Memory as a cognitive kind:
Brains, remembering dyads, and exograms

Samuli Pöyhönen*

Final draft, Oct 2015

Published in C. Kendig (2016, ed.), *Natural Kinds and Classification in Scientific Practice.* London: Pickering & Chatto, pp. 145–156

Abstract

Theories of natural kinds can be seen to face a twofold task: First, they should provide an ontological account of what kinds of (fundamental) things there are, what exists. The second task is an epistemological one, accounting for the inductive reliability of acceptable scientific concepts. In this chapter I examine whether concepts and categories used in the cognitive sciences should be understood as natural kinds. By using examples from human memory research to illustrate my argument, I critically examine some of the main contenders for a theory of natural kinds. I show that when applied to complex target domains – such as human cognition – both essentialist theories and more liberal accounts of natural kindhood (such as HPC theory) fail to simultaneously satisfy the ontological and epistemological desiderata for a theory of natural kinds. I argue, however, that natural kindhood of a category in a metaphysical sense is not necessary for its inductive reliability, and that HPC theory provides an analytically useful account of the grounds of category-based inductive inference.

1 Introduction

Philosophers often think that scientific concepts should refer to natural kinds. It has been suggested that this requirement does not only hold in the natural sciences, but applies also in the social and behavioral sciences, including the cognitive sciences, because inductive success also in the special sciences requires that the used categories must somehow track the objective structures of reality.[1] In the philosophy of psychology, debates concerning cognitive natural kinds have recently surfaced, for example, in discussions on scientific eliminativism and on extended cognition: Scientific eliminativism presupposes a view of concepts in the psychological sciences as natural kinds, and in the extended cognition debate, especially defenders of intracranialist positions have alluded to properties of natural kinds in their arguments against cognitive extension.[2]

This chapter is an attempt to uncover presuppositions about natural kinds underlying such debates, and to disentangle different argumentative roles that appeals to natural kindhood serve. Using an example drawn from memory research in psychology to illustrate my point, I argue that parallel research programs in the cognitive sciences often rely on slightly different conceptualizations of the phenomenon of interest (e.g., human memory), and it is not obvious which one (or ones) of the alternative classifications deserves a natural kind status. I suggest that there is a general trade-off between two main desiderata of theories of natural kinds: giving an adequate answer to both the ontological question ("what kinds of things are there?") and the epistemological question about natural kinds ("how does natural kindhood justify inductive inferences?"). However, I suggest that reliable category-based inductive inference does not require natural kindhood in any strong metaphysical sense, and that classificatory pluralism reflects indispensable division of cognitive labor in research on causally complex target domains.[3]

2 Where is memory? Different memory-kinds in psychological research

Let us approach the question of cognitive natural kinds by looking at an example. The scientific concept of MEMORY has been studied by several philosophers of science, and

the important episodes in the historical development of the notion are well-documented.[4] Although space does not allow a detailed rehearsal of this history, it can be said that the development of the notion of memory in the cognitive sciences seems like a series of splittings or decomposition: Up to the mid-20th century, long-term memory was considered a unitary phenomenon. However, early dissociation findings in cognitive neuropsychology (the well-known case of patient H.M) suggested that different kinds of memories are sensitive to damage in different parts of the brain, and that skill memory is a distinct type of long-term memory, sustained partly by its proprietary neural structures and processes. In later research, further dissociations have also been obtained between, for example, episodic and semantic memory systems. Based on these findings, several ways of classifying kinds of memory have emerged, but there is still no agreement about the correct way of dividing human memory into subkinds or subsystems.[5]

However, a shared assumption in memory research in cognitive neuroscience is that memory is a cognitive capacity supported by neural processes inside the human skull, with many of the relevant neural mechanisms residing in the hippocampal formation and prefrontal cortex. These intracranial memory capacities share some characteristic properties like the law of effect, primacy effect, recency effect, chunking effect, and others. Therefore, irrespective of how many subkinds of memory there ultimately turn out to be, for the purposes of this chapter, we can think of the cognitive neuroscience notion of memory as a putative cognitive kind with a well-identifiable neural basis and a cluster of characteristic kind-properties.

Now, consider a different approach. Research on *transactive memory* takes the human dyad as its fundamental unit of analysis. The approach is motivated by the insight that remembering often occurs in social contexts. Closely related people such as married couples often use each other as memory aids, and rely on collaborative recall strategies.[6] For example, when asked about personal events that took place a long time ago, couples can engage in cross-cuing, where they take turns in recalling facts about the event until one of them can retrieve the needed detail. Hence, in transactive memory research, memory processes are understood as consisting of both individual-level cognitive processes, as well as the social interaction processes between the individuals. Operations

of encoding, storage and retrieval are understood as being implemented in a socially distributed system, not in the brain of a sole individual.

More radically still, in his work on the evolution of human cognition, Merlin Donald has developed a theory of the contemporary human mind as being extensively dependent on external memory traces, exograms.[7] According to Donald, in the late Upper Paleolithic era, changes to the human prefrontal cortex allowed for increased plasticity of cognition and made possible the use of extracranial memory stores such as clay tables, papyri, printed books, government archives, and electronic data banks. This, in turn, allowed permanent storage and iterative revisal of symbols and created the basis for cumulative culture. The theory does not only claim that modern humans can successfully employ external memory stores, but that using exograms changes our internal cognitive architecture. That is, the intracranial and extracranial parts of the memory process are connected by feedback-loops and they form a unified cognitive system.

Both transactive memory research and Donald's theory suggest that the relevant kind of cognitive system to be studied encompasses elements beyond a single human brain, and that not all of its properties are sustained by intracranial factors. For the purposes of the discussion below, I identify three distinct notions of memory as a putative natural kind emerging from the research programs above: *intracranialist memory*, *transactive memory,* and *exogram memory*. Now, which of the aforementioned concepts picks out a cognitive natural kind? Or could it be that none – or all – of them are acceptable natural kinds?

While the intracranialist notion fits best with pre-theoretic intuitions and certainly represents the mainstream approach in memory research, the alternative conceptualizations have theoretical virtues of their own. As argued by the proponents of transactive memory and exograms, taking seriously the idea of the distributed nature of memory can help to explain several aspects of remembering not satisfactorily answered by intracranialist approaches.[8] All three memory concepts are bona fide scientific categories in the sense that they play a part in successful research practices in the social and behavioral sciences, and they appear to be sustained by real causal processes.[9] I now

turn to a discussion of theories of natural kinds in order to uncover possible conceptual reasons for preferring one conceptualization over another.

3 Natural kinds in the cognitive sciences

Although discussions on natural kinds in philosophy have been motivated by many different concerns, the nature of scientific categories and the principles of concept formation have been a central issue underlying the debate since Mill's seminal work on real kinds.[10] This is the shared underlying theme also between the many of the contemporary discussions on cognitive kinds. Both in the debate on scientific eliminativism and the one on extended cognition, the underlying question is the same: what are the correct units of analysis in terms of which to conduct research in the mind sciences? Are EMOTION, MEMORY or CONCEPT adequate psychological categories? Are human cognitive systems really contained by the organismic skin-bag? This *units-of-analysis problem* is of major significance for two reasons:

(1) It is generally taken to be a central task of science to tell us what kinds of things there are, what exists. Scientific kind-concepts reflect our ontological commitments, and trying to align our concepts with reality's natural kind structure is an expression of this aim. Giving an account of the nature of such fundamental furniture of reality is what I call the *ontological desideratum* for theories of natural kinds.

(2) As suggested by Nelson Goodman's famous grue problem, the way we define our scientific concepts is crucial for guaranteeing the reliability of epistemic practices in science: prediction, explanation and manipulation of phenomena. A theory of natural kinds ought to show how natural kinds differ from non-natural ones, and how this difference supports projectibility judgments. Although the philosophical explanations of the inductive reliability of natural-kind concepts often refer to laws of nature, causal powers, and mechanisms, this second question is first and foremost motivated by epistemological concerns. I call it the *epistemological desideratum* for theories of natural kinds.[11]

These desiderata articulate the reasons why we are interested in natural kindhood to begin with. In our example, why it matters whether intracranialist memory, transactive

memory, and exogram memory are natural kinds of things is that a philosophical account of natural kinds supplies us with criteria for determining which concepts (1) point to really existing categories of phenomena with (2) reliably projectable clusters of kind-properties.

I now try to show that different theories of natural kinds put emphasis differently on the ontological and epistemological desideratum. Moreover, I suggest that there is a general tension between these two desiderata.

First, consider austere theories of natural kinds, according to which natural kinds differ from less fundamental categories because they participate in laws of nature, or because they have kind-essences consisting of irreducible intrinsic dispositional properties.[12] Typical examples of such kinds come mostly from physics, and these theories typically articulate a metaphysical picture of how the kinds, their essential properties, and the regularities pertaining to them generally hang together. For ease of exposition, let us call such theories fundamentalist ones. They are often primarily motivated by the concern of uncovering the nature of the most fundamental categories of reality, and they typically provide at least coherent and intuitively plausible answers to the ontological task faced by theories of natural kinds.

According to fundamentalist theories, it is obvious that none of the memory-kinds discussed above count as natural kinds: There are no universal and exceptionless laws about cognitive kinds, nor do they (qua cognitive kinds) possess irreducible intrinsic essential dispositions. However, while these sparse conceptions of natural kinds might appear to be an ontologically prudent option, excluding most special science kinds from the class of natural kinds implies that these theories do not even begin to explain the successful epistemic functioning of kind terms in the special sciences.

One might argue that fundamentalist theories specifically do address the question of why inductive inference *in general* can be reliable (for example, by giving an account of the truth-makers of laws of nature). However, this does not yet meet the epistemological desideratum formulated above: As is illustrated by the debate on scientific eliminativism, theories of natural kinds are employed in arbitrating disputes regarding the scientific

status of particular cognitive kinds.[13] In order to contribute to such debates, a fundamentalist theory of natural kinds should either be complemented by an account of special-science kinds in terms of fundamental kinds (e.g., micro-reduction), or a entirely separate story about the epistemic reliability of special-science kinds. In the absence of such an account, due to their limited scope, fundamentalist theories of natural kinds – even if right about some fundamental categories – fail to meet the epistemic desideratum.

One way to expand the scope of essentialist theories is to relax the assumption about the irreducibility of the essential properties.[14] That is, while demanding that natural kinds must be unified by shared causal powers, one can include chemical compounds, perhaps even elms and beeches among natural kinds by arguing that causal powers need not be irreducible properties of kind-members but that they can brought about by the shared internal structure of the kind-members.[15] Such non-fundamentalist essentialism seems to imply the *prima facie* sensible choice between the three memory kinds of our example, according to which intracranial memory is the natural kind, the other conceptualizations being somehow derivative. Here the suggestion would presumably be that neural structures in the relevant brain areas of individuals constitute the internal microstructure that explains the projectibility of relevant properties of the kind MEMORY.

However, on a closer look, it turns out that non-fundamentalist essentialism does not apply to cognitive kinds either. First, in the life sciences, due to geno- and phenotypic variation and neural plasticity, the members of a kind never share an identical microstructure, and it is far from obvious how one should cash out the idea of sameness of structure required for kind membership. Secondly, as the examples discussed in the literature on extended cognition suggest, internal structure alone does not generally explain all the scientifically relevant properties of cognitive kinds.[16] Unlike physical and chemical kinds, which often tend to have relatively robust causal powers that manifest in the same way across many circumstances, cognitive kinds are evolutionarily fine-tuned to certain environmental conditions (see section 4). Hence, for many cognitive kinds, it is not the internal structure alone but the coupled system of neural structures together with the relevant parts of the environment of the organism that explain many of the observable properties of the phenomenon. This suggests that the core explanatory properties even for

the intracranialist memory include relational, non-intrinsic properties.

Therefore, it seems that for a theory of natural kinds in the cognitive sciences, both the irreducibility of essences and their intrinsicality must go. Several liberal theories of natural kinds do indeed drop these assumptions. For example, Paul Griffiths characterizes the essence of a kind as *"any theoretical structure that accounts for the projectability of a category.*"[17] However, as I argue in more detail below, such liberal theories of natural kinds no longer single out the intracranialist memory as the sole natural kind, but apply also to transcranial and exogram memory. This takes us to the initially strange idea of considering all three (or at least more than one) memory kinds as natural kinds.

Currently the most popular alternative to essence- or law-based accounts of natural kinds is Richard Boyd's theory of natural kinds as homeostatic property clusters (HPC theory).[18] In the philosophy of science, Boyd's theory has nearly achieved the status of the new received view of natural kinds.[19] According to the theory, a natural kind consists of (i) a cluster of typical properties, held together by (ii) a homeostatic causal mechanism. Boyd's own example of an HPC natural kind is a biological species: The cluster of shared morphological, physiological and behavioral properties of the species is maintained by the circulation of genetic material among its members. Interbreeding functions as the homeostatic mechanism sustaining the cluster.

Like essentialist theories, HPC theory retains a distinction between the observable properties of a kind and its explanatory core, but this core (i.e., the homeostatic mechanism) need not be either irreducible or intrinsic.[20] In fact, it seems that any organized causal pattern or structure sustaining a stable property cluster and maintaining its projectibility can function as a kind's homeostatic mechanism. This suggests that natural kinds are not scarce at all, and Boyd's theory attempts to strike a delicate balance between causal realism and classificatory pluralism:

> There are not kinds which are natural *simpliciter* but instead kinds that are natural with respect to the inferential architectures of particular disciplinary matrices. Any talk of natural kinds, properly understood, involves (perhaps tacit) reference to or quantification over disciplinary matrices. [21]

According to Boyd's somewhat opaque formulation, natural kind terms are a means for creating an accommodation between the causal structure of reality and the epistemic aims of a scientific discipline. Applied to the memory example, HPC theory suggests that the different ways of capturing the processes behind human memory phenomena can all give rise to natural kinds, as long as each of the resulting classifications has a robust cluster of properties sustained by a homeostatic mechanism. Empirical research suggests that this might be the case: As pointed out in section 2, the intracranialist concept based on the neural mechanisms sustains a property cluster of its own, and transcranial and exogram concepts, given their more complex mechanisms including also social components, are fit for the inferential needs of the other research agendas. Unlike arbitrary or purely stipulated classes of things, the unity of all three memory kinds is causal, not conventional, and all three categories support reliable prediction and manipulation of phenomena.

As the quoted passage above suggests, Boyd's theory is clearly an epistemology-first approach. I argue in the next section that the theory provides a detailed and applicable account of category-based inductive inferences and an elaborate picture of the reasoning behind projectibility judgments. However, I also argue that it faces serious difficulties with the ontological task of theories of natural kinds. Firstly, the classificatory pluralism of Boyd's theory is hard to square with many ontological intuitions about natural kinds: If all causally sustained groupings qualify as natural kinds, does this not lead to an explosion in the number of natural kinds? Are all different scientific ways of classifying the same domain of phenomena natural-kind classification systems? And how robust does a cluster of properties need to be for the kind to count as a natural one? For example, are MARKET, REVOLUTION, EYE, ENZYME, and STORM natural kinds? Are all three memory kinds really ontologically on the same footing? Understanding the naturalness of a classification as mere causal groundedness seems to water down the notion, as it no longer sustains the contrast to partly mind-dependent kinds, social kinds, and artifact kinds, all of which can be causally sustained categories as well. Frank Jackson expresses the worry of inflating our ontology by saying that metaphysics should not be concerned with any old shopping list of things that there are. Instead, when

concerned with ontology, we should seek a comprehensive account of what there is in terms of a restricted set of somehow fundamental notions.[22] HPC theory, by contrast, leads to a rich rainforest ontology. In the next section I show that even putting these scarcist intuitions on the side, there are serious worries with HPC theory understood as a metaphysical account of natural kinds, especially in causally and structurally messy domains such as human cognition.

4 Classification and complexity

Our choice problem between the three notions of memory appears to have led to an impasse: Adequate theories of natural kinds are presupposed to meet both the ontological and the epistemological desideratum for natural kinds, but essentialist theories fail to shed light on the epistemic reliability of cognitive kinds, and HPC theory fails as an ontological account of the fundamental furniture of reality. I now argue, however, that natural kindhood in an ontological sense is not a necessary condition for reliable inductive inference, unlike often assumed. In particular, I show how the mechanisms-based HPC theory provides a plausible account of inductive inference even in causally complex and non-modular domains, where it is not often even clear what a theory of natural kinds should say about the location of nature's joints. Hence, this section is a qualified defense of the HPC theory as an account of classification in the cognitive domain – not necessarily as a full-blown theory of natural kinds, but as a theory of the choice of units of analysis for research and of the foundations of category-based induction. At least for most scientific purposes, that is all we should hope for.

Many of the problems of HPC theory as an ontological account of natural kinds stem from the notion of mechanism central to the theory. First, as critics have pointed out, although mechanistic language abounds in the life sciences, it is not clear that all natural kinds (e.g., elementary particles) are united by mechanisms. Second, there's a whiff of circularity in the mechanistic approach to kinds. One way to raise the issue is the following: HPC kinds are individuated by their homeostatic mechanisms. Not, however, by token mechanisms but mechanism types – kinds of mechanisms. How should we, then, individuate kinds of mechanisms?

In the philosophy of science literature on mechanisms and causal explanation, it is widely recognized that mechanisms are always mechanisms *for* something.[23] It is the *explanandum* at hand that determines which causal variables to include in a mechanism, and which ones remain as background variables. Consequently, mechanism identity and demarcation of its boundaries always rely on judgments of causal and explanatory relevance. It could be said that mechanisms are not simply "out there," but they are parts of the causal structure of reality recruited for the purpose of explaining a particular target phenomenon, effect, psychological capacity etc.[24]

In the case of homeostatic mechanisms of natural kinds, such an analysis seems to put the cart before the horse: The "essential" property of natural kind is its homeostatic mechanism, but identifying mechanisms presupposes that the explanandum phenomenon is known beforehand. However, isn't that exactly what a theory of natural kinds should provide?

The objection might not be such a serious one in domains where mechanism boundaries are clear due to the modularity of causal structure: i.e. when it is possible to find causal bottlenecks, low-bandwidth interfaces, between relatively independent causal structures, and where the resulting mechanisms for kinds sustain clearly distinct phenomena with mutually exclusive property clusters.[25] Many traditional examples of natural kinds (such as chemical compounds or metals) appear to have such a structure: they have relatively stable clusters of properties upheld by straightforwardly demarcated mechanism (e.g. the interactions constituting the metallic lattice). However, the nature of the evolutionary design process suggests that in the cognitive domain, non-modularity is common; natural selection leads the evolved solutions to the environment challenges faced by organisms to often employ several kinds of hacks: (1) Making use of reliable environment properties in task completion and (2) recycling old biological components for novel purposes (exaptation).[26] The resulting designs often involve intricate feedback-loops between the organism and its environment, making it difficult to tell where the boundaries of the (kind of) system lie.

Likewise, in our current case example, the causal mechanisms explaining human memory

capacities extend to the environment, and the location of the boundary between the cognitive system and its environment is not obvious. Different ways of defining the concept of memory, and the respective ways of identifying its underlying causal basis correspond to different explanatory aims. Whereas the intracranialist approach focuses on the neural processes underlying remembering, transactive and exogram memory kinds illuminate the role of extracranial processes and resources in naturalistic remembering contexts. Non-modularity of the causal structures makes it necessary to resort to explanatory aims in demarcating mechanism boundaries.

This explanandum-dependence of mechanism identification dovetails nicely with Boyd's idea of natural kinds being relative to the epistemic aims of a particular discipline. However, a lot turns on the nature and number of such epistemic aims. For example, do the three different ways of classifying memory only reflect human interests, or is the set of acceptable epistemic aims constrained by some facts of the matter about mind-independent reality, which would then rule out some other possible memory concepts? Unless there are such constraints, HPC theory is driven from classificatory pluralism towards radical conventionalism – a Lockean predicament, so to say – where boundaries of kinds are only the work of human understanding.

I have now argued that, at least in non-modular domains, HPC theory avoids the circularity problem only at the price of introducing a rather strong form of conventionalism, and this raises serious worries about its role as an ontological account of *natural* kinds. However, the main reason for the popularity of HPC theory has been its usefulness as a theory of the epistemic functioning of scientific concepts and of the principles of conceptual change. The worry about conventionalism raises no conceptual hurdles for HPC theory as an account of category-based inductive and explanatory inference. Even in cases where one is left with multiple possible ways of demarcating mechanism boundaries, the core picture of category-based induction provided by HPC theory holds: Kind membership as such does not ultimately do any inferential work, but inferences to yet unknown counterfactual situations involving kind members are based on knowledge of how the underlying causal structures bring about the projectable properties of the kind. This *mechanistic extrapolation* provides a powerful inferential strategy for

the population-thinking contexts in the social and behavioral sciences. According to this view of category-based induction, counterfactual questions of whether/how a particular kind-property x gets manifested under conditions y is answered by examining the functioning of the mechanistic structures governing the appearance of the property in question.[27] For example, detailed knowledge of the neural mechanisms underlying spatial memory allows fine-grained what-if inferences of how human navigational capacities would be affected by hippocampal damage, and processes underlying transactive memory specify when and how the remembering capacities of a social dyad transcend those of isolated individuals.

Herein lies the strength of the HPC theory in comparison with empiricist and more minimalist approaches to kinds which portray them only in terms of a set of prototypical projectible properties:[28] Knowledge of category membership combined with background information about the mechanistically organized causal factors sustaining the kind, as well as knowledge concerning the invariance of such causal structures and their breakdown conditions facilitates inferences not possible with knowledge about only the observable property cluster itself.

As I have suggested elsewhere, this picture of category-based induction can be developed further by connecting the work on HPC theory to recent discussions in the philosophy of science on causality, explanation, and explanatory power.[29] There being multiple ways to classify a target domain can be understood as a reflection of a division of cognitive labor between scientific fields or research agendas. Different categorizations at various levels of abstraction result in different *profiles of explanatory and inferential power*, and it is only through coordination among a set of parallel approaches that the scientific community can come up with comprehensive accounts of complex targets. Accommodation between a causally messy world and finite cognizers is only to be had in a piecemeal fashion. Such a division of labor between different conceptual perspectives also appears as the most natural explanation for the plurality of memory concepts, and there seems to be no epistemic payoff from reducing that classificatory diversity (unlike often presupposed in theories of natural kinds).

5 Conclusion

I have argued that in the case of many cognitive kinds, the main contenders for a theory of natural kinds fail to simultaneously meet the ontological and epistemological desiderata for natural kinds. The more traditional law- or essence-based theories do not apply to cognitive kinds, and HPC theory fails to meet the ontological desideratum. As a way out, I suggest that – contrary to the presupposition in many discussions of natural kinds – natural kindhood in a strong ontological sense is not a necessary condition for the inductive reliability of scientific concepts. In particular, I argued that HPC theory provides an appealing account of the grounds of category-based inductive inference in the life sciences. Unlike the often outdated views of explanation, lawlike generalizations and inductive inference underlying many other theories of scientific concepts and kinds, the more modest commitment to causal realism assumed by HPC theory suffices to justify category-based induction in scientific research. Hence, if HPC kinds fail to count as natural, then science does not always need natural kinds. However, this is not to say that fundamental natural kinds, wherever applicable, would not be sufficient for reliable induction.

I suggest that loosening the link between a theory of scientific concepts and metaphysically-oriented theories of natural kinds can direct attention to new fruitful questions about cognitive kinds and systems: Rather than trying to find the unique correct way of classifying the cognitive domain, philosophical methods can be used for keeping track of the differences between classificatory commitments in neighboring research fields in the mind sciences. Making explicit the differences and overlaps between the kind-terms used in parallel research agendas can be useful both for avoiding misunderstandings in interfield communication and for facilitating theoretical integration in research.

---

* Thanks to Catherine Kendig, Marion Godman, and the participants of the Kinds workshop at the University of Tampere and the TINT Brown Bag seminar at the University of Helsinki for their helpful comments on earlier versions of this chapter.

[1] R. Boyd, 'Realism, anti-foundationalism and the enthusiasm for natural kinds', *Philosophical Studies*, 61 (1991), pp. 127–148; P. Griffiths, *What Emotions Really Are* (Chicago: University of Chicago Press, 1998); R. Samuels, 'Delusions as a natural kind', in M. Broome & L. Bortolotti (eds), *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives* (Oxford: Oxford University Press, 2009), pp. 49–82; P. D. Magnus, Scientific Enquiry and Natural Kinds (New York: Palgrave Macmillan, 2012): M. Khalidi, *Natural Categories and Human Kinds* (Cambridge: Cambridge University Press, 2013).

[2] For scientific eliminativism, see Griffiths, *What Emotions Really Are*; C. Craver, 'Dissociable realization and kind splitting', *Philosophy of Science*, 71 (2004), pp. 960–971; E. Machery, *Doing Without Concepts* (New York: Oxford University Press, 2009). In A. Clark and D. Chalmers, 'The extended mind,' *Analysis*, 58 (1998), pp.10-23, and in the debate that ensued, one of the central points of disagreements has been whether categories put forward by extended cognition theorists can be natural kinds.

[3] By *category-based induction* I mean inductive inference, where knowledge of categories or kinds is used to infer about the properties of yet unobserved members of the kind or related categories. For the origin of the phrase, see D. Osherson et al., 'Category-based induction,' *Psychological Review*, 97:2 (1990), pp. 185–200.

[4] For philosophical accounts, see for example R. Bechtel*, Mental Mechanisms* (New York: Routledge, 2008); Craver, 'Dissociable realization and kind splitting.'

[5] See L. Squire, 'Memory systems of the brain: A brief history and current perspective,' *Neurobiology of Learning and Memory*, 82 (2004), pp. 171–177; E. Tulving, 'Study of memory: processes and systems,' in J. Foster and M. Jelicic (eds), *Memory: Systems, Process, or Function* (Oxford: Oxford University Press, 1999), pp. 11–30.

[6] D. Wegner, T. Giuliano, and P. Hertel, 'Cognitive interdependence in close relationships,' in W. Ickes (ed), *Compatible and Incompatible Relationships* (New York: Springer-Verlag, 1985), pp. 253–276; C. Harris et al., 'We remember, we forget: collaborative remembering in older couples,' *Discourse Processes*, 48 (2011), pp. 267–

303.

[7] M. Donald, *Origins of the Modern Mind* (Cambridge: Harvard University Press, 1991), *A Mind so Rare* (New York: Norton, 2002).

[8] For observations about the theoretical limitations of intracranialist laboratory approaches to memory research, see U. Neisser, 'Memory: what are the important questions,' in Neisser, *Memory Observed: Remembering in Natural Contexts* (San Francisco: W. H. Freeman, 1981). Neisser laments: "A hundred years of psychological study on memory has produced hardly any results that would have relevance for real memory phenomena outside the laboratory."

[9] For a review of transactive memory research, see Y. Ren and L Argote, 'Transactive memory systems 1985–2010,' *The Academy of Management Annals*, 5:1 (2011), pp. 189–229. B. Sparrow et al., 'Google effects on memory,' *Science 333*, (2011), pp. 776–778 is a recent example of empirical exogram memory research (misleadingly titled as transactive).

[10] J. S. Mill, *A System of Logic* (Honolulu, Hawaii: University Press of the Pacific 2002, original 1891).

[11] A similar distinction between "metaphysics first" and an "epistemology-oriented tradition" of natural kinds has been made by T. Reydon, 'How to fix kind membership: A problem for HPC theory and a solution,' *Philosophy of Science*, 76 (2009), pp. 724–736.

[12] See for example, A. Rosenberg 2005, 'Lessons from biology for Philosophy of the Human Sciences,' *Philosophy of the Social Sciences*, 35:3 (2005), pp. 3–19; B. Ellis, *Scientific Essentialism* (Cambridge: Cambridge University Press, 2001).

[13] For an example of such a debate, see the exchange between the target article and the peer commentary in E. Machery, 'Precis of Doing Without Concepts,' *Behavioral and Brain Sciences*, 33 (2010), pp. 195–244.

[14] Intrinsicness has proved hard to analyze. In the following, I assume only that properties pertaining to the internal constitution of a system are among its intrinsic properties.

[15] See for example, T.E Wilkerson, *Natural Kinds* (Aldershot: Avebury, 1995).

[16] See A. Clark*, Supersizing the Mind* (Oxford: Oxford University Press, 2008).

[17] P. Griffiths, *What Emotions Really Are*, p.188.

[18] R. Boyd, 'Realism, anti-foundationalism and the enthusiasm for natural kinds'; 'Realism, natural kinds and philosophical methods,' in H. Beebee and N. Sabbarton-Leary (eds), *The Semantics and Metaphysics of Natural Kinds* (New York: Routledge, 2010), pp. 212–234.

[19] See R. Samuels and M. Ferreira 2010, 'Why don't concepts constitute a natural kind?' *The Behavioral and Brain Sciences*, 33 (2010), pp. 222–223; M. Ereshefsky and T. Reydon, 'Scientific kinds,' *Philosophical Studies* (forthcoming);

[20] There are now many variations of Boyd's original account, see references in R. Wilson, M. Barker and I. Brigandt, 'When traditional essentialism fails,' *Philosophical Topics*, 35 (2007), pp. 189–215.

[21] R. Boyd, 'Realism, natural kinds and philosophical methods,' p. 217.

[22] F. Jackson, *From Metaphysics to Ethics* (Oxford: Clarendon Press, 1998), p.4.

[23] See, for example, S. Glennan, 'Rethinking mechanistic explanation,' *Philosophy of Science*, 69:3 (2002), pp. S342–S353; and C. Craver, 'Mechanisms and natural kinds,' *Philosophical Psychology*, 22 (2009), 575–594. Craver makes this point about HPC theory in particular.

[24] S. Pöyhönen, 'Explanatory power of extended cognition,' *Philosophical Psychology*, 27:5 (2014), pp. 735–759.

[25] For modularity and the related notion of near-decomposability, see H. Simon, 'The architecture of complexity,' *Proceedings of the American Philosophical Society,* 106 (1962), pp. 476–482.

[26] J. Kuorikoski and S. Pöyhönen, 'Understanding non-modular functionality. Lessons from genetic algorithms," *Philosophy of Science,* 80:5 (2013), pp. 637–649.

[27] For a comprehensive account of mechanistic extrapolation in the life sciences, see D. Steel, *Across the Boundaries: Extrapolation in Biology and Social Science* (Oxford: Oxford University Press, 2008).

[28] For example, S. Häggqvist, 'Kinds, projectibility and explanation,' *Croatian Journal of Philosophy*, 13 (2005), pp. 71–87; M. Slater, 'Natural kindness,' forthcoming in the *British Journal for the Philosophy of Science.*

[29] S. Pöyhönen, 'Explanatory power of extended cognition.'