Explanatory power of extended cognition

Forthcoming in *Philosophical Psychology*

Samuli Pöyhönen*

samuli.poyhonen@helsinki.fi

Centre of Excellence in the Philosophy of the Social Sciences (TINT),
University of Helsinki

Social and Moral Philosophy PL 24,
00014 University of Helsinki

+358-40-7281096

Abstract

I argue that examining the explanatory power of the hypothesis of extended cognition (HEC) offers a fruitful approach to the problem of cognitive system demarcation. Although in the discussions on HEC it has become common to refer to considerations of explanatory power as a means for assessing the plausibility of the extended cognition approach, to date no satisfying account of explanatory power has been presented in the literature. I suggest that the currently most prominent theory of explanation in the special sciences, James Woodward's contrastive-counterfactual theory, and an account of explanatory virtues building on that theory can be used to develop a systematic picture of cognitive system demarcation in the psychological sciences. A major difference between my differential influence (DI) account and most other theories of cognitive extension is the cognitive systems pluralism implied by my approach. By examining the explanatory power of competing traditions in psychological memory research, I conclude that internalist and externalist classificatory strategies are characterized by different profiles of explanatory virtues and should often be considered as complementary rather than competing approaches. This suggests a deflationary interpretation of HEC.

Keywords: Extended Cognition; Explanatory Power; Mechanism; Pluralism

---

* I wish to thank the participants in the PPIG seminar at the University of Edinburgh (2010), the philosophy of science research seminar at the University of Helsinki (2012) and the philosophy colloquium at the Berlin School of Mind and Brain (2012) for their useful comments on earlier versions of this paper. I am also grateful to Cameron Buckner for his valuable comments and suggestions on an earlier draft.

1 Introduction

Whether human cognition is partly constituted by components residing outside the skull is the question at the heart of an extensive debate in the philosophy of psychology. I argue that examining the explanatory power of the hypothesis of extended cognition (HEC) offers a fruitful approach to the problem of cognitive system demarcation. A similar insight underlies several recent contributions to the debates on HEC, as many authors now agree that the plausibility of extended cognition depends on the epistemic value of the taxonomies generated by the suggested approach (Rupert, 2004; Clark, 2007; Barker, 2010; Sprevak, 2010). However, although both the advocates and the critics of the hypothesis have employed the notion of explanatory power in their arguments, most of these accounts have not been based on systematic theories of scientific explanation (Clark & Chalmers, 1998; Adams & Aizawa, 2001, 2008; Rupert, 2004; Wilson, 2004; Clark, 2007; Sutton, 2010b). The goal of this paper is to fill this lacuna in the literature. By building on conceptual tools from the philosophy of science, I put forward a *differential influence criterion*, which together with comparative assessment of the explanatory virtues of internalist and externalist classificatory strategies provides a scientifically plausible way of determining the boundaries of cognitive systems.

This explanationist approach to the problem of cognitive extension is not meant to be comprehensive in the sense that it would try to engage with the majority of the influential views presented in the literature. Instead, I try to isolate and elaborate a strand of thinking about extended cognition relevant to the research in the mind sciences. By working out the implications of adopting an explanationist approach, I argue that although it leads to perhaps *prima facie* counter-intuitive cognitive systems pluralism, it does provide a conceptually coherent and practically workable approach to demarcating cognitive systems. As a further difference to many of the existing accounts, my pluralistic solution to the problem of cognitive system demarcation has a deflationary tone. I argue that HEC itself should not be understood as a truth-valued hypothesis but rather as a strategy

for producing scientific classifications. Cognitive extension is thus not a single ontological problem but a collection of more local questions of scientific classification and concept formation.

I suggest that this reorientation in understanding HEC brings extended cognition back to its roots, in a sense. Philosophical discussions on extended cognition were originally inspired by the ingenious but often overlooked research traditions in scientific psychology, which emphasize the explanatory importance of external factors not visible to the mainstream intracranialist approaches (cf. Hutchins, 2010). Despite the conceptual rigor of the current philosophical debates on HEC, it is hard to avoid the feeling that the close connection to questions rising from empirical research has often been lost. A central aim of the theoretical framework developed in the current paper is to try to capture some of the important implications that the philosophical discussions on extended cognition could have for research practices in the psychological sciences.

The two first sections of the paper outline and argue for a naturalistic perspective on HEC that my solution to the problem of cognitive extension builds on. In section 2, I suggest that the hypothesis of extended cognition should be distinguished from a nearby hypothesis, the hypothesis of extended mind (HEM). I argue that HEC is most usefully formulated as a question of *cognitive system demarcation*, and that thought examples (e.g., the Otto-Inga case) are a poor source of evidence for assessing its plausibility. Section 3 rebuts a set of well-known arguments against HEC based on the concept of natural kind. Here my argument has both a positive and a negative conclusion. I argue that extended cognitive systems are compatible with contemporary theories of natural kinds, but under closer scrutiny it turns out that natural kinds are not a promising resource for answering the question of cognitive extension.

After this conceptual groundwork, sections 4 and 5 present my own solution to the problem of cognitive system demarcation. I suggest that the currently most prominent theory of explanation in the special sciences, James Woodward's contrastive-counterfactual theory, and a multi-

dimensional account of explanatory power building on this theory can be used as the foundation for my differential influence (DI) account. Section 6 collects the threads. By means of an example drawn from human memory research, section 6 illustrates how the DI criterion together with the assessment of explanatory virtues forms a workable account of cognitive system demarcation.

2 Hypothesis of Extended Cognition: Domain and evidence

Already in their seminal article, Clark and Chalmers (1998) made a distinction between the notions of extended cognition and extended mind. While the distinction has often been overlooked in the literature, keeping it in sight improves conceptual clarity and allows one to isolate the real sticking points related to HEC.[1] Following common usage, I take the hypothesis of extended cognition to mean that

> (HEC) Human cognitive processes can span the brain, body, and external world, and cognitive states sometimes comprise parts of the external world (Spaulding, 2011).

By contrast, Clark and Chalmers (1998) put forward the hypothesis of extended mind as the following claim:

> (HEM) Human mental states such as beliefs can be partly constituted by the environment.

The two hypotheses concern slightly different things: HEC is a claim about cognitive processing, whereas HEM concerns the location of human mental states. Although the difference might appear insignificant, I suggest that the hypotheses apply to different domains, address largely different issues, and depend on different sources of evidence.

Attempting to exhaustively define mentality or cognition would be a hopeless task, but both everyday use of the notions as well as the scientific and philosophical literature indicate that there are important differences between how the two concepts function. Firstly, cognitive capacities are

---

[1] For recent examples of the loose use of HEC and HEM, see Menary (ed.) 2010. Most of the contributions in the volume use 'extended cognition' and 'extended mind' interchangeably.

generally characterized as knowledge-related psychological capacities employed in intelligent action (cf. Newell & Simon, 1976; Neisser, 1976; Bechtel et al., 1998; van Gelder 1995), and therefore they form only one subclass of psychological phenomena. In contrast, the scope of mental language is broader, and it encompasses aspects of mindedness not directly related to knowledge, such as affect and sensation.

Another important difference between mental and cognitive language has to do with the contexts in which they are used. A central function for mental vocabulary is to attribute propositionally described thoughts to ourselves and fellow human beings. It thus serves an interpretative function in allowing us to describe each other as intentional and rational agents. On the other hand, cognitive language is paradigmatically employed in the explanatory contexts of the mind sciences. Cognitive states and processes featuring in these explanations are often subpersonal, and the content of the states is often not *semantically transparent* so that it could be expressed in terms of propositional contents (Clark, 1989). Furthermore, many prominent theories of explanation in psychology and neuroscience suggest that mentalistic belief-desire explanations do not play a substantial role in the explanatory practices of the mind sciences (Cummins, 2000; 2010, p.vi; Bickle, 2006; Bechtel, 2008; Chemero & Silberstein, 2008).

The general problem regarding the nature of the relationship between common-sense psychology and scientific psychology is still an open question in philosophy, and beyond the scope of this article. However, I take the considerations above to suggest that HEC and HEM have somewhat different *domains:* HEM applies to common-sense psychological language, whereas HEC primarily concerns explanatory contexts in the psychological sciences. The two hypotheses also have different sources of *evidence*. As mental vocabulary is often involved in description of conscious psychological processes, semantic and ontological intuitions regarding mindedness appear as relevant evidence for assessing HEM. Attributing mental states to extended human-artifact systems would raise difficult questions regarding agential identity, ownership, and responsibility, and should therefore be approached with caution. On the other hand, as cognitive vocabulary functions primarily in the

description and scientific explanation of knowledge-related phenomena (both in humans and artifacts), HEC appears not to be directly constrained by our *a priori* intuitions regarding mentality.

This suggests that the most widely used example in the literature, the Otto-Inga thought experiment, is a poor source of evidence for assessing HEC. The question raised by the thought experiment is whether we should treat the information stored in an external artifact (Otto's notebook) as a mental entity, as a genuine *belief*. Although the example clearly has implications for HEM, it is not directly relevant for HEC, since according to a widely shared naturalistic viewpoint, scientific theorizing should not be liable to our lay intuitions about phenomena. Instead, revisions of scientific theories and concepts should be based on the assessment of the objective correctness of hypotheses (cf. Ladyman & Ross, 2010).

Examples drawn from actual cognitive science research are a more relevant source of evidence for HEC. Its plausibility ought to be decided by comparatively assessing the correctness and explanatory power of externalist and internalist research strategies found in the cognitive sciences. Importantly, from the perspective of scientific research, the crucial question is usually not whether a particular entity or state (e.g., Otto's notebook) is cognitive *per se*, but the problem concerns which factors to include in an explanation of a cognitive capacity, or more generally, *how to demarcate the cognitive system* in question in a justified way. Like several authors in the literature, I adopt this terminological practice of phrasing HEC as a question concerning the delineation of cognitive systems (Rupert, 2009; Wilson & Clark, 2009; Weiskopf, 2010; Ladyman & Ross, 2010; Kaplan, 2012).

These differences between HEC and HEM do not mean that the hypotheses are completely independent of each other, or that there would be no important connections between the cognitivist and mentalist vocabularies. However, keeping the differences between these perspectives in sight should at least shift the burden of proof in the debate so that the identity of HEC and HEM cannot simply be assumed, but has to be argued for. Moreover, being explicit about the dissimilarities between the two hypotheses clears ground for a naturalistic and scientifically relevant approach to the

hypothesis of extended cognition: it helps in avoiding thorny ontological issues related to HEM, and allows us to focus on a set of scientifically relevant questions in the disagreements surrounding extended cognition. As suggested above, HEC concerns primarily the issues of (1) how concepts and taxonomies in psychology should be formulated, and (2) how psychological phenomena ought to be explained.

From this perspective, the notion of *explanatory power* has been considered as important to cognitive system demarcation, and it has also played a central role in the recent arguments against extended cognition. One of the most common ways to argue against extended cognitive systems has been to argue that they fail to capture genuine natural kinds, and therefore have no explanatory power. I now critically examine this group of arguments that I call *heterogeneity objections.*

3 Cognitive Kinds and Heterogeneity

In a well-known argument, Adams and Aizawa (2001, p. 58; 2008, pp. 63–68) refer to the differing properties of intracranial and extracranial cognitive components as a reason for rejecting extended cognitive kinds. Intracranial human memory, for instance, has distinctive properties (law of effect, primacy effect, recency effect, etc.) not found in extracranially supported beliefs. From this Adams and Aizawa conclude that symbols inside and outside the skull should not be lumped under the same cognitive kind. In a similar vein, Robert Rupert (2004) argues that forming "generic cognitive kinds" comprised of both intracranial and extracranial mental states deprives extended approaches of explanatory power. These heterogeneity objections can be reconstructed as consisting of roughly the following steps:

(P1) Scientific concepts refer to natural kinds.

(P2) Cognitive properties in the brain and outside it are realized by different kinds of causal processes that do not form a causally unified set.

(P3) Natural kind concepts should not refer to such internally heterogeneous sets.

(P4) HEC entails subsuming intracranial and extracranial components under the same natural kind concept.

(C) Since HEC creates defective natural kind concepts, we should not adopt the hypothesis.

The argument relies on the sensible view that natural kinds ought to be characterized by well-defined sets of causal powers (Fodor, 1987; Sterelny, 1990) and that generic kinds such as the one including both intra- and extracranial memories would not have such sets. In science we should put like with like, and if it turns out that a concept refers to a causally heterogeneous group of entities, it should be revised so that a good fit between causal structures and scientific taxonomy is achieved. Therefore, if it were shown that HEC produces causally heterogeneous classifications devoid of explanatory power, this would be a powerful naturalistic argument against the hypothesis.

As observed by Walter and Kästner (2012), heterogeneity objections presuppose a traditional understanding of natural kinds. According to such theories of natural kinds, kinds are often thought of as being defined by necessary intrinsic properties, and the notion of natural kind is understood as closely intertwined with the concept of law of nature. (cf. Bird & Tobin, 2008). However, it is widely agreed that such theories are not suitable for describing phenomena in the life sciences: in the special sciences, scientifically interesting kinds of phenomena typically have no intrinsic essences, but are often characterized by relational properties, nor do they correspond to laws of nature (cf. Boyd, 1991; Murphy, 2006, ch. 9). In consequence, it has become increasingly popular to think of natural kinds as mechanistically sustained property clusters (cf. Samuels & Ferreira, 2010). According to the homeostatic property cluster (HPC) theory, originally introduced by Richard Boyd (1991, 1999), a natural kind consists of

(α) a cluster of co-occurring properties sustained by

(β) a homeostatic causal mechanism.

When applied to psychological phenomena, HPC theory suggests that cognitive natural kinds could be described as consisting of (α) the cluster of observable properties characteristic of the phenome-

non (e.g., a cognitive capacity) and a corresponding (β) cognitive mechanism that guarantees the reliable co-occurrence of the properties in the cluster.

As has been pointed out by Paul Griffiths (1997) and Dominic Murphy (2006, ch.7), among others, psychological phenomena are often sustained by hybrid mechanisms consisting of a variety of dissimilar components. For example, the properties of many psychiatric disorders and socially constructed emotions cannot be explained by referring only to intracranial factors, but the mechanisms behind these phenomena also include important social factors. According to HEC theorists, extended cognitive phenomena generally require similar kinds of explanations: socio-artifactually scaffolded cognitive abilities typically involve tight causal couplings between the brain and the environment, and should therefore be conceptualized as being supported by causal mechanisms extending beyond the human individual.

Now, this idea of extended cognitive mechanisms does not imply postulating generic cognitive kinds, and does not make HEC susceptible to heterogeneity objections. As HPC theory suggests, including extracranial components as genuine parts of cognitive mechanisms does not require that the internal and external components would have to be causally similar. In assuming that HEC leads to lumping intra- and extracranial parts under the same natural kinds (P4), heterogeneity objections appear to suffer from a conceptual mistake that could be characterized as a grain-size error: when proponents of HEC claim that internal and external components together *constitute the mechanism* explaining the properties of a cognitive natural kind (e.g., spatial memory), they need not be committed to the idea that the components (e.g., internal and external "memory states") *themselves belong* to the same kind.[2] That is, the components c_1 ... c_n of the mechanism sustaining the kind K_1 do not have to be members of the same kind K_2. In fact, the strategy of the second-wave extended

[2] I suspect that this mistake partially stems from a conflation between HEC and HEM. As suggested in section 2, the debates about HEM typically concern the correct domain of application of common-sense psychological predicates, and one is led to ask whether intra- and extracranial "memories" could fall under the same natural kind. However, the everyday notion of memory does not play an explanatory role in contemporary psychological theories, and it has been replaced by more precise descriptions of distinct memory systems (Squire, 2004). Hence, a HEC theorist need not try to fit environmental components into the taxonomy of our common-sense mentalistic language.

cognition theorists has been to move from arguments employing the parity principle to complementarity arguments. Second-wave theorists concede that there are important differences between intracranial and extracranial cognitive components, but emphasize that these heterogeneous elements play complementary constitutive roles in integrated mechanisms sustaining cognitive capacities (Sutton, 2010b; Menary, 2007).

This compatibility of HEC with the currently most widely accepted theory of natural kinds in the philosophy of psychology implies that appealing to natural kinds is not sufficient for establishing the conclusion of heterogeneity objections. However, a recent argument by Craver (2009) suggests that theories of natural kinds cannot offer direct support to HEC, either. According to Craver, HPC theory is of little use in finding the correct classifications of phenomena. This follows from the central role of the notion of mechanism in the theory. While HPC theorists themselves have not rigorously analyzed the concept, there has been extensive discussion on mechanisms in the philosophical literature on scientific explanation. In these discussions, mechanisms are characterized as organized collections of entities put together to explain a particular phenomenon (see section 4). In other words, mechanisms are always mechanisms *for* something. Hence, while mechanisms consist of objective causal structures, the way their boundaries are drawn is explanandum-relative. This appears to reveal a problematic conventionalist aspect of HPC theory: natural kinds are where the mechanisms are, but identifying mechanisms requires that explananda have already been determined (Craver, 2009).

Craver's argument shows that prior explanatory considerations play a role in determining how natural kind classifications in the special sciences are formulated. Hence, although HPC theory is useful as a clear account of how concepts, mechanisms and phenomena are related to each other, it alone is a poor resource for answering the question of how to demarcate cognitive systems; the mere compatibility of extended cognitive systems with theories of natural kinds cannot be used as a positive argument for the explanatory power of HEC. Neither has such a satisfactory positive argu-

ment been offered by the defenders of extended cognition. Although they have suggested that HEC brings forth new interesting explananda and makes more comprehensive explanations of cognitive phenomena possible, thus increasing the explanatory power of theories, no explicit account of the central notion of explanatory power itself has been given (Clark & Chalmers, 1998; Clark, 2007; Wilson, 2004; Sutton, 2010b). Therefore, there is an important theoretical gap in the literature: both the critics and the proponents of HEC rely in their arguments on vague notions of explanatory power. This raises a possible counter-argument against the explanationist perspective on extended cognition, according to which explanatory considerations are too imprecise to count as a basis for cognitive system demarcation. In the next section, I aim to remedy this shortcoming in the explanationist approach. I present a theory of explanation in the psychological sciences that can be used as a solid conceptual foundation for the assessment of the explanatory power of extended cognitive systems.

4 Explaining Cognitive Phenomena

In the psychological sciences, a central theoretical aim of research is to understand how the cognitive capacity of a system is made possible by its material structure. As has been argued by several authors, the classical reductionist view of scientific explanation that conceives of explanation in terms of subsuming explanandum events or higher-level laws under the laws of nature gives an inadequate description of explanatory activities in the cognitive sciences (Cummins, 1983; Craver, 2007). A more plausible portrayal of constitutive explanation in psychology characterizes explanatory practices as a combination of the heuristics of decomposition and localization (Bechtel & Richardson, 2010): Complex cognitive capacities are functionally decomposed into simpler subcapacities, which are ultimately localized in concrete structural parts of the system. This process can be conceived of as a form of mechanistic explanation, because the understanding produced by such explanations arises from knowledge about the dependencies between system-level properties and the properties of the lower-level structures supporting them. For instance, by knowing mechanistic

details of how long-term potentiation (LTP) occurs in the human hippocampus, we can understand many features of the normal and abnormal functioning of human spatial memory (Craver, 2007).

In order to make explicit how knowledge of dependencies between parts and wholes explains, and to develop a theory of explanatory power, it is useful to supplement this picture of explanatory heuristics with perhaps the currently most prominent theory of causal explanation in the special sciences, the contrastive-counterfactual theory (CC-theory) (Woodward, 2003).

Contrastive-Counterfactual Theory of Explanation

A shared starting point for several theories of scientific explanation has been to distinguish explanation from other epistemic activities (e.g., description and prediction) by emphasizing that explanations offer information of a specific kind: explanations tell *why* or *how* something happened. The contrastive-counterfactual theory suggests that such questions are answered by tracking counterfactual dependencies between the relata in the explanation. In the case of causal explanation, explanations track objective relations of counterfactual dependence between things in the world. That is, had the cause been intervened upon, the effect would have been different (Woodward, 2003).

The notion of *understanding*, on the other hand, can be cashed out as the ability to answer *what-if-things-had-been-different* questions. Therefore, according to this view, understanding produced by explanation is not a mental state but an inferential ability going beyond mere knowledge of observed correlations: although descriptive knowledge makes prediction and classification of phenomena possible, counterfactual explanatory knowledge is needed in order to be able to reliably manipulate their properties. For instance, having detailed information about the counterfactual dependence between cellular mechanisms in the hippocampus and the properties of spatial memory allows one to make inferences about how a person's navigational capacities would change were there changes to the functioning of the LTP mechanism. The same applies also in non-reductivist explanatory contexts: by understanding how the maternal style of talking about past events affects

the development of a child's autobiographical memory, one can explain why there are differences between the memory capacities of grown-up individuals (Nelson & Fivush, 2004).

Explanations also appear to have a contrastive structure (van Fraassen, 1980; Schaffer, 2005; Woodward, 2003). They can usually be characterized as answers to questions of the form: *why fact rather than [foil]*, where the foil is an exclusive alternative to the fact. Explanatory knowledge thus has the following form:

(EK) *x [x'] because of y [y']* (variable X takes the value x instead of x' because Y has the value y instead of y')

This analysis of the nature of explanatory knowledge helps us understand why scientists often require that genuine explanations describe mechanisms behind phenomena. In the extensive contemporary discussions on causal mechanisms, it is generally agreed that a mechanism consists of (i) a collection of causal parts (ii) organized together to sustain a stable phenomenon (Machamer et al., 2000; Glennan, 2002). Mechanism is thus a part of the causal structure of reality, isolated for explanatory purposes. Describing a mechanism behind a phenomenon increases understanding by allowing one to answer *what-if*-questions regarding what would happen to the explanandum phenomenon, if there were changes to the parts of the mechanism or their organization. This interpretation of mechanistic explanation also illuminates the central methodological role of interlevel experiments in the mind sciences (Craver, 2007, Ch. 4; Craver & Bechtel, 2007). In these experiments, one either intervenes on a mechanistic component and observes the changes in system properties, or the other way around. Woodward's account of explanation construes the epistemic import of such research strategies as resulting from the ability of the experiments to reveal relations of counterfactual dependence between system-level properties and the mechanistic components of the system.

The contrastive-counterfactual theory has three further implications that are crucial for my argument. It suggests that (i) explanations are always aspectual, and it offers clear accounts of (ii) explanatory power and (iii) explanatory relevance.

Aspectuality of Explanation

Scientific explanations do not explain things as such (e.g., spatial memory *per se*), but they explain why the explanandum variable takes a certain value – rather than some other value – by referring to the fact that the set of explanans factors was in a certain state (and not in some other, contrastive, state). That is, although the contrastivity of explanations is rarely fully explicit, under more careful analysis explanations are often best understood as tracking change-related dependencies between values of variables. As I argue in section 6, this implies that *prima facie* competing scientific explanations and theoretical paradigms can often turn out to be compatible, and should therefore be seen as mutually enriching rather than exclusive – once the explananda are carefully identified and the implications of findings made sufficiently clear.

Explanatory Power

Secondly, CC-theory suggests that explanatory power could be understood as the degree of understanding conveyed by an explanation, which in turn can be defined as the number and importance of the counterfactual *what-if* inferences that the explanation makes possible (Ylikoski & Kuorikoski, 2010). Goodness of explanations can therefore be assessed by comparing the degrees of understanding that different explanations give. According to this inferentialist notion of explanatory goodness, explanatory power (or depth) is not a single property of a theory or an explanation, but a multidimensional comparative concept that is determined by the amount, reliability, and cognitive usability of the counterfactual inferences that it makes possible.

For example, the *non-sensitivity* of an explanation can be defined as the robustness of the explanatory dependence between the explanans and the explanandum against changes in the values of these variables and changes in background conditions. On the other hand, the *precision* of an explanation is determined by how sharply the explanandum and its contrast classes are specified (Woodward, 2003, ch.6). These two explanatory virtues, non-sensitivity and precision, are often in conflict. When we make our description of the dependence between explanans and explanandum

more precise (e.g., by providing a quantitative model of the relationship between the two variables), this often makes the explanation more sensitive to changes in environmental conditions and reduces external validity.

Building on CC-theory, Ylikoski and Kuorikoski (2010) suggest that, in addition to the above-mentioned explanatory virtues of non-sensitivity and precision, there are three further virtues that should be taken into account when comparing rival explanations: Explanations also differ on how many distorting idealizations they employ (*factual accuracy*), how well they mesh with the rest of scientific knowledge (*degree of integration*), and how easy they are to use (*cognitive salience*). While the last two dimensions might appear more pragmatic than the first ones, all five have an influence on how rich a source of counterfactual inferences an explanation is in scientific practice. And again, there are often trade-offs between the virtues. A highly idealized explanatory text might be easy to understand and work with (salience), but its low factual accuracy might decrease its usefulness for real-life applications.

It is contestable whether this list of five explanatory virtues is exhaustive. However, the dimensions nicely capture many of the properties often associated with explanatory power: Unification, the significance of causal detail, and the importance of focusing on the causally most central factors can all be cast in terms of improvements on some of these dimensions of explanatory power (ibid.). The multi-dimensionality of explanatory power also suggests that different *profiles of explanatory virtues* might be appropriate for different epistemic contexts. In the example discussed in section 6, I illustrate how this account of explanatory power can be used as a resource for the task of cognitive system demarcation. I suggest that intracranialist and externalist approaches to the study of human memory have slightly different epistemic aims, ask different explanatory questions, and thus occupy different regions in the space generated by the five virtues.

Explanatory Relevance

The contrastive nature of explanations allows us to distinguish between those putative explanantia that make a difference to the explanandum contrast and those that do not. As I show below, this idea gives CC-theory a clear criterion of *explanatory relevance* that has been missing from many theories of scientific explanation. It is commonly agreed in the philosophy of science that adding irrelevant detail to an explanation makes it worse. A good explanation has to distinguish between relevant explanatory factors and mere causal background conditions. A similar judgment also plays an important role in the literature on cognitive extension. Critics of HEC often claim that once cognition is allowed to extend beyond skin and skull, the problem of *cognitive bloat* arises: it becomes impossible to distinguish genuine parts of cognitive systems from mere background causal factors (Adams & Aizawa, 2001, 2008; Weiskopf, 2010).

The problem of cognitive bloat is a challenge that a proponent of HEC must be able to satisfactorily answer. For the hypothesis of extended cognition to have non-trivial content, it has to be possible to rule out overly liberal extensions of cognitive mechanisms, and there must be a principled way of determining their boundaries. A common strategy for answering this problem – relying on natural kinds – was rebutted in the previous section. I now put forward a more promising solution that builds on considerations of explanatory relevance.

5 The Differential Influence Criterion

In his discussion of scientific research as distributed cognition, Ronald Giere (2002, p. 294) suggests an informal yet potentially powerful criterion for cognitive system demarcation. According to Giere, we should distinguish between those features of a system that *differentially influence* its output in scientifically relevant ways from other features that merely make it possible for the system to generate any output at all. Let us consider the following example as an illustration of this idea. The beating of the heart is a necessary precondition for a cognitive capacity like spatial memory, but unlike neurotransmitters in hippocampal synapses, the heart does not appear to qualify as a genuine

component of an explanatory mechanism of a cognitive ability. This is because there are subtle dependencies between changes in the concentration of a neurotransmitter and the properties of the explanandum capacity, whereas the heart can influence cognition only in a very coarse manner by enabling it or disabling it altogether.

I propose that this plausible but somewhat vague idea of differential influence can be made more precise by employing CC-theory. According to the theory, a contrast drawn in an explanans is relevant for a particular explanandum only if it makes the difference between the explanandum and its contrastive foil (see EK in section 4). That is, Y is relevant for X iff the contrast between values y and y' explains the difference between x and x'. On the other hand, irrelevant contrasts are ones for which the difference between y and y' has no influence on X. Similarly, causal background factors are ones for which both x and x' share the same value of Y.

Hence, the reason why the beating of the heart should not be included in the cognitive mechanism for spatial memory is that from the perspective of the explananda studied in the psychological sciences, changes in the value of the heart variable fail to distinguish between the explanandum state and its contrast. The research questions asked in these sciences (e.g., concerning the properties of spatial memory) are such that both the explanandum state and its contrast require that necessary background conditions for brain function such as a functional heart are in place. On the other hand, the hippocampus qualifies as a part of the cognitive mechanism for spatial memory, because changes in its various properties (structural integrity, LTP-process, neurotransmitter levels, etc.) are relevant for explaining fine-grained differences in navigational capacities of humans and animals (cf. Squire, 2004). The notion of differential influence can thus be spelled out in terms of such subtle counterfactual dependencies. Therefore, I suggest the following demarcation criterion for cognitive systems:

(DI) Given a set of explananda, a cognitive system ought to include those, and only those, mechanistic components that are needed to explain the contrasts between explanandum states and their contrastive foils.

Above we saw how differential influence can distinguish bodily background factors from genuine parts of cognitive mechanisms. The DI criterion also avoids clear cases of implausible environmental extension. This is because environment factors that lie far causally upstream from the brain usually fail to explain scientifically interesting contrasts. For example, raindrops hitting the window of my office are causally connected to my perceptual system, but fail to make a difference between the contrasts drawn in most scientifically relevant explanandum questions. In general, differential influence implies that distal targets creating perceptual stimuli usually fall outside cognitive systems: In psychology it is generally of no explanatory interest to know how a particular target in the environment is related to cognitive processing. Instead, we want to explain general psychological capacities that, for example, allow us to visually recognize classes of objects in the outside world. Such explanations should usually not be sensitive to the manner in which the retinal images are produced (e.g., natural scenes vs. photographs) and therefore the outmost relevant variables tend to reside at the level of visual transducers.

Differential influence appears not to be excessively liberal about the scope of cognitive systems. On the other hand, in plausible cases of cognitive extension, bodily or external factors do differentially influence pertinent scientific explananda. Consider a classic example, sensory substitution. In tactile-vision substitution, a blind person's vision can be partially restored by creating a feedback system consisting of a head-mounted camera, a tongue display unit and the person (Bach y Rita & Kercel, 2003). Several important properties (e.g., spatial and temporal resolution) of the newly acquired perceptual competences depend in a non-trivial way on the properties of the external parts of the system, and thereby many variables referring to the internal architecture of the extensions pass the differential influence test.

Cognitive Systems Pluralism

Together these considerations suggest that my differential-influence solution, building on the contrastive-counterfactual theory of explanation, succeeds in blocking cognitive bloat, while in princi-

ple allowing for cognitive extension. However, these desirable consequences appear to come with a price. As can be seen from DI, my analysis implies that an entity being cognitive is not only relative to its physical context (the mechanism or the environment that it is embedded in) but relative also to the explanatory question asked. Thus, by prescribing that cognitive systems ought to be drawn in different ways in the light of different epistemic aims, the differential influence account appears to lead to a possibly problematic proliferation of cognitive systems (Rupert, 2004). That is, if each explanatory context suggests its own way to demarcate cognitive systems, this could lead to an unconstrained classificatory pluralism.

While a comprehensive defense of pluralistic approach to scientific classification is beyond the scope of this paper (cf. Kauffman, 1970; Dupré, 1993; Mitchell, 2003; Wimsatt, 2007), the kind of classificatory pluralism implied by DI appears plausible in the light of research practices in the life sciences. Different scientific fields employ different classification schemes at various levels of abstraction. I suggest that this is due to the complexity of the phenomena studied in the biological and cognitive sciences. That is, in order to produce finite and understandable theories of their targets, different research programs aim to capture different parts of the complex causal web sustaining them, because by doing so they can focus on those properties that are central to the epistemic aims of their particular discipline. In consequence, theories, models, and concepts can be seen as epistemic tools that finite agents employ in trying to understand complex systems (Griffiths & Stotz, 2008), and classificatory pluralism appears as a consequence of the *division of cognitive labor*.

Moreover, evolved complex systems are often characterized by strong environmental couplings and many of their scientifically interesting properties are not intrinsic but relational. In such cases it is likely that our everyday intuitions about natural kinds lead us astray, and actual practices of scientists might be a more reliable guide to reasonable system demarcation. In fact, Robert Richardson (2008) and Mark Couch (2009) have recently drawn attention to the existence of complicated cross-classifications in the life sciences. For instance, in the light of some epistemic aims, EYE

and ENZYME are respectable scientific kinds whereas for other more detailed purposes they do not represent sufficiently unitary phenomena. I see no reason why the situation in the cognitive sciences should be any different.

Furthermore, classificatory pluralism need not imply subjectivism or relativism. More plausibly, it amounts to moderate perspectival realism: Despite the differences between the perspectives of different scientific fields, it is reasonable to treat fields as fairly stable units constituted by groups of scientists sharing common problems, techniques and vocabularies (Darden, 1978). That is, explananda in a scientific field are not formed subjectively but are often shared between researchers working in the field and inherited from earlier research. For instance in the cognitive sciences, the separate fields studying roughly the same targets are differentiated by methods and by different levels of description. From the explanatory viewpoint we can plausibly treat these scientific fields as being characterized by finite and partly overlapping *clusters of explananda* concerning causally real targets. In consequence, given a scientific field, the set of explananda is fixed, and the DI criterion can be used to produce unambiguous judgments of the boundaries of cognitive mechanisms.

In consequence, rather than understanding the pluralist outcome as an argument against my account, I suggest that it is in accord with actual scientific research practices, and that the search for a unique correct taxonomy of systems in the cognitive sciences might be a misguided aim. Furthermore, in the next section I discuss the controversy between externalist and internalist explanatory strategies in memory research to illustrate how the DI criterion, combined with local assessments of the explanatory virtues of classification schemes, results in a workable picture of cognitive system demarcation.

6 Explanatory Virtues in Memory Research

Much of the mainstream research on human memory has traditionally been conducted by performing simple laboratory experiments of recalling lists of random digits, words, or nonsense syllables (cf. Eysenck & Keane, 2005, Ch. 6–7). The main reasons for doing the studies in the laboratory are

the need to control confounding environment variables as well as the aim of isolating the "naked" human brain-mind as the target system. Such research is a clear instance of *explanatory internalism* focusing on cognitive mechanisms bounded by the skull. Internalism has some obvious explanatory virtues.[3] In controlled experiments it is often possible to measure the values of the studied variables at a high level of accuracy without interference from background variables. Hence, laboratory studies can offer very *precise* knowledge about the counterfactual dependence between the independent and dependent variable. For instance, in studies of short-term memory, observing the effects of manipulations of the retention interval on recall yields precise information about the nature of the dependence between these variables, the forgetting curve. However, due to the artificial conditions, such explanations score much lower on the dimensions of *non-sensitivity* and *factual accuracy*. Firstly, outside the laboratory it is likely that changes in environmental conditions disrupt the dependence observed in the study, sometimes to the extent that it fails to obtain under most conditions. That is, observed dependencies are highly sensitive. Secondly, the external validity of the findings may also be compromised by the simplifications employed in the experimental set-ups: while experiments are taken to give information about human short-term memory capacities in general, this might actually be a rather inaccurate characterization of the studied variables. For tractability reasons, stimuli and reactions in the studies are often simple and unrepresentative of their real-life counterparts. Interpreting the observed dependencies between the simplified variables as applying to inputs and outputs of real-life memory capacities in general is not often a warranted generalization, but instead a distortion decreasing the factual accuracy of an explanation.

Such problems of ecological validity led cognitive psychologist Ulric Neisser (1981, pp. xi–xii) to lament that a hundred years of psychological study on memory had produced hardly any results that would have relevance for real memory phenomena outside the laboratory. Since this pessimistic comment, ecological and everyday approaches similar to Neisser's (1981, 1988, 1997) have

---

[3] See section 4: non-sensitivity, precision, factual accuracy, degree of integration and cognitive salience.

become legitimate traditions within memory research (Sutton, 2010a). These traditions differ from the laboratory paradigm at least in three distinct ways, regarding their conceptions of *what* kinds of memory phenomena should be studied, as well as *how* and *where* they should be studied (Koriat & Goldsmith, 1996).

Emphasis on the ways that the interactions between the organism and its environment constrain cognitive capacities has helped the ecological tradition to uncover several distinct explananda ('what') which had not been distinguished before. While memory always involves employing information about the past for present or future purposes, there is still no agreement about the number of distinct memory systems (Squire, 2004). Within the ecological tradition it has been suggested that even within autobiographical memory, different kinds of memories of past events serve different functions: Schematic memories of event-types purportedly prepare the individual for future occurrences of the same kind of event and are therefore employed in individual problem-solving tasks. Autobiographical memory of particular events in one's life, on the other hand, appears to be a different capacity altogether, primarily evolved for the maintenance of the stable sense of identity and for social coordination (Barnier et al., 2008; Conway et al., 2001, p. 493). Furthermore, these distinct memory functions appear to make different demands for the structures realizing them. For instance, accuracy of memory has been one of the main variables studied in laboratory research, but in real-life inferential contexts where remembering is employed, it is rarely the main goal of recall (Hyman, 1994). It has been suggested that rather than being a literal replay of the past, often the crucial function of memory is to flexibly recombine pieces of information to facilitate simulation of future events (Schacter et al., 2007).

Uncovering distinct explananda within memory has gone hand in hand with methodological commitments different from the laboratory approach ('how', 'where'). Within the ecological paradigms, much emphasis has been put on the importance of using naturalistic stimuli and studying the dynamics between the individual and environmental factors. It has been common to emphasize the constructive nature of remembering: several studies suggest that remembering ought not be con-

ceived of as retrieval of previously stored information-packets, but as active pattern-completion process where environmental input and internal memory traces play complementary roles (Schacter, 1996; McClelland & Rumelhart, 1986, p. 193; Elman, 1993). The right kind of activation pattern of environment variables is an important trigger for recall; the brain needs its environment in order to remember. Hence, not all variables relevant to memory processes reside inside the skull.

Nice examples of such an externalist approach to memory can be found in the literature on social memory phenomena (Barnier et al., 2008; Barnier & Sutton, 2008). For example, consider the *transactive memory* approach that focuses on remembering dyads or groups as its units of analysis. The approach is motivated by the insight that focusing only on isolated individuals fails to capture the mechanisms behind social remembering. Instead, closely related people such as married couples often use each other as memory aids, and even interactively construct memories in conversation (Wegner et al., 1985, 1991; Harris et al., 2011). As transactive memory focuses on systems larger than the individual – and on cognitive mechanisms extending to the environment – it falls squarely within the extended cognitive systems approach. From the point of view of DI, such extensions beyond the individual brain are plausible, because the details about the cognitive structure of the partner differentially influence the properties of the explanandum-capacity, socially supported memory. We could perhaps say that such research is a manifestation of HEC in action.

Different Profiles of Explanatory Power

This review of the two different memory research paradigms suggests that they are characterized by partly different explananda and that they demarcate their units of analysis differently – laboratory and ecological approaches corresponding to explanatory internalism and externalism, respectively. These differences in approach lead to different profiles of explanatory virtues. As was seen above, laboratory research restricted to brainbound mechanisms can often produce very precise yet sensitive characterizations of the dependencies between explanantia and explananda. The externalist approach, on the other hand, appears to increase explanatory power in two general ways: Bringing

new explananda in sight, using naturalistic stimuli, and conducting research in natural settings increase the factual accuracy of the explanations and allows for easier extrapolation to real-life cases. Furthermore, by incorporating external variables in the explanatory mechanisms, the dependencies between these variables and explananda can be made explicit, and thus the explanation becomes less sensitive to unknown background influences.

However, extending systems also creates problems of experimental control and raises worries of confounding. And as critics of HEC have argued, including contingent environment properties (other people, artifacts) as genuine parts of cognitive systems might lead to loss of generality of the explanation. In particular, Robert Rupert (2010) has argued that the ad-hoc nature of extended systems fits poorly with the epistemic aims of scientific psychology. For example in developmental psychology there appears to be little use for explaining why a system comprised of a person and a certain artifact develops in the way it does. Rather, we want to explain the developmental feats of the naked brain.

Put in terms of CC-theory, Rupert claims that externalist approaches produce explanations that fail to correspond to the cluster of explananda studied by the psychological disciplines. While the argument appears plausible at first glance, there are empirical results that suggest that the situation is more complex. Studies on cognitive development suggest that normal development of several psychological capacities often relies on reliable environmental features (Griffiths & Stotz, 2000). For instance, consider again the social-interactionist explanation of autobiographical memory mentioned in section 4, according to which the individual differences between adult autobiographical memory capacities are caused by different reminiscence styles of caregivers. A satisfying explanation of the development of autobiographical memory, a capacity apparently belonging to an isolated mind, requires focusing on a complex social-cultural-cognitive-neural system (Nelson & Fivush, 2004). A similar insight underlies perhaps the majority of the examples of cultural and artifactual cognitive scaffolding discussed by the proponents of HEC. In these cases, research focusing exclu-

sively on intracranial mechanisms would be biased to look for explanatory factors in the wrong place.

Regardless of its soundness, Rupert's argument does again remind us of the existence of trade-offs between the different explanatory virtues. Studying isolated systems under laboratory conditions results in simpler, more *cognitively salient*, explanations, whereas more complex explanations taking into account the details of interactions between the organism and its environment might ultimately turn out to be factually more accurate, more precise, and less sensitive to disturbances from unknown environment factors. That being said, even in cases of tight environmental coupling, there is a limit to how much systems should be extended: as the number of variables included in the system increases, it often becomes increasingly hard to keep track of dependencies between individual variables, and this reduces the inferential usefulness of the model. In fact, the existing division of labor between psychology and the social sciences could be seen as an attempt to deal with this problem of cognitive finiteness. Different scientific fields employ truncated mechanisms, with psychologists focusing on intracranial factors and social scientists on non-psychological variables, each keeping out of the other's territory. Such disciplinary structure of science results in a situation where modular explanations respecting organismic boundaries are often easier to *integrate* with the rest of our contemporary scientific knowledge, and thus serve as more efficient inferential devices.

Deflating HEC

This analysis of the explanatory virtues of the competing paradigms is only tentative. However, it gives support to the pluralist approach to cognitive system demarcation. How cognitive mechanisms are demarcated and how the included variables are contrastively described is an important determining factor for explanatory power. As the differential influence criterion implies, drawing the boundaries for systems and explanatory mechanisms in different ways results in changes in the profile of explanatory virtues of the theory. However, the question of which classificatory strategy is the cor-

rect one *tout court* appears ill-posed. Rather, different research programs and scientific fields are characterized by different explanatory aims (i.e. clusters of contrastive explananda), and to optimize epistemic efficiency, those explanatory demands should be met with a theory that can be used to draw the required inferences. That is, in each case we should aim for a good fit between explanatory demands and the profile of explanatory virtues of a theory. The profile of explanatory virtues, in turn, depends on how the mechanisms and corresponding systems are demarcated, and the trade-offs between the different virtues are adjusted by drawing the boundaries in different ways.

As a more general conclusion, this picture of cognitive system demarcation suggests a deflationary interpretation of HEC. By accepting the hypothesis, one is committed to the claim that the idea of extended cognitive systems is coherent. However, such a possibility claim alone gives little guidance in actual decisions of system delineation. Therefore, I suggest that HEC itself should not be understood as an empirical hypothesis concerning the scope of the human mind. It is more fruitfully interpreted as a strategy for creating classifications of cognitive phenomena, and its relationship to empirical results is indirect. In consequence, the explanationist approach to extended cognition is not undermined by the fact that HEC in itself has no direct empirical implications, and its correctness cannot be judged by employing methods such as inference to the best explanation (cf. Sprevak 2010). We should not try to determine the truth or the explanatory power of HEC itself, but the assessment must be directed to empirical hypotheses and classification systems produced by the externalist explanatory strategies, and this involves engagement with actual scientific research.

7 Counter-arguments and Alternatives

Given the picture of cognitive system demarcation presented above, at least two ways of arguing against the DI account remain open for an intracranialist. One might suggest that instead of accepting the possibility of there being extended cognitive systems and mechanisms, examples such as transactive memory should be conceived of in terms of conservative intracranialist ontology as cases of interaction between several distinct non-extended systems. As I noted above, this strategy of

dividing complex networks of interactions into smaller systems has indeed been the prevailing approach in mainstream research in psychology and the social sciences. Moreover, if low-bandwidth causal interfaces between such systems can be found, the strategy is likely to work quite well. However, in cases of genuine extended cognition where there are complicated causal couplings between the different subsystems, employing a collection of separate systems does not result in the same explanatory power as an integrated extended system. This is because the piecemeal strategy fails to offer information on how the mechanisms supporting different subsystems are causally related. Therefore, in order to answer the same set of explananda-questions as the externalist approach, the causal interactions between the elements of the different subsystems must be worked out by further research, thus resulting in a description of an extended mechanism.

Another conceivable way of defending intracranialism would be to claim that even conceding the possibility of extended mechanisms, we should stick to non-extended cognitive systems by allowing mechanisms to extend beyond system boundaries. That is, intracranialism about cognitive phenomena could be saved by allowing that although system boundaries are drawn at organismic boundaries, the constitutive mechanism could extend outside these boundaries. As above, this strategy is conceptually coherent, but abandoning the link between systems and mechanisms in constitutive explanations suggested by DI opens up several difficult questions. Firstly, in order not to beg the question against externalism, the intracranialist must come up with a justification for preferring the conservative ontology also in scientific classification. The discussion on scientific concepts and natural kinds in section 3 suggests that such a task will not be trivial. Secondly, this move raises the fundamental question of why systems matter in the first place. After all, CC-theory implies that the explanatory power of a theory depends primarily on mechanism – not system – demarcation. This suggests that often the role of system delineation might be heuristic in nature and ultimately subordinate to the search for relevant mechanistic factors: an important motivation for employing extended cognitive systems in psychological research has been to extend the search for explanatorily rele-

vant causal factors beyond the individual, and thereby counter reductionist biases that often prevail in the sciences (cf. Wimsatt, 2007; Bateson, 1972, p. 459).

In consequence, the benefits of saving intracranialism by severing system and mechanism demarcation from each other in this way appear meager. The strategy can be used to salvage the intracranialist ontology of cognitive systems, but such a decision distances the notion of system from much of its epistemic motivation. Moreover, adopting this strategy would not undermine the pluralism implied by the differential influence account: even keeping target systems constant, considerations of explanatory power and relevance entail that different epistemic aims require different demarcations of cognitive mechanisms.

Mutual Manipulability vs. Differential Influence

Not only intracranialists but also several advocates of HEC might have misgivings about cognitive systems pluralism. Due to limitations of space, it is not possible here to engage in a comprehensive comparison between the DI account and other theories of extended cognitive system demarcation (cf. Menary 2007; Rupert, 2010; Wilson & Clark, 2009; Ladyman & Ross, 2010). However, the mutual manipulability approach recently proposed by Kaplan (2012) raises a particularly interesting challenge to my account: starting from a mechanistic-interventionist theory of system membership and emphasizing the role of empirical considerations in drawing the boundaries of cognitive systems, Kaplan's approach appears to represent a largely similar perspective on cognitive system demarcation as the one outlined in the current paper. However, in contradiction to the pluralism implied by the differential influence account, Kaplan argues that the mutual manipulability criterion results in objective and unique boundaries for cognitive systems. I respond to this challenge by briefly pointing out why, *pace* Kaplan, relationships of mutual manipulability alone are not sufficient for cognitive system demarcation.

According to Kaplan, what counts as a genuine component of a mechanism is determined by the presence of relationships of mutual manipulability between the properties and activities of puta-

tive components and the overall behavior of the mechanism in which they figure. Kaplan claims that it is possible to find objective boundaries for cognitive systems by including in the system only those factors that satisfy both top-down and bottom-up intervention criteria: manipulations of putative parts must show as differences in system-level properties and the other way around. The theory relies solely on considerations of causal-constitutive relevance and promises to offer unequivocal boundaries for cognitive systems. However, I claim that the pluralist explanandum-relativity of cognitive systems sneaks in through the back door. While mutual manipulability tells us how to calibrate system-level properties with those of their mechanistic components, it does not tell which interventions qualify as cognitive ones, and fails as a criterion for demarcating *cognitive* systems.

To see the insufficiency of the mutual manipulability criterion, consider again the relationship between the heart and spatial memory. The functioning of the heart is a causal precondition for cognition, and hence it seems that the bottom-up influence exists. But so does the top-down influence. There are system-level interventions on the organism that radically influence the functioning of the heart – suffocating the person, for instance. Mutual manipulability appears to implausibly suggest that the heart qualifies as a part of the cognitive system. Kaplan's reaction to this particular case (also proposed by Craver 2007, pp. 157–158) is to resort to a more demanding notion of bottom-up manipulability: Kaplan suggests that in this case the existence of mutual manipulability requires that both inhibition and stimulation interventions on the heart must influence cognition, and because the stimulation intervention (purportedly) does not have such an influence, mutual manipulability does not hold.

I see two problems with this reply. First, Kaplan's response seems slightly *ad hoc*, because the requirement of the effectivity of both inhibition and stimulation interventions is introduced only in connection to this particular example. Secondly, I think that as a criterion for cognitive system membership, Kaplan's refined notion of bottom-up manipulability is too strict: It is plausible that often a cognitive capacity relies on the normal functioning of at least some of its components in such a way that inhibition interventions on these components would result in changes in the system

properties, but stimulation of the components above the level of normal functioning would not have such a fine-grained influence on the system. In such cases, Kaplan's refined notion of mutual manipulability would erroneously leave these genuine components outside the cognitive system.

In addition to these problems, elaborating the notion of bottom-up intervention does not help in ruling out unwanted top-down interventions like the one mentioned above (suffocating the person), and therefore mutual manipulability might also lead to too liberal demarcations of cognitive systems. Instead, to rule out non-cognitive top-down interventions on the organism, we would have to in advance specify the scope of cognitive variables we are interested in, and restrict possible interventions to these variables. Kaplan's account thus begs the question in relying on an intuitive account of what counts as cognitive – it presupposes that cognitive explananda are specified in advance.

In sum, the mutual manipulability approach falls prey to cognitive bloat in trying to base system delineation on considerations of causal relevance only, where assessment of explanatory relevance would be needed. That being said, once the explanandum-dependence of variable selection is made explicit, relations of mutual manipulability are likely to track largely similar relevance dependencies as the ones that my DI account relies on; in practice, the two theories probably often result in similar practical recommendations for cognitive system demarcation.

8 Conclusion

I have argued for an approach to HEC based on the assessment of explanatory relevance and explanatory power. Unlike most other accounts in the literature, my solution suggests that the problem of cognitive system demarcation ought not be thought of as a project for finding the one correct taxonomy of cognitive natural kinds. Instead, I advocate a moderate form of mechanisms-based cognitive systems pluralism. The resulting position is not an unconditional defense of extended cognition: I contend that choices between externalist and internalist classification strategies are necessarily more local, and based partly on the epistemic aims of the scientific field in question. The

differential influence account implies that the strict dichotomy between extended cognition and brain-bound cognition presupposed by much of the philosophical debates on the topic might not be a fruitful way to approach questions of system demarcation in the cognitive sciences. That is, although much of the excitement surrounding HEC has derived from its ability to radically question the traditional intracranialist picture of the nature and explanation of cognitive phenomena, the multi-dimensionality of explanatory power suggests that there is room in scientific psychology for various systems of classification designed for different explanatory aims.

Granted, the deflationary perspective on HEC proposed in this paper often sidesteps rather than addresses head-on many of the sticking points in the philosophical literature on extended mind and cognition. Referring to explanatory virtues as a means for determining the boundaries of cognitive systems will hardly satisfy those preoccupied with deep ontological questions regarding the extension of the mind. However, if separating HEC from HEM and approaching the former as a question of scientific classification results in a coherent approach that isolates a set of issues in extended cognition that are pertinent to the mind sciences, then the existence of alternative interpretations of the problem of cognitive extension does not count as a serious argument against the feasibility of the explanationist approach. Taking the explanatory turn in the extended cognition debate will not solve the solemn philosophical question of whether the mind really extends to the world or not, but it can provide useful conceptual tools for the systematic assessment of the epistemic power of classification schemes in the cognitive sciences. In particular, the mechanistic picture of extended cognitive systems can put ecological and mainstream psychological paradigms conceptually on an equal footing and facilitate systematic analysis of the explanatory strengths and weaknesses of the externalist and internalist approaches.

References

Adams, F., & Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology*, 14, 43–64.

Adams, F., & Aizawa, K. (2008). *The Bounds of Cognition*. Malden: Blackwell Publishers.

Bach y Rita, P., & Kercel, S. (2003). Sensory substitution and the human–machine interface. *Trends in Cognitive Sciences*, 7, 541–546.

Barker, M. (2010). From cognition's location to the epistemology of its nature. *Cognitive Systems Research,* 11, 357–366.

Barnier, A., & Sutton J., & Harris, C.B., & Wilson, R. (2008). A conceptual and empirical framework for the social distribution of cognition: The case of memory. *Cognitive Systems Research,* 9, 33–51.

Bateson, G.(1972). *Steps to an Ecology of Mind.* New York: Balentine Books.

Bechtel, W. (2008). Mechanisms in cognitive psychology: What are the operations. *Philosophy of Science,* 75, 983–994.

Bechtel, W., & Abrahamsen, A., & Graham, G. (1998). The life of cognitive science. In W., Bechtel & G. Graham (Eds.), *The Companion to Cognitive Science* (pp. 1–104). Oxford: Blackwell.

Bechtel, W., & Richardson, R. (2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Cambridge: MIT Press.

Bickle, J. (2006). Reducing mind to molecular pathways: explicating the reductionism implicit in current cellular and molecular neuroscience. *Synthese*, 151, 411-434.

Bird, A., & Tobin, E. (2008). Natural kinds, in E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/.

Boyd, R. (1991): Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies*, 61, 127–148.

Boyd, R. (1999). Kinds as the 'Workmanship of Men.' In J. Nida-Rümelin (Ed.), *Rationalität, Realismus, Revision* (pp. 52–89). Berlin: Walter de Gruyter.

Chemero A., and Silberstein M. (2008). After the philosophy of mind: Replacing scholasticism with science. *Philosophy of Science,* 75, 1–27.

Clark, A. (1989). *Microcognition. Philosophy, Cognitive Science and Parallel Distributed Processing*. Cambridge: MIT Press.

Clark, A. (2007). Curing cognitive hiccups: a defense of the extended mind. *Journal of Philosophy*, 163–192.

Clark, A. & Chalmers, D. (1998). The extended mind. *Analysis*, 58, 10-23.

Conway, M.A., & Pleydell-Pearce, & C.W., Whitecross, S.E. (2001). The neuroanatomy of autobiographical memory, *Journal of Memory and Language*, 45, 493–524.

Couch, M. (2009). Multiple realization in comparative perspective. *Biology and Philosophy,* 24, 505–519.

Craver, C. (2007). *Explaining the Brain*. Oxford: Clarendon Press.

Craver, C. (2009). Mechanisms and natural kinds. *Philosophical Psychology*, 22, 575–596.

Cummins, R. (1983). *The Nature of Psychological Explanation.* Cambridge: MIT Press.

Cummins, R. (2000). 'How does it work?' versus 'what are the laws?': Two conceptions of psychological explanation. In F. Keil & R. Wilson (Eds.), *Explanation and Cognition* (pp. 117–144). Cambridge: MIT Press.

Darden, L. (1978). Discoveries and the emergence of new fields in science. *Philosophy of Science*, PSA 1978, 149–160.

Dupré, J. (1993). *The Disorder of Things: Metaphysical foundations of the disunity of science.* Cambridge & London: Harvard University Press.

Elman, J. (1993). Learning and Development in Neural Networks: the importance of starting small. *Cognition*, 48, 71–99.

Eysenck, M. & Keane, M. (2005). *Cognitive Psychology. 5th ed.* Hove: Psychology Press.

Fodor, J. (1987). *Psychosemantics.* Cambridge: MIT Press.

Giere, R. (2002) Scientific cognition as distributed cognition. In P. Carruthers, & S. Stich & M. Siegal (Eds.), *The Cognitive Basis of Science* (pp. 285–299). Cambridge: Cambridge University Press.

Glennan. S. (2002). Rethinking mechanistic explanation. *Philosophy of Science*, 69, 3.

Griffiths, P. (1997). *What Emotions Really Are.* Chicago: University of Chicago Press.

Griffiths, P., & Stotz, K. (2000). How the mind grows: a developmental perspective on the biology of cognition. *Synthese,* 122, 29–51.

Griffiths, P., & Stotz, K. (2008). Experimental philosophy of science, *Philosophy Compass,* 3, 507–521.

Harris, C., & Keil, P.,& Sutton, J. & Barnier, A., & McIlwain, D. (2011). We remember, we forget: Collaborative remembering in older couples. *Discourse Processes*, 48, 267–303.

Hutchins, E. (2010). Cognitive ecology. *Topics in Cognitive Science,* 2, 705–715

Hyman, I. E. (1994). Conversational remembering: story recall with a peer versus for an experimenter. *Applied Cognitive Psychology*, 8, 4-66.

Kaplan, D. (2012). How to demarcate the boundaries of cognition. *Biology and Philosophy*, 27, 545–570.

Kauffman, S. (1970). Articulation of parts explanation in biology and the rational search for them. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association,* 257–272.

Koriat, A., & Goldsmith, M. (1996). Memory metaphors and the real-life/laboratory controversy: Correspondence versus storehouse conceptions of memory. *Behavioral and Brain Sciences,* 19, 167–188

Ladyman, J., & Ross, D. (2010). The alleged coupling-constitution fallacy and the mature sciences, In R. Menary (Ed), *The Extended Mind* (pp. 155–166). Cambridge: MIT Press.

Machamer, P., & Darden, L. & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science*, 67, 1–25.

McClelland, J. L., & Rumelhart, D. (1986). A Distributed Model of Human Learning and Memory, In J.L. McClelland & D. Rumelhart (Eds), *Parallel Distributed Processing,* Volume 2 (pp. 170–215). Cambridge: MIT Press.

Menary, R. (2007). *Cognitive Integration: Mind and Cognition Unbounded*. Basingstoke: Palgrave Macmillan.

Menary, R. (ed.) (2010). *The Extended Mind.* Cambridge: MIT Press.

Mitchell, S. (2003). *Biological Complexity and Integrative Pluralism.* Cambridge: Cambridge University Press.

Murphy, D. (2006). *Psychiatry in the Scientific Image*. Cambridge: MIT Press.

Neisser, U. (1976). *Cognition and Reality.* San Fransisco: W.H.Freeman & Co Ltd.

Neisser, U. (1981) Memory: What Are the Important Questions? In U. Neisser, *Memory Observed: Remembering in Natural Contexts*. San Francisco: W.H. Freeman.

Neisser, U. (ed.) (1988). *Remembering Reconsidered: Ecological and traditional approaches to the study of memory.* Cambridge: Cambridge University Press.

Neisser, U. (1997). The ecological study of memory. *Philosophical Transactions of the Royal Society*, 352, 1697–1701.

Nelson, K., & Fivush, R. (2004). The emergence of autobiographical memory: A social cultural developmental theory. *Psychological Review*, 111, 486–511.

Newell, A., & Simon, H. (1976). Computer Science as Empirical Inquiry: Symbols and Search. *Communications of the ACM,* 19, 3, 113–126.

Richardson, R. (2008). Autonomy and multiple realization. *Philosophy of Science*, 75, 526–536.

Rupert, R. (2004). Challenges to the hypothesis of extended cognition. *Journal of Philosophy*, 389–428.

Rupert, R. (2009). *Cognitive Systems and the Extended Mind.* Oxford: Oxford University Press.

Rupert, R. (2010). Extended cognition and the priority of cognitive systems. *Cognitive Systems Research,* 11, 343–356.

Samuels, R., & Ferreira, M. (2010). Why don't concepts constitute a natural kind? *Behavioral and Brain Sciences,* 33, 222–223.

Schacter, D. (1996). *Searching for Memory: the brain, the mind, and the past*. New York: Basic Books.

Schacter, D., & Addis, D., & Buckner, R. (2007). Remembering the past to imagine the future: the prospective brain. *Nature Reviews Neuroscience*, 657–661.

Schaffer, J. (2005). Contrastive causation. *The Philosophical Review*, 114, 297–328.

Spaulding, S. (2011). Overextending cognition. *Philosophical Psychology*, 1–22.

Sprevak, M. (2010). Inference to the hypothesis of extended cognition. *Studies in History and Philosophy of Science,* 41, 353–362.

Squire, L. R. (2004). Memory Systems of the Brain: a brief history and current perspective. *Neurobiology of Learning and Memory*, 82, 171–177.

Sterelny, K. (1990). *The Representational Theory of Mind.* Oxford: Basil Blackwell.

Sutton, J. (2010a). Memory. *The Stanford Encyclopedia of Philosophy* (Summer 2010 Edition), Zalta, E. (ed.), URL = <http://plato.stanford.edu/archives/sum2010/entries/memory/>.

Sutton, J. (2010b). Exograms and interdisciplinarity: History, the extended mind, and the civilizing process. In R. Menary (Ed), *The Extended Mind* (pp. 189–221). Cambridge: MIT Press.

Van Fraassen, B. (1980). *The Scientific Image.* Oxford: Oxford University Press.

Van Gelder T. (1995). What might cognition be, if not computation? *Journal of Philosophy*, 92, 345–381.

Walter, S. & Kästner, L. (2012). The where and what of cognition: The untenability of cognitive agnosticism and the limits of the Motley Crew Argument. C*ognitive Systems Research,* 13, 12–23.

Wegner, D., Giuliano, T., & Hertel, P. T. (1985). Cognitive interdependence in close relationships. In W. Ickes (Ed.), C*ompatible and Incompatible Relationships* (pp. 253–276). NY: Springer-Verlag.

Weiskopf, D. (2010). The Goldilocks problem and extended cognition. *Cognitive Systems Research*, 11, 313–323.

Wilson, R. (2004). *Boundaries of the Mind*. Cambridge: Cambridge University Press.

Wilson, R., & Clark, A. (2009). How to situate cognition: Letting nature take its course. In P. Robbins & M. Aydede (Eds.), *The Cambridge Handbook of Situated Cognition* (pp. 53–77). Cambridge: Cambridge University Press.

Wimsatt, W. (2007). *Re-engineering Philosophy for Limited Beings*. Cambridge: Harvard University Press.

Woodward, J. (2003). *Making Things Happen*. Oxford: Oxford University Press.

Ylikoski, P., & Kuorikoski, J. (2010). Dissecting explanatory power. *Philosophical Studies,* 148, 201–219.