Journal of Cognition and Neuroethics

Reverse Inference and Mind-Brain Identity

Yakir Levin
Ben-Gurion University of the Negev

Itzhak Aharon
Interdisciplinary Center (IDC) Herzliya

Biographies

Itzhak Aharon (Gingi) is a senior lecturer at the Interdisciplinary Center (IDC) Herzliya, Israel. His research focuses on the neurobiology of motivation and decision making (neuroeconomics).

Yakir Levin is a senior lecturer in philosophy at Ben-Gurion University of the Negev, Israel. His research interests include early-modern philosophy, analytic metaphysics, and philosophy of mind.

Acknowledgements

Earlier versions of the paper were presented at two conferences: (1) The Aims of Brain Research: Scientific and Philosophical Perspectives, co-organized by the Safra Center for Brain Sciences (ELSC) and the Edelstein Center, Hebrew University, the Cohn Institute, Tel Aviv University, and the Van Leer Jerusalem Institute; (2) The Nathan Stemmer Memorial Colloquium on the Philosophy of the Science of Morality co-organized by the Edelstein Center and the Center for Moral and Political Philosophy, Hebrew University. Thanks are due to the following participants in these events for their helpful comments and suggestions: Jean-Pierre Changeux, Meir Hemmo, Eva Jablonka, Tom Polger, and Adina Roskies (first conference); David Enoch, Edouard Machery, Boaz Miller, and Chandra Sripada (second conference). In developing the main argument of the paper, we deploy some material from Sects. 2 and 3 of Levin and Aharon 2011 in a significantly revised form.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). September, 2015. Volume 3, Issue 2.

Citation

Levin, Yakir, and Itzhak Aharon. 2015. "Reverse Inference and Mind-Brain Identity." *Journal of Cognition and Neuroethics* 3 (2): 23–45.

Reverse Inference and Mind-Brain Identity

Yakir Levin and Itzhak Aharon

Abstract
Reverse inference is a widespread procedure of reasoning from patterns of brain activation to the engagement of specific mental processes. One of the main attractions of reverse inference is the apparent possibility it opens for exceeding the limits of behavior based procedures in psychology. Underlying this motivation is an implicit assumption of mind-brain identity according to which behavior does not play a constititive role in the classification of mental kinds. A widely accepted consideration, however, against mind-brain identity is that while identity is a one-one relation, there appears to be mounting evidence that the mind-brain relation is one-many. In this paper we examine three recent strategies for responding to this consideration, positing that two of them fail, while the third may be successful but at the cost of giving the behavioral criteria of mental kinds a constitutive role to play in the classification of brain kinds. For this reason, the third strategy does not yield an account of mind-brain identity capable of grounding a positive answer to the question of whether reverse inference can exceed the limits of behavior based procedures in psychology. It follows that philosophical defenses of mind-brain identity may be useless for the purposes of science. Nevertheless, they are not irrelevant to science, since their failure to ground a specific scientific research strategy such as reverse inference may constitute an important negative lesson concerning this strategy.

Keywords
Reverse Inference, Mind-brain Identity, Multiple Realization, Degeneracy, Pleiotropy

1. At the Intersection of Cognitive Neuroscience and Philosophy

1.1 Forward vs. Reverse Inference

The classic strategy employed by neuroimaging researchers has been to manipulate a specific psychological function and identify the localized effects of this manipulation on brain activity. This has been referred to as "forward inference" (Henson 2005), and is the basis for a large body of knowledge that has been derived from neuroimaging research. However, since the early days of neuroimaging, there has also been a desire to reason backward from patterns of activation to the engagement of specific mental processes. This has been called "reverse inference" (Poldrack 2006), and often forms much of the reasoning observed in the discussion section of neuroimaging papers (under the guise of "interpreting the results") (Poldrack 2011, 692).

Having been so widespread in the neuroimaging literature – an "epidemic" as one writer has described it (Poldrack 2006) - reverse inference's common use has not gone unchallenged (Aguirre 2003; Poldrack and Wagner 2004; Henson 2005; Page 2006; Poldrack 2006; Christoff and Owen 2006; Poldrack 2008; Harrison 2008; Van Horn and Poldrack 2009; Bourgeois-Gironde 2010; Poldrack 2011; Fox and Friston 2012). This has led to its becoming a bad name in some quarters (Poldrack 2011, 692). However, short of undermining it, these criticisms have paved the way for a more cautious and sophisticated use which, e.g., focuses on patterns of activation rather than localized blobs, utilizes broader and more comprehensive fMRI databases, makes use of high-performance computer clusters, deploys improved techniques of statistical analysis etc. (Poldrack 2011 and 2012; Hutzler 2014). These criticisms, moreover, have certainly not deterred neuroimaging researchers – especially in areas such as neuroeconomics and social neuroscience in which the underlying mental processes are less well understood - from regarding reverse inference as a fundamentally important research tool (see, e.g., Young and Saxe 2009) – "the *sine qua non* of inference in neuroeconomics" as one researcher has put it (Harrison 2008, 535).

1.2 Behavior Exceeding Reverse Inferences

One reason that reverse inference has been considered so important in neuroeconomics is that it is often not possible to determine the correctness of cognitive theories adduced in this field solely on behavioral basis. Thus, consider the well-known tendency of consumers to behave "impatiently" today but to prefer/plan to act "patiently" in the future. For example, someone offered the choice between receiving $10 today and $11 tomorrow is likely to choose the immediate option. However, if asked today to choose between $10 in a year and $11 in a year and a day, the same person is likely to prefer the slightly delayed but larger amount. One hypothesis that has been advanced to explain this phenomenon is that it reflects the operation of two fundamentally different mechanisms, one affective, which heavily values the present and steeply discounts all future opportunities, and the other deliberative, which discounts options more consistently across time. However, it has not been possible to provide evidence for separate mechanisms from behavioral data alone, or to motivate them on the basis of purely theoretical considerations (Sanfey, Loewenstein, McClure, and Cohen 2006, 113).

It has been, therefore, the hope of neuroeconomists that neurobiological data could play here the evidential role that behavioral data does not, perhaps even cannot, play. And indeed, as a recent fMRI study has shown, choices involving the option of

an immediate reward actively engage the ventral striatum, as well as the medial and orbitofrontal areas – areas rich in dopaminergic innervation, and consistently associated with the evaluation of reward (McClure, Laibson, Loewenstein, and Cohen 2004). In addition, this study has shown that both choices involving the option of an immediate reward and those involving the option of a delayed reward consistently involve areas of the frontal and parietal cortex commonly associated with more abstract forms of reasoning and planning. And this has been taken to corroborate the aforementioned hypothesis – viz., that the difference between the short-term and long-term choices at issue reflects the operation of two different cognitive mechanisms, one affective and the other deliberative (Sanfey, Loewenstein, McClure, and Cohen 2006, 113; Camerer, Loewenstein, and Prelec 2005, § 5.1; but cf. Kable and Glimcher 2007; Bernheim 2009, § 1.5.1).

In like manner, reverse inference has been employed for dissociating social decision making of the sort often employed in behavioral economics games (e.g., trust game, ultimatum game, prisoner's dilemma game, etc.), and decision-making concerning non-social stimuli or agents, where this cannot be done on a behavioral basis (Lee and Harris 2013). Thus, it has been shown that attributions of behavioral traits to human agents on the basis of their observed behavioral patterns, rely on a distinct set of brain regions, including the medial prefrontal cortex (MPFC) – a region known to be active when value signals in a social context are created – and the superior temporal sulcus (STS). However, when the agents are anthropomorphized objects, although the same patterns of behavior as those manifested by human agents lead to the attribution of the same behavioral traits, the underlying pattern of brain activity is different. Specifically, attributions for objects do not engage MPFC but rather STS and the bilateral amygdala. And this has been taken to suggest that social and non-social decision making involve different cognitive mechanisms.

Another example of a use made of reverse inference in the absence of clear behavioral indications concerns decision making under uncertainty (Huettel et al. 2006). Thus, preferences for risk (uncertainty with known probabilities) have been shown to correlate with activation of the posterior parietal cortex. In contrast, preferences for ambiguity (uncertainty with unknown probabilities) have been shown to correlate with activation of the lateral prefrontal cortex. And this has been taken to indicate that, contrary to standard assumptions, decision making under ambiguity does not represent a special, more complex case of risky decision making; instead, these two forms of uncertainty involve distinct cognitive mechanisms.

Yet another example of a use made of reverse inference in the absence of clear behavioral indications concerns cooperative behavior (Stallen and Sanfey 2013). Thus, reciprocated cooperation in the prisoner's dilemma game and the trust game have been shown to correlate with activity in the ventral striatum and the ventromedial prefrontal cortex, brain regions consistently found to be activated by both social and monetary rewards. Relatedly, viewing the faces of individuals who had previously cooperated in a prisoner's dilemma game, as compared to faces with whom the player had no history, elicited enhanced neural activity in reward-related areas such as the striatum, nucleus accumbens, and orbitofrontal cortex. And this has been taken to suggest that people are motivated to resist the temptation to selfishly accept but not reciprocate favors, by labeling mutual cooperation as rewarding in and of itself – i.e., independent of whatever monetary gain was obtained by the cooperative action.

1.3 Mind-Brain Identity, Multiple Realization, and the Scope of Reverse Inference

A major reason, then, for the attraction of reverse inference is the hope/assumption that it makes it possible to exceed the limits of behavior based procedures in psychology/economics, and achieve what is impossible by these procedures alone. A similar hope/assumption seems to be lurking behind the claims by prominent neuroeconomists such as Camerer, Loewenstein, and Prelec that the study of the brain and the nervous system is "beginning to allow direct measurement of thoughts and feelings," or that neuroscience is making possible the measurement of "basic psychological forces … without inferring them from behavior," or enables "to observe processes and constructs which are typically considered unobservable," or opens for the first time "the 'black box' [sometimes more wittingly called the 'grey box'] … – the human mind" (Camerer, Loewenstein, and Prelec 2004, 572; Camerer, Loewenstein, and Prelec 2005, 10 and 53; Camerer 2007, c38; Camerer 2008a).

Underlying these hopes/assumptions is an implicit assumption of mind-brain identity which takes the behavioral expressions of mental features to be mere contingent indications of these features that do not play a role in their classification into mental kinds (more on this below): If mental features can be inferred on the basis of neurobiological features (such as patterns of activation) when there are no clear - or perhaps even any - behavioral indications for their presence, this would mean that only neurobiological features are constitutive of these mental features and required for their classification into kinds. It would mean, in Francis Crick's (1994, 3) provocative formulation, that "you, your joys and your sorrows, your memories and your ambitions, your sense of

personal identity and free will, are in fact no more than the behavior of a vast assembly of nerve cells and their associated molecules"; that "you are your synapses," in Joseph LeDoux's (2003, 323-324) equally disturbing words; that "mental processes actually are processes of the brain," in Patricia Churchland's (2008, 409) somewhat more prosaic phrasing. So the question of whether reverse inference can exceed the limits of behavior based procedures boils down to the question of the viability of an account of mind-brain identity which does not give behavior a constitutive role in the classification of the brain kinds with which mental kinds are identified. Thus reformulated, however, this question of the scope of reverse inference becomes related to the prototypical argument contra mind-brain identity (Polger 2011, 9; see also Polger 2009, 458), the *multiple realization objection* according to which: *The mind-brain relation is one-many* – i.e., one and the same mental kind can be realized or subserved by distinct kinds of brain structures. But, *the identity relation is one-one* – i.e., if a mental kind is realized or subserved by distinct brain kinds, it cannot be identical with any one of them. Thus, *mind isn't identical to the brain*.

Although there has been a wide consensus among philosophers ever since it was put forward by Putnam and others in the early 1960s that this objection is devastating, it has not gone unchallenged. The first wave of challenges which goes back to the late 1960s, didn't question the very phenomenon of multiple realization of mental kinds by brain kinds. Instead it sought to show that this phenomenon does not really pose a problem to mind-brain identification, or to a reduction of mind to the brain (Bickle 2010, 248-251). A second wave of challenges which arose about a decade ago, has taken a different tack attacking the very claim that mental kinds are multiply realized by brain kinds (Bickle 2010, 253-255; cf. Polger 2008, 538). A major focus of this wave of challenges has been the question of how the realization relation between realized and realizing kinds is to be analyzed, and the related question of how brain realizers are to be classified into kinds. With this focus, however, this challenge and the debate it has evoked, is particularly relevant to the question of the scope of reverse inference in its reformulation as the question of the viability of an account of mind-brain identity which does not give behavior a constitutive role in the classification of the brain kinds with which mental kinds are identified. Do the classificatory criteria of these brain kinds implied by second-wave-responses involve behavioral features? If they do, insofar as the mind-body identity defended by the second-wave-responses is concerned, reverse inference cannot exceed the limits of behavior based procedures.

1.4 Aim, Structure, and Implications

In this paper we address this question. To this end we shall first clarify what we mean by mind-brain identity-cum-reduction (§ 2). We shall then outline a set of considerations that have been widely taken to support the claim that the mind-brain relation is one-many, or that mind is multiply-realized by the brain (§§ 3-4). Finally, we shall outline the second-wave strategies of responding to these considerations, positing that two of them fail, while the third may be successful but at the cost of giving the behavioral criteria of mental kinds a constitutive role to play in the classification of brain kinds (§§ 5-7). Due to this aspect of the third strategy, it does not yield an account of mind-brain identity capable of grounding a positive answer to the question of whether reverse inference can exceed the limits of behavior based procedures (§ 8).

An important implication of this conclusion is that philosophical defenses of mind-brain identity may be useless for the purposes of science. This is not to say, however, that these defenses are irrelevant to science, since their failure to ground a scientific research strategy such as reverse inference, or some of its intended uses-cum-goals, may be an important negative lesson concerning that strategy or those intended uses-cum-goals. So the paper illustrates an interesting link between cognitive neuroscience and philosophy. The other side of the same link, which the paper also illustrates, is the heavy reliance of the second wave of challenges to multiple realization on neuroscientific practices and findings.

2. Mind-Brain Identification and Reduction

2.1 Empirical Identification

Suppose that being in a mental state of a given kind is one and the same thing as being in a brain state of a specific kind. For example, suppose that being in pain is one and the same thing as being in a state of P-fiber excitation. ('P' here stands for the parieto-insular cortex, for which there is growing evidence of its centrality for pain processing in the brain. We also use the term 'P-fiber excitation' in homage to the ubiquitous yet wrong reference that philosophers make in this context to C-fiber excitation: C-fibers are located well outside the brain, existing as a subset of the sensory neurons that project to the spinal cord [Pucceti 1977; Allen, Grau, and Meagher 2009, 129-130].) If such mind-brain identities obtain, they cannot, of course, be certifiable conceptually, or solely from the meanings of the mental terms (e.g., 'pain') and brain terms (e.g., 'P-fiber excitation') that they involve. The concept of pain and the concept of P-fiber excitation are distinct

and independent concepts, and this explains how it is possible for someone to know a lot about pains but nothing about P-fibers or their excitations. If pains are identical with P-fiber excitations, this is an empirical truth whose corroboration depends on elaborate neurobiological research. In like manner, the concept of the morning star (Phosphorus) and that of the evening star (Hesperus) are distinct concepts, and it could only have been discovered by observation and experience that they signify one and the same thing – viz., Venus. And this is also the case with "theoretical identities" in the sciences such as the identity of water and H_2O, or of heat and molecular motion, or of light and electromagnetic radiation. In being empirical, all these identities differ significantly from an identity such as "2=the smallest prime number," which are known a priori, or by merely investigating the meanings of the expressions that they involve.

2.2 Identity of Kinds and Classificatory Criteria

Prior to the discovery of an empirical identity between two kinds, then, these kinds are considered distinct and independent from each other. This being the case, each kind must be associated with a distinct and independent classificatory criteria, or a set of features that determine the things that belong to the kind – e.g., specific intricate behavior in the case of pain, and specific neurobiological features in the case of P-fiber excitation. After the discovery of the identity, however, one of these sets of features may be considered more fundamental than the other, most naturally, the set characterizing the lower level type, a level that is typically considered as of a higher explanatory value ("wherever the bottom is, that is where the real explanations are to be found" is a widely accepted view, not least so in contemporary neuroscience [Craver 2007, 11-15]). Indeed, the set of classificatory criteria considered less fundamental may no longer be considered constitutive or essential to the kind; it may no longer be considered as playing a role in determining the kind, but rather as mere contingent evidence for the occurrence of an instance of the kind.

If such a change in status of classificatory criteria occurs, however, the mere occurrence of something with features of the fundamental classificatory set may be considered sufficient for the occurrence of an instance of the kind, even if it is not accompanied by an occurrence of the other set of classificatory features. Thus, in the 17th century fluidity was considered essential to water (Locke 1975, 4.6.11), and ice, which lacks this feature, was considered a different substance (ibid., 3.6.13). So the transformation of water into ice (or vice versa) was considered then like the transformation of a wooden table to a pile of bits and pieces of wood when the table is

broken apart: Like the latter transformation the former one wasn't considered a change in a mode or state of a thing, but as a transformation of one thing to a completely different thing. After the discovery of the micro-structure of water – viz., H_2O - and the identification of the kind water with the kind H_2O that ensued, this was no longer the case, and something that had not been considered before as looking and behaving like water – viz., ice – began to be considered that way (as water in a solid phase). Indeed, its "non-watery" behavior became part of what is considered standard "watery" behavior in specific circumstances. Relatedly, once water has been identified with H_2O, things that behave and look like water but are not H_2O, cannot be considered as water anymore (Putnam 1975, 223-235).

2.3 From Kind-Identification to Reduction

An identification of one kind of phenomenon (usually of a higher level) with another kind of phenomenon (usually of a lower level) may involve, then, a change in the status of the classificatory criteria that, prior to the identification, were regarded as determining each kind. As a result of the identification only the classificatory criteria associated with the lower level kind may be viewed as constitutive of the unified kind, or as determining it. Insofar as identifications of kinds of phenomena involve such changes in status, they are tantamount to a reduction of the phenomena of the higher level kind to the phenomena of the lower level kind. The question of how common reductions of this sort are in science, may be controversial (Antony 2007, 154-156; Craver 2007, 108-109). Be that as it may, a reduction of this sort may be rather appealing in the mind-brain case. For underlying our common sense view of the mind are Cartesian intuitions which incline us to consider the behavioral criteria of mental states as mere contingent evidence for utterly different things. Cartesians consider these utterly different things as states of a non-material substance. But from here it is but a short step to consider them as states of a material substance, or the brain. So a mind-brain identification-cum-reduction can be considered a materialized version of Cartesianism (Bennett and Hacker 2003, 72).

2.4 The One-One Assumption

Having clarified what we mean by mind-brain identification-cum-reduction, the question immediately arises of what can be the basis for such an identification-cum-reduction? Well, the first and crucial step in this direction is the discovery of systematic mind-brain correlations. Indeed, all serious arguments for mind-brain identity must come up with reasons that would make it compelling, or at least reasonable, to upgrade

mind-brain correlations to identities (Kim 2005, Chap. 5). The mind side of these correlations, however, consists of mental states as these are behaviorally manifested, or can be identified by the behavioral criteria for their ascription: The systematic correlations between pain and P-fiber excitation, say, that are upgraded to an identity between the two, are in fact correlations between pain as it is behaviorally manifested, or genuine pain-behavior, and P-fiber excitation. Thus, a basic hypothesis underpinning the thesis of mind-brain identity is the One-One assumption according to which mind in its behavioral manifestations and brain are systematically – indeed, one-one – correlated – i.e., the same brain structures or processes accompany or subserve the same mental state as this state manifests itself or can be identified by observable behavior. The multiple realization thesis, to which we now turn, is directed against this assumption.

3. Enter Degeneracy and Pleiotropy

A ubiquitous property of biological systems at all organizational levels is *degeneracy*, or the ability of elements that are structurally different to perform the same function or to yield the same output (Edelman and Gally 2001; Greenspan 2001 and 2003). As a biological hypothesis, degeneracy has been posited to explain a number of studies of biological organisms, from yeast to humans, in which striking structural differences at various suborganism-levels appear to have little or no organism-level effects (see the references in Edelman and Gally 2001). To explain, for example, why mutations that eliminate the function of various genes need not cause overt harm (an explanation given in terms of overlapping networks of genes that, given appropriate conditions for gene expression, can produce the same outcome) (see Greenspan 2001 and the references therein).

Being a prerequisite for natural selection as well as a product of this process, degeneracy goes hand in hand with *pleiotropy*– i.e., degenerate structures tend to be versatile in their functions, and usually can be used differently in different processing contexts (Tononi, Sporns, and Edelman 1999; Edelman and Gally 2001; Greenspan 2001; Noppeney, Friston, and Price 2004). Thus, a given gene may subserve function F when activated within one gene network and function G when activated within another. It is also the case that biological structures can exhibit degeneracy within an individual at a time or over time, across individuals of the same species, or even across individuals of different species.

Appropriated to neural systems, degeneracy may be taken to mean that diverse brain structures can subserve or realize the same mental state as this state manifests itself or

can be identified by observable behavior. Thus defined, evidence for neural degeneracy both within and across species appears to abound (Price and Friston 2002; Greenspan 2003; Noppeney, Friston, and Price 2004; Noppeney, Penny, Price, Flandin, and Friston 2006; Aizawa 2007; Aizawa and Gillet 2009a; Aizawa and Gillett 2009b; Richardson 2009). In particular, and although they should be considered with caution, lesion and imaging studies which frequently show that entirely different anatomical areas of the brain can subserve the same cognitive functions, have been widely taken to provide strong evidence for neural degeneracy (Figdor 2010, 428-431; cf. Polger 2008, 461-469; Polger 2011, 10). (Imaging studies sometimes replicate findings of lesion studies [Dronkers, Redfern, and Knight 2000], and sometimes also combine with the latter to provide new data [Price and Friston 2002].)

It is also the case that there is ample evidence for neural pleiotropy: Different cognitive functions appear to be supported by putting the same neural circuits together in different arrangements. In each of these arrangements, an individual brain region may perform a similar information-processing operation (a single "working"), but will not be dedicated to the high-level task to which the arrangement is dedicated as a whole (Anderson 2010a; Anderson and Penner-Wilger 2013).

Neural pleiotropy does not really challenge the One-One assumption. At most, it evinces that the mind-brain correlation that this assumption posits must be - at least partially - between mental states on the one hand and patterns of neural activation rather than local neural blobs on the other hand. Nevertheless, neural pleiotropy presents a serious - although perhaps not insurmountable - challenge to reverse inference (Anderson 2010b, 295; Ramsey et al. 2010; Poldrack 2012, 1217-1218). By contrast, neural degeneracy presents a serious challenge to the One-One assumption. In consequence, it also presents a serious and perhaps insurmountable challenge to behavior exceeding reverse inferences. It is to this dual challenge that we now turn.

4. Degeneracy and Multiple Realization

A particularly natural way of viewing neural degeneracy is as showing that mind is multiply realized by the brain – i.e., that contrary to the One-One assumption, there is a one-many mind-brain correlation (cf. Fidgor 2010). Thus, semantic processing tasks (e.g., picture naming) are presumably subserved by different brain structures, in normal subjects and lesion patients, respectively (Price and Friston 2002). And how can this be taken but as showing that, contrary to the One-One assumption, the same semantic

processing task can be correlated with brain states of different kinds? Well, here are three second-wave responses to this challenge.

5. Going Deeper Down Science's Ontological Hierarchy

Going back to a suggestion made by Paul Churchland (1982) and elaborated more recently by John Bickle (2003) this response concedes that if we leave our neuroscientific understanding at the systems level, then psychoneural multiple realization will appear obvious and unavoidable, especially across species. However, as we move further down levels, into cellular physiology and into the intracellular signaling pathways, commonalities - even across widely divergent species – may be the rule – i.e., molecular pathways that underlie specific cognitive and conscious functions may be the same ones from invertebrates to mammals. Bickle's key psychological example throughout his writings on the topic has been memory consolidation, or the conversion of labile, easily disrupted short-term memories into more durable, stable long-term form. In his view, "the discovery of the shared molecular mechanisms for memory consolidation is probably not some isolated, lucky case, but rather follows from core principles of molecular evolution. As 'molecular and cellular cognition' proceeds, we should expect to discover more evolutionarily conserved examples of unitary molecular 'reducers' of shared psychological kinds. Molecular evolution suggests that they should be the rule, not the exception" (Bickle, 2010, 258).

Somewhat ironically, however, Bickle's key example of his proposal appears to refute this very proposal: As Aizawa (2007) has convincingly shown, the biochemical mechanisms of memory consolidation uncovered by molecular neuroscience reveal substantive multiple realization both across and intra species. And if so, Bickle's defense of the One-One assumption fails.

On top of that, in focusing on a low, cellular level of the brain, Bickle's strategy may be irrelevant to the purposes of neuroimaging strategies such as reverse inference which focus on a high, system level of the brain.

6. Eliminate and Split

Another response to neural degeneracy-cum-multiple realization is to eliminate the multiply realized mental kind by splitting it into uniquely realized kinds each corresponding to one of the different realizing brain kinds (Aizawa and Gillett 2011). A notable example of an actual employment of this strategy is the splitting of memory into distinct sub-types in response to neurobiological dissociation experiments: Once it was

discovered that certain sorts of brain lesions lead to the selective loss of certain memory functions, while certain other sorts of brain lesions lead to selective loss of certain other memory functions, the common assumption that there is a single over-arching type of memory has been replaced by the assumption that there are distinct subtypes of memory, declarative and nondeclarative (Squire 2004). However, as Aizawa and Gillett (2011) show, the elimination-by-splitting strategy cannot be considered a general strategy that is applicable across the board, since that would fail to reflect the nuances of actual scientific practice. Thus, as the science of color vision illustrates, differences amongst realizers may lead only to scientists positing individual differences in the same higher level property rather than to subtyping this property. More strikingly still, and again as the science of color vision illustrates, discovered variations in realizers may lead to no variation in the higher level realized properties. And in that case, not only would the subtyping of color vision by way of its lower level property instances be cumbersome, but using the lower level realizer properties to classify higher level properties into kinds may leave us without higher level theories that can track important regularities or generalizations at the higher level.

7. Reject a Presupposition of the Considerations Pro Multiple Realization

As we saw in § 3, multiple realization is supported by cases of mental functions (e.g., semantic processing tasks such as picture naming) that are subserved by different brain structures. How, rhetorically ask proponents of multiple realization, can this be taken but as showing that the One-One assumption is false?

Yet, an implicit and rather natural assumption that underlies this way of viewing multiple realization – an assumption shared by the strategies of going deeper down science's ontological hierarchy and of eliminate and split - is that diversity in realizer *structure* is tantamount to diversity in realizer *kind*: Unless the diversity in the brain *structures* that subserve picture-naming, means diversity in the *kind* of these brain structures *as* realizers of picture-naming, neural degeneracy in this case would not imply that picture-naming can be subserved by brain states of diverse *kinds*.

Thus formulated, this *structure-determines-kind* assumption, leaves open the question of how different, and in what respects, brain structures have to be in order to belong to different brain kinds (Figdor 2010, Sec. 3). Be the answer to this question as it may (Edelman and Gally 2001; Sullivan 2008, Sec. 2; Aizawa and Gillett 2009b; Figdor 2010, 435), for our purposes suffice it to point out that one can keep to the One-One assumption despite neural degeneracy by way of contesting the structure-determines-

kind assumption. It is to two important responses to the multiple realization objection along this line that we now turn.

7.1 Bechtel's and Mundale's Proposal

The first way of contesting the structure-determines-kind assumption is implied by Bechtel's and Mundale's (1999) seminal attack on the hypothesis that psychological functions are multiply realized. As part of their attempt to show that neuroscientific practice contradicts this hypothesis, they claim that brain mapping practices show that brain taxonomy makes essential use of psychological function. This claim may be contested by way of the very examples that Bechtel and Mundale bring in its support (Aizawa 2009, Section 2). Thus, it seems that brain mapping techniques that involve staining brain tissue in order to highlight different features of brain cells, and discerning differences in structure over the volume of the brain, do not make use of psychological function. For the sake of argument, however, suppose that the classification of brain structures must indeed proceed by appeal to psychological function. In that case, however diverse the brain structures that are correlated with a given mental function are, they can still be considered as forming a single unified kind by the very fact that they are all correlated with the same kind of mental function. Thus, however different the distinct brain structures that were found to subserve picture naming are, by Bechtel's and Mundale's lights these brain structures belong to the same kind due to their correlation with picture naming. Thus, Bechtel's and Mundale's strategy makes it possible to maintain the One-One assumption despite neural degeneracy by way of rejecting the structure-determines-kind assumption. This rescuing move comes with a price, however.

In taking psychological functions to play an essential role in the classification of brain structures with which they are correlated, Bechtel's and Mundale's strategy also gives the behavioral criteria of psychological functions a role in the classification of brain structures. This being the case, this strategy preserves one building block of the mind-brain identity thesis – viz., the One-One assumption – but involves a rejection of another building block of this thesis – viz., the *reductionist assumption* that the behavioral criteria of mental kinds do not play a constitutive role in determining these kinds. If the psychological kind of picture naming is identified with a kind of brain structure, and, as by Bechtel's and Mundale's strategy, this brain kind is determined by the behavioral criteria for picture naming, then these behavioral criteria also play a role in determining the mental kind of picture naming.

7.2 Shapiro's Proposal

The second way of contesting the structure-determines-kind assumption is by way of Shapiro's account of the realization relation (Shapiro 2000, 2004, and 2008). On this account, realizers should be classified on the basis of the causal mechanisms by which they yield the functional types that they realize. Thus, a waiter's corkscrew and a double lever corkscrew are different types of realizers of the function of removing corks from bottles, since they each achieve this function in different ways; each employs a different mechanism in the production of cork removal. In contrast, although steel and aluminum waiter's corkscrews differ in constitution, they should be considered of the same type, since they share the same mechanism for corkscrewing bottles. Similarly, eye types such as the octopus eye and the mammalian eye that focus light onto photoreceptive cells in the same way are considered of the same kind ("camera eye"), even if they achieve these optical characteristics by, say, different molecular structures (Shapiro 2000, 646). In other words, "it is optics that provides the level of description at which a clump of molecules constitutes an eye, and hence it is the science of optics that determines whether two eyes are instances of a single kind of realization or, rather, are instances of different realizations" (Shapiro 2004, 95).

Adopted and defended also by Polger (2008, 2010, and 2013; Polger and Shapiro 2008; Shapiro and Polger 2012), Shapiro's account has been contested by Aizawa and Gillett (2003, and Aizawa 2009a, 2009b, and 2011), who have offered an alternative account of realization according to which multiple realization would be rather pervasive. For our purposes we do not have to go into the details of this sophisticated and very interesting debate, nor for that matter even go to the fine details of Shapiro's account (though, some of its aspects – e.g., the relativity and intransitivity of the realization relation that it implies – may be relevant for the assessment of, e.g., the strategy of going deeper down science's ontological hierarchy – cf. Polger 2008, 544 n. 5). Suffice it to point out that, applied to the mind-brain relation, this account may undermine the structure-determines-kind assumption. For just as structurally different waiter's corkscrews – a steel one and an aluminum one, say – can realize the function of corkscrewing in exactly the same way, and thus belong to the same realizing type of this function, so may different brain structures realize a given psychological function in the same way and thus belong to the same realizing type of this function. Indeed, based on (1) his account of the realization relation, and (2) the claim that there may well be natural constraints on the kind of structure that is capable of rendering a humanlike psychology, a claim that gets support from instances of neural convergence (i.e., the independent evolution of similar kinds of neural structures), Shapiro argues that (3) it seems plausible that any organ that

exhibits humanlike psychological capacities must also possess various humanlike brain properties (Shapiro 2004, Chaps. 3-4).

However, in taking realizing brain structures to be classified by the way they bring about their realized psychological functions, Shapiro's strategy gives the latter a constitutive role in the classification of the former. So like Bechtel's and Mundale's strategy it also gives a role in the classification of brain structures to the behavioral criteria of psychological functions. Thus, and again like Bechtel's and Mundale's strategy, Shapiro's strategy for defending the One-One assumption must involve a rejection of the reductionist assumption of the mind-brain identity thesis according to which the behavioral criteria of mental kinds do not play a role in determining these kinds. To recapitulate, if the psychological kind of picture naming is identified with a kind of brain structure, and, as by Shapiro's strategy, this brain kind is determined by the behavioral criteria for picture naming, then these behavioral criteria also play a role in determining the mental kind of picture naming.

7.3 A Tint of Functionalism

A *functional property* is a property specified by a job description, or by a certain function this property can perform. Thus, showing the time is a functional property of clocks. According to the functionalist conception of the mind, or *functionalism*, mental kinds are determined by functional properties of the body, which are defined in terms of the role they play as causal intermediaries between perceptual input, other mental states, and behavioral output (Kim 1996, Chap. 5; Antony 2007; Levin 2013). For (an avowedly simplistic) example, a functionalist theory might identify the state of believing that it is raining with the functional property of being in a state that tends to be produced when it is raining, and, given one's belief that by using an umbrella one can avoid getting wet as well as one's desire not to get wet, leads one to take an umbrella. Alternatively, such a theory might identify the state of believing that it is raining with a brain state, all of whose concrete instantiations tend to be produced when it is raining, and, given one's belief that by using an umbrella one can avoid getting wet as well as one's desire not to get wet, causes one to take an umbrella.

Identifying the mental state of believing that it is raining with a higher-level functional property, the first example illustrates the so called *role* version of functionalism (Levin 2013, § 3.4). Identifying the same mental state with a brain state the classificatory criteria of which are constituted by the higher-level functional property that it realizes, the second example illustrates the so called *realizer* version of functionalism (ibid.).

In identifying mental states with brain states that are determined by aspects of higher-level psychological functions and their behavioral classificatory criteria, both the proposal of Bechtel and Mundale and that of Shapiro have significant affinities with realizer functionalism. This is rather ironic, since these proposals seek to defend functionalism's main rival – viz., mind-brain identity theory – by undermining the multiple-realization objection to this rival position.

8. Back to Reverse Inference

The question of whether reverse inference can exceed the limits of behavior based procedures boils down, as we have seen, to the question of the viability of a mind-brain identity thesis which does not give behavior a constitutive role in the classification of the brain kinds with which mental kinds are identified. The question of the viability of the latter thesis depends, in turn, on whether the One-One assumption can be defended against the multiple realization objection. Of the three defense strategies of this assumption that we outlined, the first two – going deeper down science's ontological hierarchy and eliminate and split – are unsuccessful, while the two versions of the third strategy – Bechtel's and Mundale's on the one hand, and Shapiro's on the other hand – may succeed but at the cost of giving behavior a role in the classification of the brain kinds with which mental kinds are identified. This being the case, these two strategies do not yield an account of mind-brain identity capable of grounding a positive answer to the question of whether reverse inference can exceed the limits of behavior based procedures. It follows that philosophical defenses of mind-brain identity may be useless for the purposes of science. Nevertheless, they are not irrelevant to science, since their failure to ground a specific scientific research strategy such as reverse inference or some of its intended uses-cum-goals may form an important negative lesson concerning this strategy.

References

Aguirre, G.K. 2003. "Functional Imaging in Behavioral Neurology and Cognitive Neuropsychology." In *Behavioral Neurology and Cognitive Neuropsychology*, edited by T.E. Feinberg and M.J. Farah, 85–96. New York: McGraw-Hill.

Aizawa, K. 2007. "The Biochemistry of Memory Consolidation: A Model System for the Philosophy of Mind." *Synthese* 155: 65–98.

Aizawa, K. 2009. "Neuroscience and Multiple Realization: A Reply to Bechtel and Mundale." *Synthese* 167: 493–510.

Aizawa, K. and Gillett, C. 2009a. "Levels, Individual Variation, and Massive Multiple Realization in Neurobiology." In *The Oxford Handbook of Philosophy and Neuroscience*, edited by J. Bickle, 539–581. New York: Oxford University Press.

Aizawa, K. and Gillett, C. 2009b. "The (Multiple) Realization of Psychological and Other Properties in the Sciences." *Mind and Language* 24: 181–208.

Aizawa, K. and Gillett, C. 2011. "The Autonomy of Psychology in the Age of Neuroscience." In *Causality in the Sciences*, edited by P.M. Illari, F. Russo, and J. Williamson, 202–223. New York: Oxford University Press.

Allen, C., Grau, J.W., and Meagher, M.W. 2009. "The Lower Bounds of Cognition: What do Spinal Cords Reveal?" In *The Oxford Handbook of Philosophy and Neuroscience*, edited by J. Bickle, 120-142. New York: Oxford University Press.

Anderson, M.L. 2010a. "Neural Reuse: A Fundamental Organizational Principle of the Brain." *Behavioral and Brain Sciences*, 33: 245–266.

Anderson, M.L. 2010b. "Cortex in Context: Response to Commentaries on Neural Reuse." *Behavioral and Brain Sciences* 33: 294–313.

Anderson, M.L., and Penner-Wilger, M. 2013. "Neural Reuse in the Evolution and Development of the Brain: Evidence for Developmental Homology?" *Developmental Psychobiology* 55: 42–51.

Antony, L. 2007. "Everybody Has Got It: A Defense of Non-Reductive Materialism." In *Contemporary Debates in Philosophy of Mind*, edited by B.P. McLaughlin and J. Cohen, 143–159. Oxford: Blackwell.

Bechtel, W. and Mundale, J. 1999. "Multiple Realizability Revisited: Linking Cognitive and Neural States." *Philosophy of Science* 66: 175–207.

Bennett, M.R. and Hacker, P.M.S. 2003. *Philosophical Foundations of Neuroscience*. Oxford: Blackwell.

Bernheim, B.D. 2009. "On the Potential of Neuroeconomics: A Critical (but Hopeful) Appraisal." *American Economic Journal: Microeconomics* 1: 1–41 (NBER Working Paper Series, 13954).

Bickle, J. 2003. *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht: Springer.

Bickle, J. 2010. "Has the Last Decade of Challenges to the Multiple Realization Argument Provided Aid and Comfort to Psychoneural Reductionists?" *Synthese* 177: 247–260.

Bourgeois-Gironde, S. 2010. "Is Neuroeconomics Doomed by the Reverse Inference Fallacy?" *Mind & Society* 9: 229–249.

Camerer, C.F., Loewenstein, G., and Prelec, D. 2004. "Neuroeconomics: Why Economics Needs Brains." *Scandinavian Journal of Economics* 106: 555–579.

Camerer, C.F., Loewenstein, G., and Prelec, D. 2005. "Neuroeconomics: How Neuroscience Can Inform Economics." *Journal of Economic Literature* XLIII: 9–64.

Camerer, C.F. 2007. "Neuroeconomics: Using Neuroscience to Make Economic Predictions." *Economic Journal* 117: c26–c42.

Camerer, C.F. 2008a. "Neuroeconomics: Opening the Gray Box." *Neuron* 60: 416–419.

Christoff, K. and Owen, A.M. 2006. "Improving Reverse Neuroimaging Inference: Cognitive Domain Versus Cognitive Complexity." *Trends in Cognitive Sciences* 10: 352–353.

Churchland, P.M. 1982. "Is Thinker a Natural Kind?" *Dialogue* 21: 223–238.

Churchland, P.S. 2008. "The Impact of Neuroscience on Philosophy." *Neuron* 60: 409–411.

Craver, C.F. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Clarendon Press.

Crick, F. 1994. *The Astonishing Hypothesis: The Scientific Search for the Soul*. New York: Touchstone Press.

Dronkers, N.F., Redfern, B.B., and Knight, R.T. 2000. "The Neural Architecture of Language Disorders." In *The New Cognitive Neurosciences*, edited by M. Gazzinga, 949–958. Cambridge, MA: MIT Press.

Edelman, G.M. and Gally, J.A. 2001. "Degeneracy and Complexity in Biological Systems." *Proceedings of the National Academy of Sciences of the USA* 98: 13763–68.

Figdor, C. 2010. "Neuroscience and the Multiple Realization of Cognitive Functions." *Philosophy of Science* 77: 419–456.

Fox, P.T. and Friston, K.J. 2012. "Distributed Processing; Distributed Functions?" *NeuroImage* 61: 407–426.

Gillett, C. 2003. "The Metaphysics of Realization, Multiple Realizability, and the Special Sciences." *The Journal of Philosophy* 100: 591–603.

Greenspan, R.J. 2001. "The Flexible Gnome." *Nature Reviews Genetics* 2: 383–387.

Greenspan, R.J. 2003. "Darwinian Uncertainty." *KronoScope* 3 (2): 217–225.

Harrison, G.W. 2008. "Neuroeconomics: A Rejoinder." *Economics and Philosophy* 24: 533–544.

Henson, R.N. 2005. "What Can Functional Imaging Tell the Experimental Psychologist?" *Quarterly Journal of Experimental Psychology A* 58: 193–233.

Huettel, S.A., Stowe, C.J., Gordon, E.M., Warner, B.T., and Platt, M.L. 2006. "Neural Signatures of Economic Preferences for Risk and Ambiguity." *Neuron* 49: 765–775.

Hutzler, F. 2014. "Reverse Inference is not a Fallacy Per Se: Cognitive Processes Can be Inferred from Functional Imaging Data." *Neuroimage* 84: 1061–1069.

Kable, J.W., and Glimcher, P.W. 2007. "The Neural Correlates of Subjective Value During Intertemporal Choice." *Nature Neuroscience* 10: 1625–1633.

Kenning, P., and Plassmann, H. 2005. "Neuroeconomics: An Overview from an Economic Perspective." *Brain Research Bulletin* 67: 343–354.

Kim, J. 1996. *Philosophy of Mind*. Boulder: WestviewPress.

Kim, J. 2005. *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

LeDoux, J. 2003. *Synaptic Self: How Our Brains Become Who We Are*. London: Penguin Books.

Lee, V.K. and Harris, L.T. 2013. "How Social Cognition Can Inform Social Decision Making." *Frontiers in Neuroscience* 7: 1–13.

Levin, J. 2013. "Functionalism." *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/functionalism/.

Levin, Y. and Aharon, I. 2011. What's On Your Mind? A Brain Scan Won't Tell." *Review of Philosophy and Psychology* 2: 699–722.

Locke, J. 1975. *An Essay Concerning Human Understanding*. Edited by P.H. Nidditch. Oxford: Clarendon Press.

McClure, S.M., Laibson, D.I., Loewenstein, G., and Cohen, J.D. 2004. "Separate Neural Systems Value Immediate and Delayed Monetary Rewards." *Science* 306: 503–507.

Noppeney, U., Friston, K.J., and Price, C.J. 2004. "Degenerate Neuronal Systems Sustaining Cognitive Functions." *Journal of Anatomy* 205 (6): 433–442.

Noppeney, U., Penny, W.D., Price, C.J., Flandin, G., and Friston, K.J. 2006. "Identification of Degenerate Neuronal Systems Based on Intersubject Variability." *NeuroImage* 30: 885–890.

Owen, A.M. and Coleman, M.R. 2008. "Functional Neuroimaging of the Vegetative State." *Nature Reviews Neuroscience* 9: 235–243.

Page, M.P.A. 2006. "What Can't Functional Neuroimaging Tell the Cognitive Psychologist?" *Cortex* 42: 428–443.

Poldrack, R.A. and Wagner, A.D. 2004. "What Can Neuroimaging Tell Us About the Mind? Insights from Prefrontal Cortex." *Current Directions in Psychological Science* 13: 177–181.

Poldrack, R.A. 2006. "Can Cognitive Processes Be Inferred from Neuroimaging Data?" *Trends in Cognitive Science* 10: 59–63.

Poldrack, R.A. 2008. "The Role of fMRI in Cognitive Neuroscience: Where Do We Stand?" *Current Opinion in Neurobiology* 18: 223–227.

Poldrack, R.A. 2011. "Inferring Mental States from Neuroimaging Data: From Reverse Inference to Large-Scale Decoding." *Neuron* 72: 692–697.

Poldrack, R.A. 2012. "The Future of fMRI in Cognitive Neuroscience." *Neuroimage* 62: 1216–1220.

Polger, T.W. 2008. "Two Confusions Concerning Multiple Realization." *Philosophy of Science* 75: 537–547.

Polger, T.W. 2009. "Evaluating the Evidence for Multiple Realization." *Synthese* 167: 457–472.

Polger, T.W. 2010. "Mechanisms and Explanatory Realization Relations." *Synthese* 177: 193–212.

Polger, T.W. 2011. "Are Sensations Still Brain Processes?" *Philosophical Psychology* 24: 1–21.

Polger, T.W. 2013. "Realization and Multiple Realization, Chicken and Egg." *European Journal of Philosophy* 21 (1): 1–16.

Polger, T.W. and Shapiro, L.A. 2008. "Understanding the Dimensions of Realization." *Journal of Philosophy* 105: 213–222.

Puccetti, R. 1977. "The Great C-Fiber Myth: A Critical Note." *Philosophy of Science* 44: 303–305.

Price, C.J. and Friston, K.J. 2002. "Degeneracy and Cognitive Anatomy." *Trends in Cognitive Sciences* 6 (10): 416–421.

Ramsey, J.D., Hanson, S.J., Hanson, C., Halchenko, Y.O., Poldrack, R.A., and Glymor, C. 2010. "Six Problems for Causal Inference from fMRI." *NeuroImage* 49: 1545–1558.

Richardson, R.C. 2009. "Multiple Realization and Methodological Pluralism." *Synthese* 167: 473–492.

Sanfey, A.G., Loewenstein, G., McClure, S.M., and Cohen, J.D. 2006. "Neuroeconomics: Cross-Currents in Research on Decision-Making." *Trends in Cognitive Sciences* 10: 108–116.

Shapiro, L.A. 2000. "Multiple Realizations." *Journal of Philosophy* 97: 635–654.

Shapiro, L.A. 2004. *The Mind Incarnate*. Cambridge MA: The MIT Press

Shapiro, L.A. 2008. "How to Test Multiple Realization?" *Philosophy of Science* 75: 514–525.

Shapiro, L.A. and Polger, T.W. 2012. "Identity, Variability, and Multiple Realization in the Special Sciences." In S. Gozzano and C.S. Hill (eds.) *New Perspectives on Type Identity: The Mental and the Physical*, 264-287. Cambridge: Cambridge University Press.

Squire, L.R. 2004. "Memory Systems of the Brain: A Brief History and Current Perspective." *Neurobiology of Learning and Memory* 82: 171–177.

Stallen, M. and Sanfey, A.G. 2013. "The Cooperative Brain." *The Neuroeconomist* 19: 292–303.

Sullivan, J.A. 2008. "Memory Consolidation, Multiple Realization, and Modest Reductions." *Philosophy of Science* 75: 501–513.

Tononi, G., Sporns, O., and Edelman, G.M. 1999. "Measures of Degeneracy and Redundancy in Biological Networks." *Proceedings of the National Academy of Sciences of the USA* 96: 3257–3262.

Van Horn, J.D., and Poldrack, R.A. 2009. "Functional MRI at the Crossroads." *International Journal of Psychopisiology* 73: 3–9.

Young, L. and Saxe, R. 2009. "An FMRI Investigation of Spontaneous Mental State Inference for Moral Judgment." *Journal of Cognitive Neuroscience* 21: 1396–1405.