognitively penetrated does not, by itself, entail the cognitive penetration of experience. Experience may depend on a wider class of processing and, in principle, the cognitive influences on perceptual processing (at some particular stage or other) may not ultimately influence conscious experience. Moreover, some aspects of perceptual processing may not give rise to a conscious experience but rather, for example, to the sub-personal guidance of motor performance.

However, cognitive penetration of perceptual experience does entail cognitive penetration of perceptual processing *at some level*. There is much to be said here. The only assumption we need regarding the relation between perceptual experience and

[3] So to be clear, and to foreshadow some of the discussion to come, we make no claim that perception and cognition are indistinguishable or perfectly continuous (so we do not advocate the 'continuity thesis' that is sometimes, perhaps wrongly, attributed to Jerome Bruner and New Look Psychology). Indeed, the very notion of cognitive penetration seems to presuppose some perception/cognition distinction, even if no one knows exactly how to mark that distinction.

perceptual processing is this. Whether one takes experience to be identified with, constituted by, supervenient upon, or the output of perceptual processes, a difference in perceptual experience implies a difference in perceptual process. This will be generally true—albeit for different reasons—no matter how one's metaphysics of mind varies according to these alternatives. So if experience is penetrated, then information relevant to cognitive systems, or the processing of cognitive systems, directly influences the processing of perceptual systems. It is this entailment relation that is important for the criticism offered below.[4]

One final point: Although the cognitive penetration of experience entails the penetration of processing at some stage, the cognitive penetration of experience is compatible with the cognitive impenetrability of processing at some stage or even most (but not all) stages. This point is instructive: one cannot argue from the alleged fact that some particular perceptual module is impenetrable to the claim that perception broadly or perceptual experience (in that modality) is impenetrable. Zenon Pylyshyn, for instance, argues that ''early vision'' is not penetrable by cognition (Pylyshyn 1999). Pylyshyn's empirical claim about early visual processing, even if true, is insufficient to support the thesis that perception is cognitively impenetrable. Indeed, Pylyshyn admits that the output of this component of the visual system, as he and most theorists understand it, does not (alone) determine perceptual experience. His defence of cognitive impenetrability is thus consistent with the cognitive penetrability of perception; one might accept that the computations performed by the early visual system are impenetrable by cognitive states but maintain that perceptual processing is

[4] It might be added here that if one makes a distinction between *descriptive vision* and *motion-guiding vision* (this is one terminological way of marking the distinction, see Matthen 2005), then these claims all concern descriptive vision. Accordingly, an effect on perceptual experience is an effect on descriptive vision which—according to a dominant theory in cognitive neuroscience—is plausibly an effect on processing in the neural pathway known as the *ventral stream* (see Milner and Goodale 1992, 1995).

penetrated elsewhere such that the resulting perceptual experience is causally dependent upon cognition[5]

Although there is ample evidence for "top-down" processing in the brain—where information is exchanged between various areas of the cortex, including those areas believed to process higher-level or conceptual information—current neuroscience lacks an uncontroversial mapping from conceptual mental states (like belief) onto brain structures. And some such mapping would be necessary for neuroscience to provide a verdict on the actuality of cognitive penetration.[6] Consequently, empirical evidence for cognitive penetration must be obtained at the behavioural or psychological level, rather than merely the neurological level. Predictably, there are a number of possible alternative interpretations of this data, and so the inference structure is abductive. Critics of cognitive penetrability appeal to these alternative interpretations as better explaining alleged cases of cognitive penetration. We identify four such general skeptical strategies. With these strategies in hand, a working definition of cognitive penetration can be devised, in hopes of isolating the target phenomenon in a way agreeable to both sides of the debate.

First, for some experimental and/or anecdotal cases, critics claim that what is affected by the subject's cognitive states is the subject's memory rather than her perceptual experience. Subjects recall the stimulus to be some way as a result of some other cognitive state, and report a memory of the stimulus rather than a perceptual

[5] A number of critics have questioned Pylyshyn's conclusions in this general way (Bermudez 1999; Macpherson 2012; Moore 1999; Noë and Thompson 1999). It is also worth noting that Pylyshyn's empirical claim can be challenged (see Boynton 2005; Kamitani and Tong 2005).

[6] Some have argued that evidence for reentrant neural pathways is evidence for cognitive penetration (Churchland 1988; 1989). Others have argued against this line of reasoning (Fodor 1988; Gilman 1991; Raftopoulos 2001). For purposes of this discussion, we simply assume that the neurological evidence is presently insufficient to count in either direction.

experience. This evidences cognitive penetration of cognition, and this is uncontroversial. Call this the *memory interpretation*.[7]

A second strategy is the *attention-shift interpretation*. This interpretation maintains that in some of the cases in question, cognitive states of the experimental subjects cause a shift in attention, generally involving some overt action, which then results in the change in perceptual experience. This is no different in kind, critics urge, from an ordinary perceptual scenario where one, for example, has some belief about one's environment and this belief causes some action, which in turn results in a changed perceptual experience. This familiar cognitive-behavioural dynamic is important to everyday life, but unless cognitive penetration is trivially rampant, this is not cognitive penetration. Cases involving shifts of attention, this interpretation suggests, are to be treated similarly: these scenarios lack an appropriate internal connection, and so there is nothing in this causal chain to properly call 'penetration' (see Pylyshyn 1999: 343).[8]

Third, critics have suggested that experimental subjects have only a cognitively influenced judgement of the perceived stimulus, while the perceptual experience of the stimulus remains unaffected. Thus different reports of the experimental subjects versus the control subjects in particular studies indicate differences in judgement, not perception. Call this the *judgement interpretation*.[9]

---

[7] For one example, see McCurdy 1956. Note also that if memory is *factive* as some have argued (Williamson 2002), then the memory-interpretation amounts to something like a *quasi*-memory interpretation.

[8] Fodor also appeals to this general response in his debate with Paul Churchland on the theory-ladenness of perception/observation (see Fodor 1988; Churchland 1988; see also Fodor 1983). Note also that both Fodor and Pylyshyn focus on intentionally caused *shifts* in attention, and offer little if any discussion of sub-personal attentional mechanisms. Accordingly, the attention-shift interpretation is described here in the way that it is given by the relevant critics. It remains an open question, and one very much in need of further discussion, how evidence for non-intentional attentional effects on perception bears on questions about cognitive penetration. For some related discussion, see Connolly (forthcoming), Mole (forthcoming), Stokes (2014).

[9] For additional discussion of these and other strategies for the cognitive impenetrability theorist, see Macpherson 2012; Stokes 2012, 2013.

Finally, an *intra-perceptual interpretation* claims that some of the evidenced effects are not cognitive ones but instead occur as adjustments or adaptations within the perceptual system. There are many ways to develop this alternative. One might claim that certain types of stimuli are such that the human perceptual system is appropriately "tuned" so as to represent these objects more quickly or in some enhanced way. These effects would not be learned, but would instead be artefacts of the evolution of human sensory systems. Or one might claim that a subject acquires a non-cognitive association with the stimulus type, and this association affects how perceptual information regarding tokens of this type are processed and/or how one acts in response to the stimulus. Or one may argue that sensory systems are sufficiently plastic to "learn" new ways to process sensory information, perhaps for some adaptive advantage.[10] Differences to one side, the thread common to these interpretations is that some cases may be better explained by changes in the sensory system that do not depend upon background cognitive states.

Grant that if any alleged case of cognitive penetration can be interpreted in one of these alternative ways, then the critics are correct: it is *not* a genuine case of cognitive penetration of experience. We can then define cognitive penetration so as to rule out these interpretations, and ask if any case plausibly meets the definition. If the answer is 'yes', then the critics must secure some alternative interpretation to deflect the case/s. Here, following Stokes (2013), is such a definition:

> *(CP) A perceptual experience E is cognitively penetrated if and only if (1) E is causally dependent upon some cognitive state C and (2) the causal link between E and C is internal and mental*.

[10] Against Churchland's appeal to subjects' adjustment to inverting lens as evidence for diachronic cognitive penetration, Fodor appeals to an intra-perceptual interpretation (see Fodor 1988: 193). For a more recent use of this kind of interpretation, see Deroy (2013), who analyzes some of the research also discussed below.

The definition requires a few qualifications. First, clause (2) says that if an unscreened internal cause involves a cognitive state—that is, the causal chain runs from experience back to a belief, desire, or some other cognitive state without deviating from internal mental processes—then the perception depends (internally) upon a cognitive state.[11]

Second, (CP) excludes obviously non-genuine cases of cognitive penetration. For example, a desire to see the symphony, coupled with a true belief about the location of the symphony, may result in a perceptual experience of the symphony. But this should not count as an instance of cognitive penetration of experience, else the concept 'cognitive penetration' becomes trivial. (CP) delivers the appropriate result. In cases like this, a cognitive state (or some cognitive states) motivates an action (or set of actions) which eventually results in the relevant experience. The perceptual experience thus causally depends upon the relevant cognitive state/s. Clause (2) ensures, however, that this is not an instance of cognitive penetration, since the cognitive state (or states) is screened from being *internally*, causally efficacious: the cognitive state causes an (external) action which eventually results in the experience.[12]

Importantly, a perceptual experience that satisfies this definition cannot be interpreted in any of the four ways described above. Clause (1) of (CP) rules out the memory, judgement, and intra-perceptual interpretation, since it requires a *cognitive* influence *on perception*, rather than just an influence on some other cognitive state in the system or an intra-sensory adjustment. Clause (2) of (CP) rules out the attention-shift interpretation (in a way that parodies treatment of the symphony example just above),

[11] The preferred notion of causation is of little matter so long as the internal causal dependence is maintained. For example, one could characterize the causal relation counterfactually or probabilistically. One should also note that C is a non-sufficient cause of E. There are other relevant causal factors.
[12] In this way, CP is consistent with other recently proposed definitions of cognitive penetrability: Macpherson 2012; Siegel 2011; Wu 2013.

since it requires a non-externally mediated causal link between the cognitive state and the perceptual experience. The question now becomes: are there any experimental cases that satisfy (CP)? We now consider two sets of studies that strongly suggest that the answer is 'yes', preceded by some relevant historical background.

Both sets of studies to be discussed are importantly informed by experimental work in the 1940s and 50s, identified now with the *New Look* movement in psychology. New Look psychology was important both for its explicit opposition to its behaviourist predecessors—contra behaviourism, the New Look argued that the proper explananda of psychology include internal mental states and processes—and for the *way* it made that opposition—the New Look theorized perceptual experience as an active construction of representations of the environment, substantially informed by the perceiver's expectations, needs, values, desires, and other higher-level mental states. This last feature of the theory is often described as tantamount to denouncing the perception/cognition distinction: perception and cognition are entirely "continuous". And although the pioneering new look psychologist, Jerome Bruner, explicitly rejected this characterization (see Bruner 1957), it led largely to a dismissal of New Look theorizing. But this dismissal is unfortunately overstated, and much has been and still can be learned from New Look studies.[13]

In what is probably the most famous as well as the initiating study for New Look, Jerome Bruner and C.C. Goodman (1947) employed a methodology that stands the best chance of isolating cognitive effects on perception. Experimental subjects were asked to estimate the size of currently visually perceived coins (held in one hand) by adjusting a the circumference of a small patch of light positioned six inches to the right of the grasped coin. In these studies, experimental subjects systematically overestimated the

[13] See Balcetis and Dunning 2006, Stokes 2012, 2013, and van Ulzen 2008 for brief historical discussions of the rise and fall of the New Look movement, as well as (discussion of) new studies in the New Look spirit.

size of the coins, and this effect was more pronounced for poor subjects by comparison to rich subjects.[14]

Crucial to disarming the alternative skeptical interpretations is the online nature of Bruner and Goodman's experimental procedure. Many recent studies in psychology are suggestive of cognitive penetration[15], but the most convincing studies follow some of Bruner and Goodman's basic methods. We now present two such sets of studies—the first on colour perception of natural and artificial objects, the second on the influence of racial categories on visual experience.

In a recent study, Thorsten Hansen and colleagues tested colour perception of objects with high "colour diagnositicity", objects the concepts of which are partly constituted by a distinctive colour concept: YELLOW for bananas, RED for strawberries, ORANGE for carrots, and so on (Hansen et al 2006). The procedure involved the presentation, on a computer monitor, of digital photographs of natural fruits/vegetables, presented in their typical colour, set against a uniformly grey background. The subject's task was to adjust the fruit image to what she judged to be a neutral (achromatic) grey. What in fact happens is that subjects adjust the image past achromatic grey and into the opponent colour range (e.g. adjusting a banana image past grey into the bluish hue). The researchers describe this as the *memory colour effect*. The researchers quantify this effect with a *memory colour index* (*MCI*), which in simplest terms provides a measure of the

[14] A number of theorists were critical of particular details and the broad scope of the New Look approach (Klein, Schlesigner, and Meister 1951; Carter and Schooler 1949; Lysak and Gilchrist 1955). These critics challenged some of the strongest New Look claims, and by simply acquiring evidence for cases where cognition apparently fails to affect perception. But this evidence fails to undermine the more modest implication that cognition sometimes influences perception in the relevant ways. Moreover, the Bruner and Goodman 1947 results have been broadly replicated by a number of similar studies at least insofar as these studies all evidence some higher-level effect on perceptual experience. See Bruner and Postman 1948; Postman, Bruner, and McGinnies 1948; Bruner, Postman, and Rodrigues 1951; Dukes and Bevan 1952; Bruner and Rodrigues 1953; Bruner and Minturn 1955; Blum 1957; Holzkamp and Perlwitz 1966.

[15] For example, each of the following studies present data that may be plausibly explained in terms of cognitive penetration: Balcetis and Dunning 2006, 2010; Payne 2001, 2005; Stefanuci and Proffitt 2008, 2009; Witt and Dorsch 2009. However, the experimental controls in these studies are such that the results could also be plausibly explained in terms of one (or more) of the mentioned alternative interpretations.

achromatic adjustment, towards the colour typical of the stimulus object (negative index) or away from it into the opponent hue range (positive index), relative to the typical colourfulness of the object. So for example, for a banana, the MCI is the ratio of the distance of shift past the perceiver's grey point into the bluish hue range *to* the distance of the shift from the typical yellow of the banana to the grey point, (with both of these distances measured along the same axis of typical adjustment for subjects).[16] For all of the experimental conditions, the MCI ranges from +4 to +13%, with a mean effect of +8.23%. As the researchers clarify, this quantification corresponds to an effect that is approximately three to five times above the threshold of discrimination. As a control, subjects perform the same task with uniformly coloured discs, and there is no memory colour effect: subjects adjust the discs to achromatic grey with perfect accuracy. We should emphasize that in this study (and those discussed immediately below), the task was clearly perceptual and online. Subjects took as much time as they felt necessary to make the adjustments, thus making adjustments to the perceptual stimuli in real time.

This case plausibly meets definition (CP). As the researchers hypothesize, a fruit/vegetable image, say a banana, still appears yellow to the subject at the point of achromatic grey. This hypothesis explains the fact that the subject adjusts the image into the bluish range, to compensate for the residual yellow, and then reports the fruit to be grey (when in fact it is slightly blue). This colour experience seems to depend, in a direct way, upon beliefs or conceptual associations with the relevant fruit/vegetable

[16] More specifically, the researchers clarify the calculation of the MCI as it is used in all three of the colour perception studies discussed here, as follows. "For the MCI the achromatic adjustments are projected on the axis of the typical adjustments that leads through the subjective grey point. The distance of this projection from the subjective grey point measures how strong the shift along this axis was. For the MCI this measure is divided by the length, i.e. the saturation, of the typical adjustment. In this way, the MCI represents the ratio of achromatic shift relative to the colourfulness of the typical colour. The sign (+/–) of the MCI reflects the direction in which the adjustment is shifted away from the subjective grey point. A positive MCI indicates an achromatic adjustment opposite to the typical adjustment. A negative MCI implies, contrary to the memory colour effect, that there is a shift of the achromatic adjustments towards the same direction as the typical adjustments. The MCI has been calculated separately for each participant using their subjective grey point" (Witzel et al 2011: 37).

objects. Because the testing procedure involves online adjustment of the target stimuli itself, the memory interpretation is not appropriate. For similar reasons, the attention-shift interpretation fails: there is no plausible explanation whereby subjects, in experimental but *not* control conditions, execute overt (or covert) attention to get the relevant effects.[17] And the judgement interpretation would require that, as the subject visually inspects and adjusts the target stimulus, she veridically perceives the stimulus (e.g. a banana image as slightly blue) but then reports a judgement that it is perfectly grey. So she sees the stimulus accurately but reports it erroneously. And this error has to be explained in a way that current (veridical) perception is bypassed or ignored as informing the subject's report, in spite of the task being an explicitly perceptual one. This looks much less plausible than an explanation where a non-veridical experience, itself causally dependent on background cognitive states, causes a judgement and report that the target is perfect grey (when it, the banana image for example, is in fact objectively, slightly blue). Here the report is erroneous—as the data make clear—but the error in report is explained by perceptual error, and the perceptual error is explained by cognitive penetration.

It is useful to briefly consider a possible rejoinder from the cognitive impenetrability theorist. She might respond by invoking instances where perception and judgment do come apart in just this way. So, for example, although one sees the Müller-Lyer lines as being of different lengths, one believes (if one knows the illusion) that the lines are of the same length. And indeed one cannot manage to see them accurately in spite of this background knowledge (Fodor 1983, 1985, 1988; Pylyshyn 1999). So, the critic would argue, a consistent mismatch between simultaneous experience and judgment is not so uncommon, and perhaps these experimental subjects can be explained similarly. However, the subjects in these experiments are importantly different from standard

[17] This is by contrast, for example, with Fodor's favoured explanation of the way one can shift, by attentional changes, one's experience of the Necker cube or the duck-rabbit. See Fodor 1988: 190.

perceivers of the Müller-Lyer and other such illusions. When one judges and reports that the Müller-Lyer lines are of the same length, one bases this report *not* on current perceptual experience, but on knowledge of the illusion. The subjects in the Hansen et al 2006 studies (like those in Bruner and Goodman's 1947 studies) are different in this regard: they intend for their report to be one of what they presently see. Indeed, if asked, the subjects would certainly confirm that their report—that the image is perfect grey—is based on what they see. To treat these subjects like perceivers of the Müller-Lyer illusion requires that they are systematically mistaken about this: the subjects are not reporting on the basis of what they see.

What finally of the intra-perceptual interpretation? One might worry, that since the target stimuli are all natural objects, the memory colour effect is symptomatic of hard-wired sensitivities of the human perceptual system. An enhanced perceptual sensitivity to ripe fruit and vegetables would plausibly be an evolutionary advantage for humans. And so granting that subjects still see the banana image as slightly yellow even when it is objectively grey, one might argue that this is best explained by facts about human perceptual processing and how it has evolved, without any needed appeal to cognition. This interpretation may appear even more plausible in the light of a second study performed by some of the same researchers, where the memory colour effect was most pronounced for realistic images of fruits/vegetables (e.g. those depicting texture) and mostly absent for mere fruit/vegetable outline shapes (Olkonnen et al 2008).[18]

However, this interpretation is easily dispelled by a more recent study (Witzel et al 2011). These studies involve artificial, human-made objects as stimuli. In a preliminary study, the researchers identify artificial objects with maximal *colour diagnosticity*, the blue Smurf, the Pink Panther, the red Coca-Cola logo, a green ping pong table, and so on. Images of these objects are then included in an experiment where the task is the

[18] See Deroy 2103 for an analysis that partly focuses on the Olkonnen et al 2008 study.

same as the above two studies, plus a few additional controls. Target objects are initially presented in a random colour (e.g. a fire extinguisher might appear as blue rather than its typical red colour) against a uniformly grey background, and subjects then adjust the object to what they perceive to be achromatic grey. Additionally, control objects that typically vary in colour (e.g. a sock) and control objects that are typically achromatic (e.g. a golf ball) are presented in a random colour where the task is the same. Under these conditions, there is no evident effect for control objects, and a significant effect for colour diagnostic objects. The mean MCI for fourteen colour diagnostic stimuli was +3.31%, with a high of +10.3% (for the blue Nivea tin). Just as in the earlier studies, the results provide strong evidence for a cognitive effect on perceptual experience. And importantly, the intra-perceptual interpretation is less compelling in the face of this most recent study. One version of the intra-perceptual explanation clearly does not apply: there is no story to be told about an evolved perceptual sensitivity to cartoon icons or soda logos. The opponent might attempt to maintain that ''pure'' perceptual associations are at work in these phenomena, denying any operative cognitive learning. One further reason to doubt this explanation is that these effects are culturally-variant.[19] And so, among the relevant population/s, images of smurfs, for example, are imbued with a range of semantic content. This factor lends additional plausibility to the effect being a cognitive one: subjects learn (and sometimes differently in different places) what colours *and other features* are typical of artificial objects. This is hardly conclusive, but as stated above, current empirical data is insufficient to determine some of the explanatory decisions here. Absent some further method of

[19] In their initial study to identify colour diagnosticity for artificial objects, which was performed in Germany, Witzel et al (2011) found that some stereotypically German images were highly colour diagnostic (as measured by reaction time and accuracy of typical colour identification)—for example the orange Die Maus (a German television character), the yellow German mailbox, the yellow (German-made) UHU glue tube. But some non-German objects were not sufficiently colour diagnostic (relative to German subjects)—for example, the yellow Ferrari symbol and the red Soviet flag. These researchers did not run the study using these non-colour diagnostic (relative to German subjects) objects, but presumably if they had, any memory colour effect would have been insignificant at best.

adjudication, we maintain that it is most plausible that subjects' beliefs or conceptual associations about these artificial kinds of objects (by contrast to pure, non-cognitive colour associations) affect colour experience of (images of) those objects.

Next consider a recent study on racial stereotypes and face perception (Levin and Banaji 2006). In Experiment 1, subjects were presented, on a computer monitor, with realistic greyscale images of male faces with features stereotypical of either black or white persons (with hair removed). The task was to match the luminance of an adjustable greyscale face to the target face (in some conditions the adjustable face was of the same racial prototype as the target, and in others of the opposite racial prototype). Although the luminance of the two target prototypes was (objectively) identical, subjects consistently adjusted to a lighter grey for the stereotypical white faces and to a darker grey for the stereotypical black faces (in both mixed-race and same-race conditions).[20]

In Experiment 2, Levin and Banaji first, in a preliminary study, created a racially ambiguous face by morphing a range of prototypical black and white-face features, and then confirmed the ambiguity of the face by appeal to racial classification results across 15 subjects. On an instruction screen, the ambiguous face (call this 'BW') was then paired with either an unambiguously white face (call this 'W') or an unambiguously black face (call this 'B'). And in each condition, both faces were labelled, either 'Black' or 'White'. So for example, when paired with an unambiguously white face, the ambiguous face (BW) was labelled 'Black' and the unambiguous white face (W), labelled 'White'. Taking this condition as our example—Levin and and Banaji call this the ''BW/W condition''—the task phase proceeded as follows. Subjects were presented with a series of trials, where each trial involved either the ambiguous face (i.e. the one

[20] More specifically, for example, with a black-face prototype as target, subjects adjusted a white-face prototype to 4.65 levels darker (out of 256 possible greyscale levels for the computer monitor) than a white-face prototype target (where, again, both targets are of identical luminance levels).

labelled 'Black' in the instruction phase of the BW/W condition) or the unambiguous white face (labelled 'White' in the instruction phase), both of identical luminance, coupled with an adjustable rectangular region of uniform grey. The task in each trial was to adjust the grey report patch to match the face simultaneously perceived. Result: the racially ambiguous face is reported in a way that strongly correlates with the semantic labelling prime. So, in the BW/W condition, the lightness report for the ambiguous ('black'-labelled BW) face was .465 levels darker (than the objective luminance of the target) and 17.85 levels lighter for the unambiguous ('white'-labelled W) white face. And here is perhaps the most striking result: when the same ambiguous face BW is labelled 'White' (in the opposite ''B/BW condition'') the report for BW is 15.95 levels lighter. So: present a face identical both with respect to luminance and facial features, but change the label from 'Black' to 'White', and the reported match goes from .465 levels darker to 15.95 levels lighter (than the objective luminance of the ambiguous face)(2006: 505-6).[21] This is not an effect explained just by optics or (intra-) perceptual features; the linguistic label is clearly playing an operative role in the subject's perceptual experience.

To conclude discussion of this final set of studies, consider the remaining alternative interpretations. The Levin and Banaji results are not well-explained by memory since this study involves online perceptual matching tasks. Nor is there any reason to think that overt shifts in attention are explanatory of the effects.[22] Thus both the memory and attention-shift interpretations fail to apply. What about the judgement interpretation? Considering Experiment 2, this interpretation would require that a subject, upon

[21] See footnote 19 for clarification regarding the grey measures.

[22] In fact, the researchers devise a third experiment explicitly devoted to discounting an explanation where attention is drawn to facial contours (e.g. of the stereotypical black face) in a way that explains the perceptual differences that appear in the results. They construct greyscale line drawings—with either white lines or black lines providing the facial outlines, but with no other shading of facial features—of the white and black prototypes. The results are relevantly the same and statistically significant: subjects choose darker samples for the black prototype faces and lighter samples for the white prototype faces. See Levin and Banaji 2006: 506-8.

confirming her report and moving to another trial, has veridical experiences of the target face and the report region of grey—for example, where the grey report region would appear as (on average) 15.95 levels lighter than the simultaneously perceived ('white'-labelled) ambiguous face. But then somehow the subject, in spite of perceiving this difference, judges and reports the face and report region as matching. This is far less plausible than the opposing cognitive penetrability thesis. The best explanation, here and above, is that the subject is having a non-veridical experience. She sees the prototypical white and prototypical black face as lighter and darker, respectively, and in Experiment 2, this effect is exaggerated by a linguistic labelling prime. The non-veridical experience is a result of penetrating cognitive states, in this case, racial stereotypes or beliefs.[23]

Summarizing, there are crucial methodological features common to all of these sets of studies. First, in all experiments, subjects must perform *online perceptual tasks*. Thus the target stimulus—a coin, a Smurf image, a greyscale face—is present and perceivable while the subject makes her report. One might understand this as a way of extending a now familiar methodological approach in current philosophy of perception, namely, the *method of phenomenal contrast* (Siegel 2007). In this context, the method specifies that the apparent contrast between two distinct perceptual phenomena be explained abuctively—by inference to the hypothesis that best explains the contrast. Because in these experiments perceptual experience is online and available for perceptual report, explanations in terms of memory look dubious. And the attention-shift interpretation would require substantial differences in active attention between control and experimental subjects, and this simply doesn't show up. Second, the method for reporting involves some direct kind of manipulation, either of the target stimulus itself or of some match disc or region. It is this methodology that, coupled with

[23] Macpherson 2012 briefly discusses both Hansen et al 2006 and Levin and Banaji 2006. She also provides a detailed analysis of an earlier study on colour perception, Delk and Filenbaum 1965.

the online methodology, disarms the judgment interpretation. (Compare: Many other experiments use verbal reports of some kind. And a task involving a verbal report—for example, providing a numerical estimate of the distance of a perceived object—opens space for judgement about perception and, in turn, encourages the judgement interpretation.) Finally, the stimuli used in these studies are all ones about which we learn and form beliefs, desires, and other cognitive states. It is this methodology that, at least partly, disarms the intra-perceptual interpretation.[24] As a point of methodology, we prescribe that any experimental attempts to test for cognitive penetration should employ, at minimum, this combination of features. And this approach can be traced back to Bruner and Goodman's important work of over 60 years ago.

Now, what does all of this imply for a strong modularity theory, any theory that commits to informationally encapsulated perceptual modules? If the above discussion is successful, then the standard alternative strategies fail to deflect the discussed cases as genuine evidence for the cognitive penetration of perception. In each case, whether it is a desire, value, belief, or some other higher-level mental state, there is evidently *some* cognitive state (internally) influencing experience.[25] These cases are best described as meeting the conditions of (CP). And therefore, as we will now argue, perceptual systems are not informationally encapsulated.

Recall that perceptual systems are paradigms for modular systems. And recall further that if one is an encapsulated modularity theorist, then one commits to the informational

[24] Recall that because the inference method here is abductive, to "disarm" an hypothesis means, *at best*, to render the hypothesis highly implausible and, *at least*, to show that the hypothesis is less plausible than competing alternatives, all things considered.

[25] As one anonymous reviewer notes, a number of questions about the relevant background cognitive state remain insufficiently answered in existing literature. For example, it is not made clear whether the influencing cognitive state must be an occurrent mental state. And must the effect be synchronic or may it take place diachronically? We agree that these are important questions. However, note that for our purposes, the answers don't matter. So long as the background state is cognitive and has an effect on perceptual processing (and thereby experience), then that state can be occurrent or non-occurrent, and the effect synchronic or diachronic. In other words, if a phenomenon meets the conditions specified by CP, then such a phenomenon will count against informational encapsulation no matter the answer to these other questions.

encapsulation of modules. The informational encapsulation of perceptual modules entails cognitive impenetrability. Finally, the cognitive penetration of perceptual experience entails, at some level, the cognitive penetration of perceptual processing. Therefore, any legitimate case of the cognitive penetration of experience undermines the alleged informational encapsulation of the relevant perceptual systems, and in turn challenges any theoretical architecture of those systems that commits to informational encapsulation as necessary for modules.

Here, finally, is the first horn of the dilemma for modular architectures of the mind. There are legitimate cases of the cognitive penetration of experience. We have defended two sets of studies against the relevant alternative interpretations. And so perceptual systems—in these cases *vision*—are not informationally encapsulated. Any modularity hypothesis about the architecture of perceptual systems that commits to the necessity of informational encapsulation (and by implication: cognitive impenetrability) for modularity is therefore threatened.

To clarify our critique, it is useful to briefly consider a hypothetical defence for the encapsulated modularity theorist in response to this first horn of the dilemma. Our suggestion is not that perception (or vision, more specifically) is, as it were, unencapsulated through-and-through. As discussed above, the entailment relations between the cognitive penetration of perceptual experience and the cognitive penetration of perceptual processing would not support this last inference. So, the modularity theorist might retort, the penetration of experience is compatible with the impenetrability (and thus encapsulation) of *some* (but not all) components or systems in perceptual processing, which means that some components of perceptual systems may be strongly modular.[26]

[26] We thank XXXX for pressing us to consider this reply for modularity theory.

Reply: the fact that some aspects of perceptual processing can be explained by encapsulated modularity does nothing to save a modular approach to theorizing the *architecture* of perception. For example, feature detecting components like groups of simple and complex cells in the primary visual cortex are likely encapsulated, as are many other neural circuits and low-level components in the overall visual system. In fact, it may be that certain sub-systems in vision—for example, Pylyshyn's early vision—are encapsulated in spite of the penetration of visual experience. This would be to maintain the commitment to informational encapsulation and thus a *strong* notion of 'module'. But note that the *scope* of modularity on such a view, that is, the kinds of systems to which the conceptual framework can be successfully applied, is significantly weakened. Such a modularity theorist can only claim that *some* of the visual system is modular and, importantly, cannot claim that vision is, generally, modular. This last claim *is* inconsistent with genuine cases of cognitive penetration.

It is important, moreover, not to overemphasize a characterization of (perceptual) modularity as concerning only perceptual processing. This is because an interest in mental architecture is guided not merely by goals of psychological modelling. Another crucial issue of relevance is epistemic: a modularity theory of perception promises a preferable epistemology, one where perceptual systems rapidly deliver perceptual representations in a way not prone (or less prone) to errors introduced by the cognitive agent. As Fodor puts the point, the "function of perception is to deliver to thought a *representation* of the world." And since here the goal is to represent "[n]ot the distant past, not the distant future and not…what is very far away…it is understandable that *perception* should be performed by fast, mandatory, encapsulated, etc. systems…" (Fodor 1985: 5; emphasis added). The systems in question are sub-personal modules, but the representations they provide or give rise to are personal-level experiences. (Fodor's fondness for citing the Muller-Lyer and other "persistent" illusions as examples of

cognitive impenetrability further highlights this point.) Given the epistemic role that such representations are supposed to serve, and the supposed epistemic advantage of modular perceptual systems, the modularity theorist should be no happier with evidence for penetrated experience than he is with evidence for unencapsulated perceptual processing. In short, a central motivation for modular perceptual systems entails a concern with perceptual experience.[27]

Where does this leave the view? The claim that some individual low-level circuits are encapsulated and thus strongly modular is largely uncontroversial among cognitive scientists. And the claim that some sub-systems in perception are strongly modular is insufficient to support the claim that the general structure of perception (or, more specifically, vision) is strongly modular. In turn, these weakened claims are insufficient to secure the putative epistemic benefit of modular perceptual systems. In short, one cannot save a modular architecture of perception by appeal to encapsulated perceptual components or sub-systems. To do so would be to opt for strength of modules over scope, in turn undermining the modularity hypothesis as an *architecture* of perceptual systems.

II. Informationally unencapsulated modules: A challenge for the massive modularity hypothesis

A number of recent theorists have weakened the notion of modularity with respect to Fodor's original characterization and, in particular, with respect to informational encapsulation. This change in the notion of modularity tends to accompany a

[27] Fodor makes similar suggestions elsewhere; see, for example, his discussion of "perceptual identifications" (1983: 68-71). And Pylyshyn (1980) makes similar commitments, claiming that the reliability of perception requires cognitive impenetrability. For further discussion of the epistemic consequences of cognitive penetrability, see Lyons 2011; Siegel 2012, 2013; Stokes 2012, 2013.

broadening of the scope of modular theories. Thus, massive modularity theorists take much if not the whole of the human mind to be modular, including higher level conceptual and cognitive systems. If, as we have argued in the previous section, informational encapsulation is too strict a requirement on the modularity of perception, then it makes sense to not require it of higher-level cognitive systems (Sperber 2001 is a notable exception). Weakening modularity in this way, however, comes with significant costs to any modular account of cognition. First, it greatly weakens the explanatory value of modular architectures. Second, it threatens the internal coherence of modularity theories.

Peter Carruthers, a massive modularity theorist, argues that

> *if a thesis of massive mental modularity is to be even remotely plausible*, then by 'module' we cannot mean 'Fodor-module'. In particular, the properties of having proprietary transducers, shallow outputs, fast processing, significant innateness or innate channelling, and *encapsulation* will very likely have to be struck out. (Carruthers 2006: 12; emphasis added.)

According to Carruthers, massive modularists should expect most (if not all) central cognitive modules to be *un*encapsulated. He writes:

> …even where a system has been designed to focus on and process a particular domain of inputs, one might expect that in the course of its normal processing it might need to query a range of other systems for information of other sorts. (Carruthers 2006: 10).

In other words, an unencapsulated module, in order to perform its task, will often need to compute over information that is processed and made available by other systems. For example, the mind-reading system "may need to query a whole range of other systems for information relevant to solving the task in hand" (Carruthers 2006: 11).

Evolutionary psychologists, many of whom subscribe to the massive modularity hypothesis, also tend to argue for (or assume) the compatibility of modularity with unencapsulation[28]. Hagen (2005) explicitly states what is often implicitly assumed in this field:

> Why, except when processing speed or perhaps robustness is exceptionally important, should modules not have access to data in other modules? Most modules should communicate readily with numerous (though by no means all) other modules when performing their functions, including querying the databases of selected modules (163).

Any such modularity theorist thus claims that systems, like the mind-reading system, can be modular *in spite of* being informationally unencapsulated. As Carruthers suggests, this might be a necessary adjustment of a general modular architecture for the simple reason that anything stronger is implausible.

One main theoretical advantage of, and indeed motivation for, proposing modular architectures is that they explain behavioural (or task) dissociations. A cognitive task *A* is said to be dissociated from cognitive task *B* when some individuals are observed who show a significant deficit with respect to *A* in the absence of a corresponding deficit in *B*. *A* and *B* are said to be *doubly* dissociated when, in addition, we observe individuals in whom *B* is significantly impaired without a corresponding deficit in *A*. Double

[28] Sperber (2001) is a notable exception to the view that the plausibility of massive modularity entails giving up the encapsulation requirement.

dissociations are taken as evidence of modularity because cognitive modules are, in the weakest sense, dissociable (or separately modifiable) functional components (Carruthers 2006, Sternberg 2001, 2011, Shallice & Cooper 2011)[29], and cognitive architectures composed of dissociable systems do produce behavioral double dissociations if damaged in different ways (Coltheart 2001).

To see this, suppose that two cognitive systems *A* and *B* are dissociable because *A* has a subsystem S_a that *B* doesn't have and *B* has a subsystem S_b that *A* doesn't have. If S_a is damaged while B is left intact, we should expect performance on behavioral task T_a (which depends on *A*) to be impaired while performance on behavioral task T_b (which depends on *B*) *not* to be impaired. Similarly, if S_b is damaged and *A* is left intact, then we should expect performance on behavioral task T_b (which depends on *B*) to be impaired while performance on behavioral task T_a (which depends on *A*) *not* to be impaired.

Consider, for example, the double dissociation between face recognition and object recognition—i.e. observing patients with intact visual object recognition but impaired face recognition (De Renzi & Di Pellegrino 1998), and patients with intact face recognition but impaired visual object recognition (Rumiati & Humphreys 1997). These findings suggest that the system used to recognize faces is not identical to the system used to recognize objects, and that each of the two systems has at least one subsystem that the other doesn't have[30]. This, in turn, suggests that the face recognition and object

[29] Carruthers (2006) states that "in the weakest sense, a module can just be something like: a dissociable functional component", and that "understood in this weak way, the thesis of massive modularity would… predict that the components should be separately modifiable" (p.2). Sternberg (2011) provides the following definition of cognitive module: "two sub-processes A and B of a complex process (mental or neural) are modules if and only if each can be changed independently of the other" (p. 159). It is worth mentioning that modules have also been minimally conceived as domain specific systems (e.g. Coltheart 1999 defines 'module' as "a cognitive system whose application is domain specific", p. 118). Interestingly, however, both Sternberg and Coltheart have argued that domain specificity implies separate modifiability (Coltheart 2011, Sternberg 2011).

[30] It would not, however, be reasonable to infer from this behavioral double dissociation that the face recognition and visual object recognition systems are *completely* distinct (or disjoint, see Lyons [2003]), since the two systems evidently share some of their subsystems (e.g. the subsystems responsible for low-level visual feature analysis).

recognition systems are dissociable, or in other words, that the architecture of visual recognition is modular (Coltheart 1999).

This reasoning from dissociation data to modularity—call it the *functional modularity inference* (Bergeron 2007)—has been central to the development of modern neuropsychology (Shallice 1988, Vallar 2000). In the last thirty years, philosophers and cognitive scientists have refined concepts of dissociation and narrowed the scope of the inference, and there is an emerging consensus among proponents of this approach that the reasoning should be understood as an inference to the best explanation (Coltheart 2001, Davies 2010, Shallice 1988). As we just saw, behavioral double dissociations naturally occur when modular architectures are damaged in different ways, so one can infer the existence of cognitive modules from a behavioral double dissociation on the grounds that modularity *best explains* this kind of data.

The legitimacy of inferring dissociable systems (or modules) from double dissociation data is very much a matter of debate (Coltheart 2001, Davies 2010, Dunn & Kirsner 1988, Juola, & Plunkett 2000, Machery 2012; Plaut 1995, Van Orden, Pennington, & Stone 2001). However, grant for the moment that the inference is generally sound. We then want to question whether an inference to the best explanation is supported *when modules are assumed to be unencapsulated*.

Let us suppose, then, as the weakened modularity theory we're considering does, that modules are *not* encapsulated. Suppose, for example, that the alleged face recognition and object recognition modules are not encapsulated, that they both often need to compute over information made available by other systems in order to perform their tasks.[31] This means that a double dissociation between face and object recognition

[31] Strictly, a cognitive system could be unencapsulated with respect to a whole range of systems and never have to compute over information made available by any of them. However, this is more of a conceptual possibility than an empirically plausible one, since it is hard to see why evolution or development would invest in building the relevant connections (so that the system can have access to whatever information these other systems make available) if these are never used.

could easily occur *even if* both alleged modules remained intact (i.e. were not damaged). This would happen if damage occurs (in patient 1) to any of the systems that the face recognition module needs to access, or to the connections between the face recognition module and any of these systems, and damage occurs (in patient 2) to any of the systems that the object recognition module needs to access, or to the connections between the object recognition module and any of these systems.

What this shows is that if modules are unencapsulated, there are a lot more ways to obtain a double dissociation between face and object recognition than by separately damaging (or modifying) the face recognition and object recognition systems. In fact, it shows that a double dissociation between face and object recognition could easily occur even if both functions are produced by a single system—face recognition fails when the system cannot access certain systems and object recognition fails when the system cannot access certain other systems. Thus, if we assume that cognitive modules are unencapsulated, observing a behavioral double dissociation can hardly be taken as strong evidence for the existence of dissociable systems. Put in terms of the functional modularity inference, since it is the assumed dissociability of cognitive modules that is supposed to explain behavioral double dissociation data, an appeal to *un*encapsulated modularity does not *best* explain such data.[32]

By contrast, behavioral double dissociations are adequately explained by encapsulated modularity. Importantly, if modules are informationally encapsulated, their normal functioning does not thereby depend on information made available by other systems (other than their proprietary inputs of course). Thus, a behavioral double

[32] The alleged theory of mind module is another case where unencapsulated modularity might not best explain behavioural dissociations between mind reading and other cognitive capacities. See Gerrans and Stone (2008) for a discussion of this case.

dissociation does, in this case, strongly suggest that two different systems have been separately damaged.[33]

This point about the explanatory weakness of unencapsulated modularity is worth further emphasis, since it reveals that, and in fact may be partly explained by the fact that, there is conceptual tension between the notions of unencapsulation and modularity.

To see this, recall that modules are, at the very least, dissociable systems. This minimal conception of modularity is what gives the functional modularity inference its theoretical force (see also Robbins 2013). Behavioral double dissociations are explained by modular architectures—and thus suggest the existence of dissociable systems—since cognitive architectures composed of dissociable systems give rise to behavioral double dissociations when damaged in different ways. We agree that a minimal notion of modularity should include dissociability, since without it the modular approach would seem to reduce to functional decomposition (see e.g. Barrett and Kurzban 2006 for a view of modularity that boils down to functional specificity; see also Cowie 2008 and Wilson 2008). And functional decomposition—understanding the mind in terms of functional components and sub-components—is uncontroversial as an approach, except perhaps in some connectionist quarters (more on this later).

Stated most generally, there is conceptual tension between the notions of unencapsulation and dissociability (at the level of systems) because the two properties work in opposite directions. The more unencapsulated a cognitive system is, the more likely it is that it will need to compute over information made available from other systems, and therefore the less likely it is that this system will be dissociable from these other systems. We now argue that this conceptual tension raises doubts about the

[33] This is not to say, however, that encapsulated modularity is the only plausible explanation of a double dissociation. Even with encapsulated modularity, the functional modularity inference remains abductive (Shallice 1988, Coltheart 2001).

overall soundness of (weak) modular theorizing that go beyond the use of the functional modularity inference in cognitive (neuro)psychological research.

On the one hand, two cognitive systems *A* and *B* are dissociable if *B* can be damaged (or modified) without affecting *A*'s functions, and *A* can be damaged (or modified) without affecting *B*'s functions.

On the other hand, if *A* is unencapsulated relative to *B* and *A* needs to compute over information made available by *B* in order to perform its task, then *B* cannot be damaged (or modified) without affecting *A*'s functions, so *A* and *B* are not dissociable.

Of course, *A* can be unencapsulated relative to *B* and *not* need to compute over information made available by *B* when performing a particular task at hand. Thus the claim is not that unencapsulation is incompatible with dissociability. Rather, the claim is that the more unencapsulated systems are in a given cognitive architecture, the more likely it is that these systems will need to compute over information made available by other systems (recall footnote 30), and so the less likely it is that dissociability will be a general characteristic of this architecture. That is, the less likely it is that the architecture will be (massively) modular.

To illustrate this, consider a cognitive architecture composed of several unencapsulated systems.[34] For any such architecture, there might exist some systems A and B that can be dissociated by disrupting some of their respective parts X and Y, such that disrupting X will not disrupt B, but will disrupt A and several other systems (e.g. C, D, E…) that need to access information provided by X, and disrupting Y will not disrupt A, but will disrupt B and several other systems (e.g. D, E, F…) that need to access information provided by Y. So, even though we might find some specific pairs of systems that are functionally dissociable (here A and B), dissociability does not appear to be a general characteristic of this largely unencapsulated architecture. But this is

[34] It does not matter here whether the systems share some parts.

exactly one of the core claims of massive modularity.

Here's a more concrete illustration. Consider a team of software developers responsible for creating an app for a smart phone. Each member of the team has a specific task to do, for example, building the database, creating the graphic interface, programming, integrating the app with the phone's operating system, app security, and so on. Each member must communicate frequently with several other members to make sure that the component they are working on will be compatible with other components, so as to ensure that the app will run smoothly. In a sense, then, we can say that each member of the team is unencapsulated relative to (several) other members; each must obtain specific information from other members (through regular meetings and one-to-one consultations) in order to accomplish her task. What this means, in turn, is that if a team member fails to perform her task according to plan (say, she falls behind schedule, makes a last minute change, or withdraws from the project), this would likely have a negative impact on several other members' ability to accomplish their tasks, so each member's task is likely not dissociable from other members' tasks.

Nonetheless, we may imagine ways in which a part of a task carried out by one member might be dissociable from a part of a task carried out by another member. For example, suppose the app is a fitness app (designed to help users know how much calories they burn while engaging in different physical activities) and that the person responsible for building the database needs to figure out which kinds of physical activity will be made available and for each of these what the typical energy consumption is. Suppose further that the person responsible for the app's security needs to make sure that users will have the option of turning off the location services feature. Surely, these particular subtasks could be carried out completely independently of each other and would therefore dissociate (failure to add the locations services option would not affect, for example, the decision to add brisk walking as one of the activities, and

vice versa). We would not, however, want to say that task dissociability is one of the core features of the team's overall software development project.

The main intuition behind the massive modularity hypothesis is that large complex systems must be decomposable into distinct functional components. We share this intuition, since it might be the only way to build systems capable of performing complex tasks reliably (Simon 1969, Marr 1976, Carruthers 2006). However, the building of such systems—whether by a human designer, or in the course of natural evolution or normal development—does not require that the systems' components be dissociable (see e.g. Bergeron, forthcoming, for a notion of functional independence that does not require dissociability.) What our argument shows is that when such systems are assembled from unencapsulated functional components, we cannot generally expect these components to be dissociable. To put it differently, what our analysis suggests is that the functional decomposition of large complex systems into unencapsulated functional components cannot generally rely on the notion of dissociability, but must rely instead on other forms of functional separability or functional individuation. A return to Fodor's original argument against the modularity of central systems (i.e. against massive modularity) may help further clarify this point.

Fodor, as discussed above, characterizes the specialization of perceptual systems within the strong (encapsulated) modularity framework. According to this approach, individual brain areas can be ascribed specific perceptual functions when they constitute "domain-specific computational systems characterized by informational encapsulation, high speed, restricted access, neural specificity, and the rest." (Fodor 1983: 101). It is when brain areas can be characterized in this way that, according to Fodor, we should expect to find stable (i.e. lawful) relationships between structure and function. Or, to put it differently, we should expect to find stable relationships between particular brain areas and specific cognitive functions when brain areas can perform

their computational functions independently of other brain areas (by virtue of being encapsulated).

By contrast, Fodor was much less optimistic about the prospect of finding stable structure-function relationships in the case of *un*encapsulated computational systems.

> Consider, by contrast, [unencapsulated] systems, where more or less any subsystem may want to talk to any other at more or less any time. In this case, you'd expect the neuroanatomy to be relatively diffuse. At the limit, you might as well have a random net, with each computational subsystem connected, directly or indirectly, with every other; a kind of wiring in which you get a minimum of stable correspondence between neuroanatomical form and psychological function. (Fodor, 1983: 118).

On Fodor's model, you get neural specificity, and thus stable structure-function relationship, *only* when cognitive systems perform their computations autonomously and locally. In the case of unencapsulated systems, the computational and informational resources needed to perform the task at hand are distributed across a wide range of systems, which is why (according to Fodor) we should not generally expect to find stable correspondence between unencapsulated computational systems and specific cognitive functions.

Our claim is considerably weaker than Fodor's. We do not claim that encapsulation is a requirement for the successful functional decomposition and modelling of cognitive systems. Rather, we've argued that the tension between the notions of unencapsulation and dissociability makes it unlikely that cognitive systems (e.g. the visual recognition system, the working memory system) can generally be assembled from unencapsulated and dissociable functional components. And since cognitive modules can plausibly be

minimally conceived as dissociable functional components, we doubt that modular theorizing is the right approach to the functional decomposition and modelling of such systems.

This, finally, is the second horn of our proposed dilemma, challenging modularity theorists that expand the scope of modularity *by* weakening the strength of modules so as not to require informational encapsulation. Weakening modularity to this degree greatly weakens the explanatory value of modular architectures, which in turn greatly weakens the functional modularity inference that is so frequently used in cognitive (neuro)psychological research. Second, the claim that cognitive systems could be composed (massively) of unencapsulated systems is at odds with the core idea behind modular theorizing, that is, the idea that modules are dissociable functional components. These two points are related in the following way. As we argue above, unencapsulation and dissociability work in opposite directions. And so the less encapsulated systems are the weaker the inference from behavioral dissociations to modularity (qua dissociability). As the unencapsulated computations posited increase, so do the range of possible explanations for a given set of behavioral dissociation data.

Recall further that this weakened modularity may be partly motivated by—in addition to broadened scope—acknowledgement of the apparent failure of encapsulated modularity to explain various perceptual phenomena. This was the first horn of our dilemma: perceptual systems are not well explained as encapsulated modules (or systems of encapsulated modules) if perception is cognitively penetrated. And there is compelling empirical evidence for phenomena best explained by cognitive penetration.

***

This concludes our proposed dilemma for modular approaches to theorizing the architecture of the human mind. Strong, encapsulated modularity approaches are challenged by the first horn; weakened (but broadened) unencapsulated modularity

approaches are challenged by the second horn. Either way, modularity is significantly challenged as an *empirical* strategy for studying the mind. As we have argued, informational encapsulation seems essential to a substantive and empirically well-motivated modularity, while at the same time implausible in the face of a variety of evidence for unencapsulation in perceptual and cognitive systems.

Acknowledgements

The authors gratefully acknowledge helpful comments and criticism from members of audiences at Carleton University and the Canadian Philosophical Association and in particular from Steve Downes, Matt Haber, Matthew Ivanowich, Susanna Siegel, Wayne Wu, and two anonymous referees.

References

Balcetis, E. and Dunning, D. (2006) 'See what you want to see: Motivational influences on visual perception,' *Journal of Personality and Social Psychology* 91(4): 612-25.

_____(2010) 'Wishful Seeing: Desired objects are seen as closer' *Psychological Science*: 21: 147-52.

Barrett, H. C., & Kurzban, R. (2006). Modularity in cognition: framing the debate. *Psychological review*, *113*(3): 628-647.

Bergeron, V. (2007). Anatomical and Functional Modularity in Cognitive Science: Shifting the Focus. *Philosophical Psychology*, **20**: 175–95.

Bergeron, V. (forthcoming). Functional independence and cognitive architecture. *The British Journal for the Philosophy of Science*.

Bermudez, J. (1999) 'Cognitive impenetrability, phenomenology, and nonconceptual content,' *Behavioural and Brain Sciences* 22(3): 367-8.

Blum, A. (1957) 'The value factor in children's size perception, *Child Development* 28: 14–18.

Boynton, G. M. (2005) 'Imagining orientation selectivity: Decoding conscious perception in V1,' *Nature Neuroscience*, 8: 541–542.

Brewer, W. F., & Lambert, B. L. (2001) 'The Theory-Ladenness of Observation and the Theory Ladenness of the Rest of the Scientific Process,' *Philosophy of Science* 68: 176-86.

Bruner, J.S., & Goodman C.C. (1947) 'Value and need as organizing factors in perception,' *Journal of Abnormal and Social Psychology* 42: 33-44.

Bruner, J. S., & Minturn, A. L. (1955) 'Perceptual identification and perceptual organization,' *Journal of General Psychology*, 21-28.

Bruner J. S., & Postman, L. (1948) 'Symbolic value as an organizing factor in perception,' *Journal of Social Psychology* 27: 203-208.

Bruner, J.S., Postman, L., and Rodrigues, J. (1951) 'Expectation and the Perception of Color' *American Journal of Psychology* 64: 216-227.

Bruner, J.S., & Rodrigues, J. S. (1953) 'Some determinants of apparent size,' *Journal of Abnormal and Social Psychology* 48: 17-24.

Carruthers, P. (2006) *The Architecture of the Mind*, Oxford: Oxford University Press.

Carter, L. F., & Schooler, K. (1949) 'Value need and other factors in perception,' *Psychological Review* 56,: 200-207.

Churchland, P. M. (1988) 'Perceptual Plasticity and Theoretical Neutrality: A Reply to Jerry Fodor,' *Philosophy of Science* 55: 167-187.

Coltheart, M. (1999) 'Modularity and Cognition,' *Trends in Cognitive Science* 3: 115-20.

Coltheart, M. (2011) 'Methods for Modular Modelling: Additive Factors and cognitive neuropsychology, *Cognitive Neuropsychology* 28: 224-240.

Connolly, K. (forthcoming) 'Perceptual Learning and the Contents of Perception' *Erkenntnis*

Cowie, F. (2008) 'Us, Them and It: Modules, Genes, Environments and Evolution,' *Mind and Language* 23: 284-92.

Davies, M. (2010) Double Dissociation: Understanding Its Role in Cognitive Neuropsychology. *Mind and Language*, 25:500–540.

Delk, J. L. and Fillenbaum, S. (1965) 'Differences in Perceived Colour as a Function of Characteristic Color,' *The American Journal of Psychology,* 78: 290-93.

De Renzi, E. & Di Pellegrino, G. (1998) 'Prosopagnosia and alexia without object agnosia', *Cortex* 34: 403-415.

Deroy, O. (2013) 'Phenomenal contrast without the cognitive penetrability of perception,' *Philosophical Studies*, 162: 87-107

Dukes, W. F., & Bevan, W. (1952) 'Size estimation and monetary value: a correlation,' *Journal of Psychology* 34: 43-53.

Dunn, J. C. & Kirsner, K. (2003). What can we infer from double dissociations? *Cortex*, *39*(1), 1-7.

Farah, M. (1994) 'Neuropsychological inference with an interactive brain: A critique of the locality assumption,' *Behavioural and Brain Sciences* 17: 43-104.

Fedorenko, E., Patel, A., Casasanto, D., Winawer, J. and Gibson, T. (2009) Structural intergration in language and music: Evidence for a shared system. Memory and Cognition 37(1): 1-9.

Fodor, J. (1983) *Modularity of Mind*, Cambridge, MA: MIT Press.

_____(1985) 'Precis of *The Modularity of Mind*', *The Behavioural and Brain Sciences* 8: 1-5

_____(1988) 'A reply to Churchland's ''Perceptual plasticity and theoretical neutrality,''' *Philosophy of Science* 55: 188-19

Gentaz, E. and Rossetti, Y. (1999) 'Is haptic perception continuous with cognition?' *Behavioural and Brain Sciences* 22(3): 378-9.

Gerrans, P. and Stone, V. E. (2008) 'Generous or parsimonious cognitive architecture? Cognitive neuroscience and theory of mind'. *British Journal for the Philosophy of Science* 59: 121-41.

Goodale, M. A. and Milner, D. (1992) Separate visual pathways for perception and action. *Trends in Neurosciences* 15(1): 20-25.

Hagen, Edward H. (2005). Controversial issues in evolutionary psychology. In D. M. Buss (Ed.), *The handbook of evolutionary psychology*. (pp. 5-67). Hoboken: John Wiley & Sons.

Hansen, T., M. Olkkonen, S. Walter & K.R. Gegenfurtner. 'Memory modulates color appearance.' *Nature Neuroscience* 9 (2006) 1367-8.

Holzkamp, K. and Perlwitz, E. (1966) 'Absolute oder relative Größenakzentuierung? Eine experimentelle Studie zur sozialen Wahrnehmung,' *Zeitschrift für experimentelle und angewandte Psychologie* 13: 390-405.

Juola, P. & Plunkett, K. (2000). Why double dissociations don't mean much. In G. Cohen, R. A. Johnston, & K. Plunkett (Eds.), *Exploring cognition: damaged brains and neural networks*. (pp. 319-327). Psychology Press.

Kamitani, Y., & Tong, F. (2005) 'Decoding the visual and subjective contents of the human brain,' *Nature Neuroscience* 8: 679–685.

Klein, G. S., Schlesinger, H. J., & Meister, D. E. (1951) 'The effect of personal values on perception—an experimental critique,' *Psychological Review* 58, 96-112.

Levin, D. and Banaji, M. (2006)'Distortions in the Perceived Lightness of Faces: The Roleof Race Categories,' *Journal of Experimental Psychology: General* 135: 501-12.

Lyons, J. C. (2003) 'Lesion studies, spared performances, and cognitive systems', *Cortex*, **39**, pp.145–7.
_____(2011) 'Circularity, Reliability, and Cognitive Penetrability of Perception', *Philosophical Issues* 21(1): 289-311.

Lysak, W., & Gilchrist, J. C. (1955) 'Value, equivocality, and goal availability as determinants of size judgments,' *Journal of Personality* 23: 500-501.

Machery, E. (2008), 'Massive Modularity and the Flexibility of Human Cognition,'*Mind and Language* 23: 263-72.

_____ (2012) Dissociations in neuropsychology and cognitive neuroscience. *Philosophy of Science*, 79: 490-518.

Macpherson, F. (2012) 'Cognitive Penetration of Colour Experience: Rethinking the Issue in Light of an Indirect Mechanism,' *Philosophy and Phenomenological Research* 84 (1):24-62.

Marr, D. (1976). Early Processing of Visual Information. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, *275*(942), 483-519.

Matthen, M. (2005). *Seeing, doing, and knowing: A philosophical theory of sense perception*. Oxford University Press.

McCurdy, H. G. (1956) 'Coin perception studies and the concept of schemata,'*Psychological Review* 63: 160–168.

Mole, C. (forthcoming) 'Attention-mediated cognitive penetration.' in A. Raftopoulos and J. Zeimbekis (Eds) *Cognitive Penetrability*, Oxford: Oxford University Press.

Moore, C. (1999) 'Cognitive impenetrability of early vision does not imply cognitive impenetrability of perception,' *Behavioural and Brain Sciences* 22(3): 385-6.

Noë, A. and Thompson, E. (1999) 'Seeing beyond the modules toward the subject of perception,' *Behavioural and Brain Sciences* 22(3): 386-7.

Olkkonen, M., T. Hansen, and K.R. Gegenfurtner (2008)'Colour appearance of familiar objects : effects of object shape, texture and illumination changes.' *Journal of Vision* 8: 1-16.

Payne, K. (2001). 'Prejudice and Perception: The role of automatic and controlled processes in misperceiving a weapon.' *Journal of Personality and Social Psychology* 8: 181-92.

Payne, K., Shimizu, Y., Jacoby, L. (2005) 'Mental control and visual illusions: Toward explaining race-biased weapon misidentifications.' *Journal of Experimental Social Psychology* 41: 36–47.

Plaut, D. C. (1995). Double dissociation without modularity: evidence from connectionist neuropsychology. *Journal of Clinical & Experimental Neuropsychology*, *17*(2), 291-321.

Postman, L., Bruner, J. S., & McGinnies, E. (1948) 'Personal values as selective factors in perception,' *Journal of Abnormal and Social Psychology* 43: 142–154.

Pylyshyn, Z. (1980) 'Computation and cognition: issues in the foundations of cognitive science,' *Behavioural and Brain Sciences* 3:111-132.

_____(1999) 'Is vision continuous with cognition? The case for cognitive impenetrability of visual perception,' *Behavioural and Brain Sciences* 22 (3):341-365.

Robbins, P. (2013). Modularity and mental architecture. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(6), 641-649.

Rumiati, R. I. & Humphreys, G. W. (1997). Visual object agnosia without alexia or prosopagnosia. *Visual Cognition*, *4*, 207-217.

Shallice, T. (1988) *From Neuropsychology to Mental Structure*, Cambridge: Cambridge University Press.

Shallice, T. & Cooper, R. P. (2011) *The organization of mind*. Oxford University Press.

Siegel, S. (2011) 'Cognitive Penetrability and Perceptual Justification,' *Nous* 46 (2).
_____(2013) 'The Epistemic Impact of the Etiology of Experience,' *Philosophical Studies* 162 (3):697-722.

Simon, H. A. (1969). *The sciences of the artificial*. M.I.T. Press.

Sperber, D. (1994) 'The Modularity of Thought and the Epidemiology of Representations,' in L.A. Hirschfeld and S.A. Gelman (eds), *Mapping the Mind: Domain Specificity in Cognition and Culture*, New York: Cambridge University Press.

_____(1996) *Explaining Culture: A Naturalistic Approach*, Oxford: Blackwell.

_____(2001) 'Defending massive modularity,' In E. Dupoux (ed.) *Language, Brain and Cognitive Development: Essays in Honor of Jacques Mehler*, Cambridge, MA: MIT Press

Stefanuci, J.K., and D.R. Proffitt. (2008) 'Skating down a steeper slope: Fear influences the perception of geographical slant.' *Perception* 37: 321-323.

_____(2009) 'The Roles of Altitude and Fear in the Perception of Height.' *Journal of Experimental Psychology* 35: 424–438.

Sternberg, S. (2011) Modular processes in mind and brain. Cognitive Neuropsychology, 28 (3 & 4): 156-208.

Sternberg, S. (2011) 'Modular processes in mind and brain, *Cognitive Neuropsychology* 28: 156-206.

Stokes, D. (2012) 'Perceiving and Desiring: A new look at the cognitive penetrability of experience.' *Philosophical Studies* 158: 479-92.

_____.(2013) 'The cognitive penetrability of perception' *Philosophy Compass* 8: 646-63.

_____.(2014) 'Cognitive penetration and the Perception of Art.' *Dialectica* 68: 1-34.

Vallar, G. (2000) The methodological foundations of human neuropsychology: studies in brain-damaged patients. In *Handbook of Neuropsychology* (Boller, F. and Grafman, J., eds), pp. 305–344, Elsevier.

Van Orden, G. C., Pennington, B. F., & Stone, G. O. (2001). What do double dissociation prove? *Cognitive Science*, *25*, 111-172.

van Ulzen, N.R., Semin, G.R., Oudejans, R., Beek, P. (2008) 'Affective stimulus properties influence size perception and the Ebbinghaus illusion', *Psychological Research* 72: 304–310.

Wilson, R. (2008), 'The Drink You Have When You're Not Having a Drink,' *Mind and Language* 23: 273-83.

Witt, J.K. and T.E. Dorsch. (2009) 'Kicking to bigger uprights: Field goal kicking performance influences perceived size.' *Perception* 38: 1328-1340.

Witzel, C., Valkova, H., Hansen, T., Gegenfurtner, K. (2011) 'Object knowledge modulates colour appearance' *i-Perception* 2: 13-49.

Wu, W. (2013) 'Visual Spatial Constancy and Modularity: Does Intention Penetrate Vision?' *Philosophical Studies*: 165 (2):647-669

Young, A. W. (1996). Face recognition. In J. G. Beaumont, P. M. Kenealy, & M. J. C. Rogers (Eds.), *The Blackwell dictionary of neuropsychology*. Blackwell.