Behavioral and Brain Sciences

The Reification Objection to Bottom-Up Cognitive Ontology Revision

--Manuscript Draft--

Manuscript Number:	
Full Title:	The Reification Objection to Bottom-Up Cognitive Ontology Revision
Short Title:	The Reification Objection to Bottom-Up Cognitive Ontology Revision
Article Type:	Commentary Article
Corresponding Author:	Edouard Machery, Ph.D University of Pittsburgh Pittsburgh, PA UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Pittsburgh
Corresponding Author's Secondary Institution:	
First Author:	Joseph B. McCaffrey
First Author Secondary Information:	
Order of Authors:	Joseph B. McCaffrey
	Edouard Machery, Ph.D
Order of Authors Secondary Information:	
Abstract:	Anderson proposes a bottom-up approach to cognitive ontology revision: Neuroscientists should revise their taxonomies of cognitive constructs on the basis of brain activation patterns across many tasks. We argue that such bottom-up proposal is bound to commit a mistake of reification: It treats the abstract mathematical entities uncovered by dimension reduction techniques as if they were real psychological entities.

Author of the book: Michael Anderson

Word count:

Abstract: 59

Main text: 1974

References: 315

Total: 2441

Title: The Reification Objection to Bottom-Up Cognitive Ontology Revision

Full Names: Joseph B. McCaffrey and Edouard Machery

Institution:

Joseph B. McCaffrey

Department of History and Philosophy of Science

University of Pittsburgh

Edouard Machery

Department of History and Philosophy of Science

University of Pittsburgh

Institutional/mailing address

Edouard Machery

Department of History and Philosophy of Science

University of Pittsburgh

1017CL

Pittsburgh PA 15217

USA

Telephone #: 4126245883

Emails:

jbm48@pitt.edu

machery@pitt.edu

Homepage URL:

http://www.josephbmccaffrey.com

http://www.hps.pitt.edu/profile/machery.php

Abstract

Anderson proposes a bottom-up approach to cognitive ontology revision: Neuroscientists should revise their taxonomies of cognitive constructs on the basis of brain activation patterns across many tasks. We argue that such bottom-up proposal is bound to commit a mistake of reification: It treats the abstract mathematical entities uncovered by dimension reduction techniques as if they were real psychological entities.

Main Text

Reverse inference consists in inferring that a task recruits a psychological process (*P*) on the grounds that a brain structure (*S*) is activated during this task (as observed by, e.g., fMRI). It is often assumed that reverse inference is valid only if activation is *selective*, that is, if the ratio

$P(activation\ of\ S\ |P\ is\ recruited) \Big/ P(activation\ of\ S\ |P\ is\ not\ recruited)$ is high

(Poldrack, 2006). Since brain areas are typically multi-functional, cognitive neuroscientists have grown skeptical of area-based reverse inference. Anderson endorses this pessimistic conclusion—"It should go without saying that we must also curtail the common practice of reverse inference" (Anderson, 2014, 113)—and the first two chapters of *After Phrenology* extensively review the multifunctionality, hence low selectivity, of brain regions.

One can address the problem raised by multifunctionality in three different ways. First, reverse inference can be reformulated so as to depend on *diagnosticity* instead of selectivity (Machery, 2014). In this approach, reverse inference is valid only if the activation discriminates between the recruitment of a first psychological process, *P*, and of a second psychological process, *P'*, that is, only if the ratio

$P(activation\ of\ S\ |P\ is\ recruited) \Big/ P(activation\ of\ S\ |P'\ is\ recruited)$ is high.

Second, one can increase the selectivity of brain activation by revising cognitive neuroscientists' *brain ontology*: Instead of focusing on regional activation, one can reverse infer on the basis of activation in other brain structures (e.g. networks) that may be selectively associated with psychological processes (e.g., Glymour & Hanson,

forthcoming). In Chapter 4 of *After Phrenology*, Anderson rejects this second approach on the grounds that brain networks too can be multifunctional. Anderson's concern here is speculative, and more evidence is needed before discrediting brain ontology revision. Large-scale brain networks (e.g., effective connectivity networks), or activation patterns within those networks (e.g., as measured by MVPA), may be far more selective or diagnostic than individual regions. Third, one can increase the selectivity of brain activation by revising cognitive neuroscientists' *cognitive ontology*: On this approach, activation of brain structures is not selective because cognitive neuroscientists lack the right set of cognitive constructs for describing the functions or computations that these structures perform (e.g., Poldrack, 2010).

This third approach has led to a lively debate about cognitive ontology revision (Klein, 2012; Lenartowicz, Kalar, Congdon, & Poldrack, 2010; McCaffrey, 2015; Poldrack, Halchenko, & Hanson, 2009; Price & Friston, 2005). As Anderson perspicuously notes, most "revisionists" have a *conservative* goal: Taking current cognitive ontology as their starting point, they attempt to validate cognitive constructs by investigating whether they can be selectively associated with brain activation patterns (e.g., Lenartowicz et al., 2010). By contrast, Chapter 4 of *After Phrenology* advocates a *revolutionary* goal. Anderson's project is not to determine which members of current cognitive ontologies are valid and which are invalid, but rather to propose entirely new cognitive constructs by mining fMRI datasets. Before describing and assessing Anderson's proposal, we note that it is unclear whether his goal is to revolutionize the constructs *psychologists* are working with (e.g., recommending they stop using the construct of working memory) or, less

ambitiously, whether he is proposing a new cognitive ontology for *cognitive neuroscientists*: In this case, the idea would be to develop novel ways of characterizing what neural structures do.

Anderson's central idea is that cognitive neuroscientists should not characterize *the intrinsic function* of each brain region—i.e., the operation the region performs independently of its neural context (e.g., its computational function); instead, they should quantitatively characterize each region's *disposition* to be involved in a given set of tasks. Anderson calls such dispositions "neural personalities." Neural personalities allegedly vary with respect to some fundamental psychological dimensions (or "neuroscientifically relevant psychological (NRP) factors"), exactly as personality varies with respect to a few dimensions (e.g., extraversion). The dimensions of neural personality need not correspond to existing cognitive constructs, and they must be discovered by examining brain activation across many tasks (more on this below).

Several points about Anderson's proposal are noteworthy. First, the focus on neural personalities instead of intrinsic functions is a radical change of heart for Anderson, who previously advocated characterizing regions' *workings*—roughly, their context-insensitive computational functions (Anderson, 2010). Second, it is not clear whether Anderson denies that brain regions have intrinsic functions or merely thinks the best strategy for cognitive neuroscientists is to characterize their neural personalities, while conceding that future efforts could identify their intrinsic functions. The anti-computationalist rhetoric in *After Phrenology* suggests the former, but more guarded

remarks support the latter. Third, Anderson mainly resists the call to revise brain ontology, focusing mostly on the brain structures—i.e., individual regions—that cognitive neuroscientists have traditionally studied. In this respect, *After Phrenology* is surprisingly conservative. Fourth, Anderson's focus on neural personalities implies that, in contrast to Poldrack's approach, the search for selective activation plays no role in cognitive ontology revision: A "central point of this book is not just that we don't *get* selectivity in the brain, but that *we don't need it*. *We can stop looking for it*" (2014, 141, emphasis in the original). Fifth, Anderson proposes to identify the dimensions of neural personalities (the NRP factors) in a strictly *bottom-up* manner: The proposal is to infer these new cognitive constructs from the brain's "behavior"—its activation patterns—across many tasks. In this respect, *After Phrenology* is surprisingly radical. Cognitive neuroscientists typically impose existing cognitive constructs onto the brain to interpret task-related activation. Instead, Anderson proposes using brain activation patterns across tasks to determine their psychological nature—what the tasks have in common and how they differ from a psychological point of view: "[O]ne can (…) use these data [i.e., the data from imaging experiments] to let the brain tell us something about these experiments—to reveal the underlying attributes of the task situation to which the brain differentially respond" (2014, 138).

How should researchers interpret NRP factors (the dimensions along which neural personalities vary) and neural personalities themselves? There are two ways of interpreting them: an *instrumentalist* or a *realist* interpretation. According to the instrumentalist interpretation, these dimensions (NRP factors) are just a way of

summarizing how similar the brain activation patterns elicited by the tasks under consideration are, and ascribing a neural personality to a brain area is just nothing more than a way of summarizing the data showing how this area is differentially active in a set of tasks. According to the realist interpretation, the dimensions of neural personality are *real* psychological constructs: That is, they can feature in causal explanations. *After Phrenology* is unclear about which of these two interpretations is correct, but Anderson appears to view NRP factors as explanatory and causal (2014, 151): "NRP factors should be understood as a region's disposition to help shape an organism's interaction with its environment, or to manage some aspect of the organism-environment relationship." These two interpretations of neural personalities should be familiar to readers acquainted with the history of psychology: Psychologists have long debated whether traits such as IQ or personality dimensions should be interpreted instrumentally or realistically.

Our main contention is that, just like other attempts at revising cognitive ontologies in a strictly bottom-up manner, Anderson's revolutionary endeavor to develop new cognitive constructs—the NRP factors and the neural personalities—can only be interpreted *instrumentally*, and that this is in tension with his goal of developing a new set of causally explanatory cognitive constructs. To characterize brain areas' dispositions, Anderson first appeals to the notion of a *functional fingerprint* developed by Passingham, Stephan, and Kotter (2002) (Anderson, 2014, section 4.2; Anderson, Kinnison, & Pessoa, 2013; Uddin, Kinnison, Pessoa, & Anderson, 2014). Identifying a region's functional fingerprint begins with categorizing the tasks in the fMRI literature on this area as recruiting one of several psychological processes. Anderson and colleagues typically use

a coarse-grained categorization scheme, distinguishing about 20 processes such as vision, attention, phonology, semantics, learning, or working memory. This allows them to represent quantitatively how often, according to a given literature, a given area is activated when one of these 20 processes is recruited by an experimental task, for instance how often articles studying working memory report activation in the dorsal anterior insula. The pattern of recruitment of a given area, given a particular set of fMRI articles and a categorization scheme, is its functional fingerprint. While, unsurprisingly, areas tend to be activated by many processes, their functional fingerprints vary. Importantly, a functional fingerprint is a mere *summary* of a data set: It does not explain why the area is activated the way it is.

Following Poldrack et al. (2009), Anderson (2014, Sections 4.3 and 4.4) proposes to use dimension reduction techniques (factor analysis, MDS, PCA, etc.) to identify a few dimensions explaining why an area has its functional fingerprint. Instead of merely summarizing the involvement of a given area in a set of tasks, as functional fingerprints do, neural personalities *explain* this involvement: They allow cognitive neuroscientists to claim that *because* an area has a given neural personality (its score is *i* on NRP factor *1*, *j* on NRP factor *2*, etc.), it is involved more in some tasks than others.

However, dimension reduction techniques are ill suited for discovering new cognitive constructs (Gould, 1996; Glymour, 2001). These statistical techniques project high-dimensional spaces onto spaces with fewer dimensions. On their own, the resulting dimensions cannot be interpreted realistically; they merely provide convenient ways of

summarizing high-dimensional data. Three main arguments support this deflationary understanding of dimension reduction techniques. First, the outcome of these techniques is *underdetermined*. A given set of vectors in a high-dimensional space can be projected onto different spaces with different dimensions. To highlight merely three issues, there are many non-equivalent dimension reduction techniques, the number of dimensions is typically arbitrarily chosen, and these dimensions can be oriented in different manners. None of the possible spaces should be interpreted realistically since it would be arbitrary to treat one of them as real to the detriment of the others. Second, just like causally-based correlations, *accidental* correlations can be projected onto a lower-dimensional space, resulting in meaningless dimensions (e.g., Gould, 1996, 280). Thus, that a high-dimensional space can be projected onto a lower-dimensional space does not justify interpreting the resulting dimensions realistically. Finally, the capacity of dimension reduction techniques such as factor analysis to identify causes has not been validated (Glymour, 2001, chapter 14). These three arguments bear on Anderson's project, exactly as they bear on IQ and personality research: On their own, dimension reduction techniques do not justify interpreting the dimensions of neural personalities realistically. Forgetting their limitations is committing the error of *reification*—viz., presuming that the abstract mathematical entities uncovered by dimension reduction analyses correspond to real psychological entities.

Naturally, the products of dimension reduction techniques can sometimes be interpreted realistically instead of as mere instruments for summarizing high dimensional data. To do so scientists need to bring their broader empirical knowledge to bear on the interpretation

of the dimensions of the lower-dimensional space. In the present context, this means that a purely bottom-up approach to cognitive ontology revision is unlikely to succeed: Some other information beyond the activation of brain areas across a range of tasks and their dimension reduction is needed to interpret the resulting dimensions realistically. Perhaps it is also worth noting that establishing the predictive validity of neural personalities does not justify understanding them realistically.

Anderson's approach to cognitive ontology revision is not the only one to fall prey to this *reification objection*; in fact, we speculate that in general purely bottom-up cognitive ontology revisions commit the error of reification (e.g., Poldrack et al., 2009). Such approaches must reduce the very high-dimensional space defined by the number of voxels considered in order to identify cognitive constructs defined solely by brain activation patterns. Doing so probably requires using techniques whose product cannot be interpreted realistically. In our opinion, the reification objection reveals a fundamental shortcoming of bottom-up cognitive ontology revision.

References

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences, 33*, 245-261.

Anderson, M. L. (2014). *After phrenology: Neural reuse and the interactive brain*. Cambridge, MA: MIT Press.

Anderson, M. L., Kinnison, J., & Pessoa, L. (2013). Describing functional diversity of brain regions and brain networks. *Neuroimage, 73*, 50-58.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT Press.

Glymour, C., & Hanson, C. (Forthcoming). Reverse inference in neuropsychology. *The British Journal for the Philosophy of Science*.

Gould, S. J. (1996). *The mismeasure of man*. New York: WW Norton & Company.

Klein, C. (2012). Cognitive ontology and region-versus network-oriented analyses. *Philosophy of Science, 79*, 952-960.

Lenartowicz, A., Kalar, D. J., Congdon, E., & Poldrack, R. A. (2010). Towards an ontology of cognitive control. *Topics in Cognitive Science, 2*, 678-692.

Machery, E. (2014). In defense of reverse inference. *British Journal for Philosophy of Science, 65*, 251-267.

McCaffrey, J. (Forthcoming). The brain's heterogeneous functional landscape. *Philosophy of Science*.

Passingham, R. E., Stephan, K. E., & Kötter, R. (2002). The anatomical basis of functional localization in the cortex. *Nature Reviews Neuroscience, 3*, 606-616.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data?. *Trends in cognitive sciences, 10*, 59-63.

Poldrack, R. A. (2010). Mapping mental function to brain structure: how can cognitive neuroimaging succeed? *Perspectives on Psychological Science, 5*, 753-761.

Poldrack, R. A., Halchenko, Y. O., & Hanson, S. J. (2009). Decoding the large-scale structure of brain function by classifying mental states across individuals. *Psychological Science, 20*, 1364-1372.

Price, C. J., & Friston, K. J. (2005). Functional ontologies for cognition: the systematic definition of structure and function. *Cognitive Neuropsychology*, *22*, 262-275.

Uddin, L. Q., Kinnison, J., Pessoa, L., & Anderson, M. L. (2014). Beyond the tripartite cognition–emotion–interoception model of the human insular cortex. *Journal of cognitive neuroscience*, *26*, 16-27.