What Is a Cognitive System? In Defense of the Conditional Probability of Co-contribution Account

Rob Rupert, University of Colorado, Boulder

draft 5/15/2019

I. Introduction

The theory of cognitive systems has played a central role in recent debates in philosophy of mind and the philosophy of cognitive science, perhaps most notably in the debate about extended cognition (Clark and Chalmers 1998, Wilson 2002, Wilson 2004, Rupert 2004, 2009). In *Supersizing the Mind*, Andy Clark argues for the Principle of Ecological Assembly: "the canny cognizer tends to recruit, on the spot, whatever mix of problem-solving resources will yield an acceptable result with a minimum of effort" (Clark 2008, 13). In Clark's hands, this principle supports the hypothesis of extended cognition, that is, the claim that material beyond the boundary of the human organism partly constitutes a significant proportion of human cognitive states and processes; and the principle does so by allowing the human to be a growing-and-shrinking cognitive system, the components of which vary with context and purpose (Clark 2011). Such a system is centered on the organism, yet its material extent can vary – sometimes extending into the environment, sometimes not – as the organismic core recruits and discards resources (Clark 2007).

Discussion of the nature and individuation of cognitive systems has also informed debates about group-level mental or cognitive states. Edwin Hutchins suggests the following characterization of a cognitive system's boundary: it is fixed by a "steep gradient...in the density of interaction among media" (Hutchins 1995, 157), where such interaction consists in the computational transformation of representational units. Hutchins famously applies this "steep-gradient" criterion to crews of naval vessels, concluding that the crew of a boat, together with

many of its organismically external tools, sometimes constitutes a single cognitive system, a group mind of sorts.

In the context of debates about extended cognition and group minds, discussion of cognitive systems has sometimes focused on the kind of integration or statistical dependence required in order that a collection of causes or mechanisms form a coherent, functioning network, one that might be identified as an agent or a self (Sporns et al. 2004; Rupert 2009, chapter 3; Rowlands 2010, chapter 6; Friston 2013; Hohwy 2016) or, more generically, as the "unit of cognition" (Mandelblit and Zachar 1998, 254). But, questions about systemic integration bear on other issues as well, for instance, on the attempt to characterize the integrity of cognitive *sub*systems, such as modules or other domain- or task-specific components of larger, unified cognitive systems (Sporns et. al. 2004, de Brigard 2017, 224); in this vein, one might ask what sort of internal integrity determines that a collection of mechanisms or causes has a distinct identity within the broader architecture of a single cognitive agent. And, to come at matters from a somewhat different direction, one leading theory of consciousness – the Integrated Information Theory (Tononi 2008) – appeals to a measure of unity, *phi*, that, setting questions about consciousness aside, provides a potentially promising account of the relation that binds together components within a single cognitive system to the exclusion of other causal contributors to the production of intelligent behavior.

It would be a massive undertaking to canvas and evaluate all extant measures of systems integration. This essay serves a more specific purpose, to situate and refine what I now refer to as the 'conditional probability of co-contribution' account of cognitive systems individuation (CPC) (Rupert 2009, 2010) and to defend it against criticisms (Klein 2010, Clark 2011, de Brigard 2017).

II. Talmy and the Overlapping Systems Model of Cognitive Organization

In recent work, Leonard Talmy has developed the Overlapping Systems Model of Cognition (Talmy 2015). He takes as given the presence of various cognitive subsystems in humans – subsystems for language, affect, and culture, among others – and explores the extent to which the operation of various such systems depends on or incorporates structure that they share with each other. He examines systems pair-wise, identifying shared structure as well as structure that is distinctive of one or the other member of the pair. And, given his expertise in linguistics, Talmy focuses specifically on pairs one of which is the language system: language and the visual system, language and the somatosensory system, and language and the affect system, and so on. Talmy's results deepen our understanding of the ways in which various subsystems are integrated as well as of the distinctive aspects of each subsystem.

To compare subsystems, one must, on Talmy's view, divine the structuring principles of each of the subsystems in question. It is certainly true that both language and vision can represent, for instance, a chair, but that is a matter of shared content, not shared structure. To find structure in the case of the language system, Talmy looks to closed-class linguistic devices – as revealed by prepositions, conjunctions, or inflections, for example, and also embodied in syntactic structure itself. This introduces significant constraints, for, as Talmy notes, although it is typical for language to inflect nouns in way that indicates number (to have a device such as English's addition of an 's' to indicate the plural), no language includes noun inflections that indicate color. Even with regard to number, closed-class devices mark some distinctions but not others: in no language are nouns inflected so as to indicate reference to a nondenumerable, as a opposed to a denumerable, class of objects of a given type (*ibid.*, 13).

Consider the English prepositions ‘in’ and ‘along’. They manifest two different aspects of figure-ground structure as it is “built in” to the language system. In the sentence, “The radio is in the dumpster,” the figure precedes ‘in’ and the ground follows it, specifically in a containment relation with it; the figure appears within the boundaries of a container that serves as ground. And, similarly in the case of “Some water is in the vase” (*ibid.*, 20). In the case of ‘along’, the preposition indicates an entity’s linear movement, as figure, on a path that serves as ground, as in “The ball rolled along the ledge,” and “The hunter walked along the trail” (*ibid.*). Talmy proposes that “visual processing may produce schematic representations of the geometric relations of the two principle objects in each scene [that is, in each of the four scenes just described]” (*ibid.*). The suggestion is that, in these ways, the structure of the visual system overlaps with that of the language system. This all stands in contrast to the case of source-marking (Tosun et al. 2013), which is a component of linguistic structure in some languages (marking, for example, the distinction between a proposition’s being asserted on the basis of first-hand observation and a proposition’s being asserted on the basis of the testimony of someone else), but is not reflected in the structure of the visual system (Talmy 2015, 15–16).

Talmy’s search for overlapping structure among various subsystems could ground a theory of cognitive integration – that is, a theory of that in virtue of which various subsystems constitute a single cognitive system. In fact, there is a distinct affinity between the picture he paints of overlapping systems and the conditional probability of co-contribution view presented below. It is, however, not Talmy’s goal to provide such a theory. Although he dubs the view the “Overlapping Systems Model of Cognitive Organization,” there is no evaluative or criterial dimension to Talmy’s view. He does not claim, for example, that, in the absence of an overlap between systems A and B, those two are not subsystems of the same cognitive entity, that is, the

same cognitive self, subject, or agent. Talmy identifies a number of pairs of systems – language and affect, language and culture – that share little to no structure. And this passes without remark. He gives no indication that lack of such shared structure forces a wedge between these two systems to such an extent that they are not part of the same self or subject. Thus, although Talmy's work could provide the material for a theory of cognitive systems individuation – according to which, for instance, subsystems A and B are part of the same cognitive system only if there is substantial overlap in their structure – Talmy himself simply does not seem to have such a project in mind. Rather, he takes for granted a set of cognitive subsystems and takes for granted that they are all part of a single human cognitive system, and then he catalogues the kind and extent of overlap between various pairs of these systems, with some pairs exhibiting high degrees of overlap and some little at all.

Notice, too, that Talmy is neutral with respect to the underlying physical or neural mechanisms responsible for whatever overlap exists between cognitive subsystems. It might be that overlap exists because subsystems have evolved underlying mechanisms that happen to have the same structure, or because one system "borrows" the workings of a neural mechanism that is part of another, or because two systems both share a single resource that is not properly a part of either of the systems (*ibid.*, 11) (cf. Anderson 2010, 2014; Zerilli 2019). As I see things, a more bottom-up approach is called for if we wish to characterize integration in a way that does not presuppose knowledge of the boundaries of cognitive systems, that is, if we are after a location-neutral criterion of cognitive systems individuation.

III. Conditional Probability of Co-contribution

3.1 The conditional probability of co-contribution account

What, then, individuates a cognitive system? In this subsection, I build upon my earlier work on the topic (Rupert 2004, 2009, 2010, 2011), presenting a refined version of what I now call the 'conditional probability of co-contribution' account of cognitive systems individuation (or CPC).

There is strong evidence that *at least one* coherent, integrated cognitive system appears inside the skin. A wide range of successful research programs in cognitive science treat organisms as the locus of cognition and identify consistency in their output across varying materials and experimental conditions; these include programs investigating everything from the false-belief task (Wellman, et al. 2001) to visual neglect (Driver and Mattingly 1998) to the processing of narrative (Gernsbacher and Robertson 2005) to the illusory truth effect (Dechêne et al. 2010). Regardless of what might be mainstream cognitive science's other failures, these successes demand explanation; and if there were not some kind of relatively stable and coherent cognitive system inside the skin, such successes would be mysterious. Note that one cannot immediately infer from these successes that there is only one human cognitive system, the one inside the skin. Not at all. That would clearly beg the question against advocates for an extended view of cognition. Rather, the best (available) explanation of such successes is that *one* coherent, functional, integrated cognitive system appears inside the skin. (This might be called the 'at least one' step in the argument against extended cognition.) That some integrated, cognition-related system appears within the boundary of individual organisms best explains why various modeling approaches that focus on the individual organism – indexing behavioral data to persisting organisms while varying extra-organismic materials – have been as successful as they have been.

As a methodological matter, however, once we are committed to the existence of *a* (not necessarily *the*) cognitive system within the boundary of the organism, we should attempt to get as much causal-explanatory mileage out of it as we can. And, much mileage can be gotten from

it; by tracking regularities in the ways in which this system interacts with external materials, we see that one internal cognitive system suffices to account for the data that impress proponents of the extended view. After all, the proponent of the extended view must recognize an important interface between the organismic system, already admitted by all parties, and the external materials that might or might not, for all we know initially, be part of *a* genuinely cognitive system. All of the structure presupposed by my view – that there is an internal system that manages the use of external problem-solving materials – must be reproduced in a model formulated by an extended theorist. Thus, there is no advantage to gained by positing an additional cognitive system. The extended model includes the internal cognitive system plus the relevant external materials and the processes at the interface. It multiplies entities beyond necessity to replicate all of the structure at issue – including the internal cognitive system and the interface with the environment and the environmental materials – and further add that the entire collection is *another* cognitive system. (This might be called the 'at most one' step in the argument against extended cognition.)

There is, however, a catch. We should want to understand in virtue of what the internal cognitive system is integrated such that it can support flexible, adaptive, intelligent behavior of a wide variety of forms. Whatever the proper analysis, it shall leave open the door to an extended view of cognition. In attempting to identify the properties in virtue of which the internal cognitive system is cognitive, we characterize a *kind* of integrated system. And, generally speaking, once we have committed ourselves to the existence of a token *k*, in virtue of its playing causal-explanatory role *K*, we're committed to the existence of a *K*-type system wherever there appears a system with *K*-type causal-explanatory properties. Thus, if CPC successfully characterizes the *kind* of system to the existence of which we're already committed, we must

then accept that any system – including, possibly, an extended system – satisfying this characterization is a cognitive system, in addition to the internal system the analysis of which motivated that characterization. It follows, then, that my argument thus far speaks only against the introduction of a whole other kind of system, integrated in different way. The internal one does all of the work needed in cognitive science, and it does it in a simpler fashion than an extended approach that would explain the same data by introducing an additional system (say, a hybrid system) of a different kind. Thus, the data of interest to proponents of the extended view do not motivate the commitment to an extended system. What *could* motivate such a commitment would be the discovery of an extended system that happens to satisfy the same criterion that, I will argue, properly characterizes the internal system of interest.

How should we characterize the internal cognitive system? I have argued that virtually all successful organism-based forms of cognitive modeling – computational, brute biological, robotics-based, connectionist, and dynamicist – distinguish between, on the one hand, the relatively integrated, relatively persisting architecture, and, on the other hand, the more transient causal contributors that, together with aspects of the architecture, produce intelligent behavior (*cf.* Wilson's [2002] distinction between obligate and facultative systems). Think of this as another inference to the best (available) explanation: that a persisting, structured system consisting of integrated components is the distinctively *cognitive* thing beneath the skin best explains why the various forms of successful individual-level modeling posit such a structure. This doesn't take us far beyond the simplicity-based argument given above, but it reinforces the depth of the distinction between the integrated architectural components and the passing parade of transient contributors, and it pointedly raises the question of what binds together the elements of an architecture, allowing them function as a single cognitive system.

Can we say anything more precise about the integrated nature of the system inside the skin, anything that sheds light on its role as a cognitive system, that is, a system that flexibly produces a wide range of forms of intelligent behavior? I have proposed that a cognitive system consists of a collection of mechanisms that co-contribute in overlapping subsets to the production of a wide range of forms of intelligent behavior; and as a way to cash out the requirement "in overlapping subsets," I have proposed a mathematical measure of integration among the mechanisms that constitute a cognitive system (Rupert 2009, 2010), a refined version of which is presented below. Though motivated by an attempt to characterize the internal cognitive system, the measure is location-neutral; it distinguishes between two kinds of causal contributor, wherever they appear, to the production of intelligent behavior.

Here, then, is the conditional probability of co-contribution account of cognitive systems integration and individuation (CPC):

1. For a subject at a time, form each non-singleton subset of the mechanisms that have distinctively causally contributed to the production of any form of intelligent behavior.

2. For each subset, relative to each form of intelligent behavior,[1] there is, for each of its proper subsets, a probability of its being a causal contributor to the production of that form of behavior

[1] Relativizing the conditional probability to task-type is not, contrary to what de Brigard (2017) suggests, designed to identify the cognitive *sub*system responsible for performing a certain kind of task or family of tasks; it is not motivated by an attempt to find *the* system for x-ing, where x-ing is, e.g., processing language or recognizing faces. Rather, the idea is to identify a single cognitive system, something roughly analogous to the subject or self. And, in this context, contribution to the execution of multiple task-types should weigh in favor of a mechanism's being part of the cognitive system as a whole; adverting to conditional probabilities relative to each kind of task increases the likelihood, *ceteris paribus*, that a mechanism deployed across task-types will count as part of the overall cognitive system, which is at it should be. It does so because it gives the mechanism in question more chances to show up in a set with a high conditional probability (see step 5., below). And the more such chances it has, the more likely it is to appear numerous times on the list of sets with high conditional probabilities, and thus to count as part of the cognitive system.

conditional on every member of the complement of that set's contributing causally. (Here's an illustration: Take an edge-detection mechanism. It causally contributes to the avoidance of obstacles. So does a mechanism that computes distance from retinal disparity and so does a mechanism that calculates shape from detected shading [Marr 1982]. Relative to a given kind of behavior, say, obstacle avoidance, each of these three mechanisms can appear in two different two-member sets, and each can appear in one three-membered set. For each two-membered set, two conditional probabilities are relevant: the first-mechanism's contributing conditional on the second's, and vice versa. For the three-membered set, there are six relevant conditional probabilities: each single mechanism's contributing conditional on the other two's, and each combination of two's contributing conditional on the third's. Now go through this procedure – in principle! – for every grouping of all causally contributing mechanisms relative to each form of intelligent behavior that has been exhibited by the subject in question [so long as the subject has exhibited a reasonably wide range of forms of intelligent behavior – if not, all bets are off, for this richness of repertoire is one of the central features of the *explananda* of cognitive science].)

3. Rank order all such conditional probabilities.

4. Take the natural cut-off between the higher probabilities and lower ones. (If something's being an integrated system is a natural kind, and the current proposal is on the right track, we should expect such a statistically significant gap to appear.)[2]

[2] It's possible that multiple significant gaps appear on this list, which raises difficult questions concerning, for example, whether an individual subject might consist of multiple, nested cognitive systems. For my part, I'm inclined to see the question about extended cognition as a question about kinds of causes. For millennia, it's been recognized that intelligent behavior has internal causes and external causes. Typically, at least some subset of internal causes have been thought to be of a distinct type – to be mental, cognitive, or part of the agent or self. The claim that cognition extends can be understood as the claim that causes of *that type*, whatever we call them or however we conceive of them, frequently appear beyond the boundary of the organism. If multiple significant gaps

5. For each mechanism appearing on the list of sets with higher conditional probabilities (that is, the sets above the gap referred to at Step 4.), count the number of times that mechanism appears and rank order individual mechanisms accordingly (that is, according to their number of appearances on the list produced by Step 4.).

6. A statistically significant gap separates those mechanisms that appear frequently on the list from those that do not.

7. The integrated cognitive system comprises all and only those mechanisms appearing above that gap.

CPC was initially formulated in an attempt to adjudicate claims about extended cognition, in particular, the claim that contemporary cognitive science has revealed human cognition to be extended in a deep and theoretically important way (Clark and Chalmers 1998). If cognition must occur within the cognitive system, as delineated by CPC, then it would seem that for most individual human subjects at most times, cognitive processing occurs within the boundaries of the subject's body; for, generally speaking, the preceding characterization of the cognitive system cuts against the inclusion of special-purpose tools and one-offs, which tends to be the status of causal contributors beyond the boundary of the body. The location of individual human

---

appear, then the set of causes of intelligent behavior divides into three or more types. I think the most interesting version of the hypothesis of extended cognition projects causes of the "narrowest band" – causes from the most tightly integrated cognitive system – into the world beyond the organism. But weaker versions of the hypothesis of extended cognition might make reference to causes of intermediate status – those not below the lowest significant gap on Step 4.'s list but not above the highest gap. Matters are further complicated by the fact that parallel questions arise in connection with Step 5.; it, too, constructs a list that could have more than one statistically significant gap in it. So, possibilities for the identification of different kinds of cause ramify. Thanks to Luke Roelofs for pressing me to address the possibility of multiple, significant gaps.

cognition is largely an empirical matter, though. The systems-based proposal CPC leaves open the possibility that a tool – perhaps an iPhone (Chalmers 2008) – that consistently contributes to the production of a variety of forms of intelligent behavior across a variety of contexts, alongside a shifting set of co-collaborators that themselves have similar standing, is part of a human's cognitive system.

But *why* think CPC is correct? Flexibility is the heart of cognition and intelligence – flexibility in learning, in the acquisition of concepts and skills, in problem-solving, and in the deployment of a variety of resources in the pursuit of and revision of goals in an oft-changing environment. It is this flexibility that attracts attention to certain forms human behavior and performance, and motivates the development of a distinctive science (cognitive science) to study them, in contrast to tropes and other stereotyped forms of behavior. It is the lack of such flexibility that drives continuing complaints about extant forms of artificial intelligence. "It's not intelligence at all," one is tempted to say about systems, "It wouldn't have any idea what to do if an unexpected situation were to arise! It does only that one thing!" – whether that one thing is play chess, answer quiz-show questions, or control an automobile.

CPC is grounded in the idea that flexibility is achieved in humans only by the presence of many units poised to work together in various combinations. There's plentiful evidence that this sort of thing happens in the human brain (Anderson 2010, 2014, Cole et al. 2013, Botvinick and Cohen 2014). On some accounts of this sort of process, subnetworks with overlapping members wrest control from each other via competitive processing. When two functional subnetworks have overlapping members, it may take only a bit of differential stimulus to shift the agent's activity from the performance of one task to the performance of a different one. On this approach, a shift in task doesn't require an entirely new network to take control from a

previously dominant one; more subtle shifts in the co-activation of elements, some of which are already active, can more smoothly effect such a transition. The systems-based view CPC emphasizes what seems likely to be a central trait of such a system – that any given mechanism is capable of cooperating with various other subsets of mechanisms to complete a variety of tasks.

IV. Objections, Replies, and Refinements

IV. A Modularity

Andy Clark objects to CPC on the following grounds: "It is unclear...exactly how the proposed criteria would fare internally should, for example, the inner neural story turn out to be either strongly modular or (worse still, as thus affecting so-called 'central processing' too) massively modular" (2011, 456). The gist of the objection seems clear enough. If CPC's story about integration tells a tale of active collaboration – in which various resources causally interact with each other to produce behavior – CPC appears hostile to modular cognitive architectures. Relatively isolated processing streams – of the sort identified with modules – do not causally interact with each other. Thus one would not expect a measure based on active causal collaboration to group various modular subsystems into a single cognitive system. And, if CPC precludes modular cognitive architectures, something is deeply wrong with CPC.

This objection rests on a misunderstanding of CPC. The conditional probability of co-contribution account does *not* appeal to causal interaction among mechanisms; rather, CPC appeals only to the probability of co-contribution of mechanisms to the production of intelligent behavior: "For each such type of mechanism, relative to each kind of cognitive phenomenon that it helps to produce, there is a conditional probability...of its use relative to each of the other mechanisms, abilities, etc. in the set, as well as a conditional probability of its use relative to

each subset thereof" (Rupert 2009, 42). The materials out of which CPC builds cognitive integration and thus cognitive systems are threefold: instances of behavior, causal contributions of individual mechanisms to the production of those instances of behavior, and probabilities of co-contribution of various mechanisms to the production of intelligent behavior (which co-contribution does not require any causal interaction among the mechanisms in question), as these co-contributions are represented by conditional probabilities. In other words, CPC appeals, in the first instance, to the causal contribution of individual mechanisms to instances of intelligent behavior as produced by a given individual. Beyond that, the relevant question concerns only the probability that one mechanism is causally contributing to a specific form of intelligent behavior given that some others are. (Compare: there might be a statistical correlation between Pat's contributing to a charity and Terry's contributing to the same one, without there being any causal interaction between Pat and Terry.) To be fair, in informal descriptions, I have sometimes used language that suggests causal interplay among mechanisms (e.g., "a collection of persisting physical properties, the integrated interaction of which..." – *ibid.*, 39). The formal construction makes no mention, however, of causal relations among mechanisms, only of the conditional probability that one mechanism contributes to the production of a given kind of intelligent behavior given that some other mechanism(s) is contributing. And, this probability can be very high in a case in which the mechanisms in question are components of two distinct modules. Even if, for instance, verbal production and visual processing are entirely casually isolated from one another – implausible though that may seem – the probability of some bit of the visual system causally contributing to the production of a verbal response given that the parser contributes may be quite high, simply because whenever the parser contributes, the subject in question is visually tracking the target of her response.

Of course, it's not uninteresting to ask, when conditional probabilities are high, why they are high. In some such cases, the correct explanation of high conditional probability will be that the mechanisms in question causally interact during the process of producing the kind of behavior in question. But, in other cases, causal interaction among the mechanisms in question does not explain the high conditional probability in question, because the statistical relation results, for example, from a common cause.[3]

IV.B. Changes over Time

Felipe de Brigard (2017) has recently published an extended critique of CPC. He takes aim at various aspects of the view and, in his closing section, offers what he takes to be a competing account. I see a deep affinity between CPC and his alternative, though discussion of such affinity is left for another occasion. Here I focus on de Brigard's critical perspective, much of which rests on an erroneous conception of the motivation and structure of CPC. Even after such misconceptions have been cleared away, however, a significant critical point remains, connected to de Brigard's concern about the way in which the brain's pattern of contributions changes over time and about the accompanying changes in the contribution of cognitive mechanisms associated with various areas of the brain.

The purpose of CPC was to provide a location-neutral, empirically defensible way to individuate a given subject's cognitive system – her entire cognitive system, the sort of thing that

[3] This suggests a shortcoming in the way modularity is sometimes explained, in terms of informational encapsulation (Fodor 1983). On one influential notion of information, all that's required for one state or process to carry information about another is statistical dependence that obtains non-accidentally (that is, because of natural law –e.g., Dretske 1981, 38); no causal interaction is required. On this view, it's highly likely that even if Fodor (1983) had been right about the role of causally encapsulated modules in the human cognitive architecture, those modules would not have been informationally encapsulated. It's very implausible that, across the board, the activity of every module of the sort Fodor had in mind, would be statistically independent of the activity of every other module – or of central processing, for that matter.

might be identified with the cognitive self (Rupert 2004).[4] And, for this reason, I have sometimes referred to the subject's cognitive system as the 'architecture' (Rupert 2009, 42, 45, 51; cf. Wilson 2002, Weiskopf 2010). De Brigard takes me to be talking about subsystems, however – systems used for a specific kind of task, such as for face recognition, reading, or motor planning – and this sends the discussion off course. Recall that a core feature of the cognitive system as a whole is its flexibility. Humans can solve a wide range of problems; they effectively shift between tasks as well as between strategies for completing a given task, as the situation demands. In contrast, a subsystem responsible for performing a specific kind of task may well not exhibit this same sort of flexibility. Thus, integration within a distinct subsystem, *qua* individual subsystem, may well be a different phenomenon, resting on different internal relations, than the kind of integration that stitches together the entire cognitive system and is responsible for human intelligence. This observation marks a significant gap between my interests and de Brigard's focus on cognitive subsystems.

Consider now the complaint de Brigard raises in connection with the Posterior-Anterior Shift with Aging (PASA). What is PASA? Broadly speaking, as a human ages, the neural resources used to accomplish certain cognitive tasks shift from those located in more posterior areas of cortex to those located in more anterior areas. Brigard seems to think that CPC is out to identify, at the level of types, *the* network for performing specific tasks, for instance, face recognition, and that, once CPC has identified that subsystem in a subject, the system cannot change. De Brigard then brings the hammer down; given PASA, we know that it's common for

[4] The use of 'subject' here is not meant to suggest anything about subjectivity or subjective conscious experience. Rather, it is used in a sense closer to the one in which it's used in such statements as "the experimenters recruited 18 subjects, 11 female and 7 male." And, bear in mind that this use of 'subject' is consistent with the possibility that, within the subject, there is no master executive or Cartesian theatre (Dennett, 1991), that is, it is consistent with an account of the architecture of the cognitive system according to which control is distributed. Compare this to Rodney Brooks's robots: A distributed architecture controls them, yet, at the same time, it is crystal clear where the boundary lies between the persisting architecture and, in contrast, the other causal contributors to the production of its behavior (the soda cans, the wall, and so on – see Brooks 1999).

the subsystem responsible for performance of a given task by a particular subject to change over time, and thus CPC fails to reflect the reality of the internal human cognitive system.

But, de Brigard imputes to CPC a goal it simply doesn't have. It is no part of my brief to lay down conditions for the identification of cognitive subsystems, and I do not presuppose that each individual has one fixed subsystem for performing a given kind of task; neither do I hold that the individual has a single cognitive system, fixed for its life. Rather, CPC grows from a much messier vision of cognitive systems: for any given subject, at any given time, there is a certain causal history of mechanisms contributing to the production of instances of behavior. The conditional probabilities determined by that history, and the clustering of those conditional probabilities, individuates the entire cognitive system associated with a given subject or organism, regardless of whether there are any well-delineated subsystems for performing certain tasks. My reasons for relativizing the calculation of conditional probabilities to task types (see footnote 2) reflect my attempt to capture the flexibility of intelligence; they have nothing directly to do with, for example, standard practice in cognitive neuroscience that emphasizes the identification of the specific networks, shared across subjects, responsible for completing various tasks.

Part of what drives de Brigard's concern about CPC is his interpretation of CPC's conception of integration. As he sees things, CPC is out to identify the integrated set of mechanisms that are *necessary* for the subject to perform the task in question (de Brigard 2017, 228–229). Allowing a shift in the components of a cognitive subsystem that continues to perform the same task *would* be puzzling, if the idea behind CPC had been to characterize the necessary components (given some specification of that modality – cf. Rupert 2009, section 2.2.3) for the performance of the task in question by the subject in question. Consider a young person

performing task T1 using resources in posterior location. If those resources are truly necessary for that person to perform T1, then it can't be the case that the person performs T1 at a later age using different resources, now in anterior locations.

But, CPC makes no claims about a subject's necessary use of certain resources. Its framework is probabilistic. It is consistent with CPC that completely disjoint subsystems, within the very same subject, control the performance of a given type of task on different occasions – even near to each other in time, never mind over the course of a lifetime. And, that seems to me the right feature for a theory of cognitive systems individuation to have. I see no motivation for thinking that human cognitive systems are so inflexible; at least, such inflexibility shouldn't be presupposed at the outset. It might be that a given subject's brain employs only one specific network for the performance of a given task type; but there's also evidence of neural redundancy and short-term plasticity, which suggests that the same task can be performed by a single subject deploying different mechanisms on different occasions. These questions about the human case should be left open – awaiting data (such as the data supporting PASA) – and CPC does, in fact, leave them open. It's not part of the operative notion of integration that the integrated mechanisms that constitute a cognitive system (or subsystem) are the ones that *must always* work together or contribute to the performance of a given task.

A further point of misunderstanding concerns frequency of use. On a reasonable assumption about the correlation between neural resources and mechanisms, PASA entails a significant shift in the mechanisms typically used by the aging subject to complete tasks of the relevant types. As de Brigard sees things, CPC is not appropriately sensitive to these changing patterns of activity because it is "dependent exclusively on frequency" (de Brigard 2017, 237) of the use of a given mechanism or on frequency of co-activity of multiple mechanisms (*ibid.*, 233).

This is not how conditional probabilities work, however, not even conditional probabilities determined by past relative frequencies. Given my personal history, the conditional probability of seeing the vocalist Jon Anderson given that I'm at a concert by the band *Yes* equals unity. I've seen the band perform twice, and Anderson was present in both cases: 2/2 = 1. So, one gets a maximal past relative frequency, and thus a maximal conditional probability, from two very rare event types. (I've seen Anderson three times in total, once when he performed without *Yes*). On the other hand, the probability of say, being in Seattle given that I see police around me is fairly low, given that I've lived in many places and seen many, many police officers in all of those places. So, one gets a very low conditional probability from two very common event types. I've found myself in Seattle on thousands of occasions, and I've seen police officers on thousands of occasions; but given that the latter thousands are many more than the former, the probability of being in Seattle, given that I see police is low (roughly .25, I would guess; I've spent about 25% of my adult life in Seattle, and I assume that police sightings are roughly equally common across the various places I've lived). So, questions of absolute frequency play no role in the probabilistic foundation of CPC.

Questions about frequency of co-occurrence might seem more relevant, but even in that case, mathematics are not on de Brigard's side.[5] Consider that I have frequently seen police in Seattle – that is, seeing a police officer and being in Seattle have frequently co-occurred. Nevertheless, by manipulating the number of police sightings in other cities where I've lived, it's easy enough to drive the conditional probability in question up or down. Imagine that my current home of Denver is overrun by police. In that case, the conditional probability "being in Seattle"

[5] I suspect that the root of the misunderstanding at issue lies in de Brigard's conception of CPC, not in any mistake on his part with respect to the mathematics of probability. In the present context, though, it is worth working through the mathematics alongside the presentation of CPC, to try to prevent both forms of potential misunderstanding on the part of readers.

given “I’m seeing police” plummets, approaching zero. If, instead, we virtually eliminate police from all of the cities in which I’ve lived other than Seattle, the condition probability in question shoots to the ceiling. If there are, in fact, no police in the other cities in question, then the probability of being in Seattle, given that I’m seeing a police officer = 1. The upshot: The conditional probabilities relevant to CPC can vary quite significantly without any change in the absolute frequency of the co-occurrence of the event-types in question (thought of as, say, the span of the subject’s life divided by the number of token instances of the conjoint event type in question).

Thus, we must keep questions about conditional probabilities, and only conditional probabilities, in our sights, not absolute frequency either of events of a single type or of pairings of events of two different types. Even when we do, however, something important remains of de Brigard’s critique. Infrequently occurring events can result in what we might think of as excessively small sample sizes, at least if one is focused on a subject’s actual history. If there haven’t been many instances of a given event type (e.g., my being at a *Yes* concert), the actual past relative frequencies of its occurring give that an instance of some other event type does might not reflect the actual causal dispositions or tendencies of the kinds of events in question. Perhaps both times I see a band perform, the band members bring the same guest performer onto the stage, even though that performer has joined the band only twice in the band’s history (I just happen to have been there on both occasions). As a result, I’ve gotten the wrong impression about the status of the guest in question in relation to the band’s performances. Moreover, remaining closer to de Brigard’s worry, past events can make for bad samples when the causal dispositions underlying the events in question have changed significantly since the events took place. If one conceives of actual events as samples, meant to reveal the current level of

integration of the subject's cognitively relevant mechanisms, one should want to avoid positing a measure that yields determinations on the basis of outdated samples.

In defense of CPC, notice that disuse can change conditional probabilities reasonably quickly, to such an extent that an unused mechanism is quickly eliminated from the subject's cognitive system, by CPC's measure. Here's how. Relative to a given subject, time, and task type, take a posterior mechanism *m1* that has high probability of use given the use of *m2*, a high probability of use given *m3*, and so on for, say, ten other mechanisms *and* is such that for each of those twelve mechanisms, the probability of their use is high given the use of *m1*. (This perfect symmetry is unrealistic, but it will suffice to make the point. Note, too, that in order to streamline the discussion, I'm considering only cases of two-member sets.) Now shift attention to a time in that subject's future when *m1* has fallen out of use in favor of the use of an anterior mechanism, while the remaining mechanisms *m2–m13* all continue to be in used in essentially the same fashion. In which case, the number of times *m1* appears on the list associated with Section III's Step 5. is reduced by half, from 24 to 12 (relative to the task type in question), making *m1* much less likely to "make the cut" and be included as part of the subject's cognitive system. Even though the probability of the use of, for instance, *m3*, conditional on the use of *m1*, might remain high (simply because, when *m1* was, in the past, in use, *m3* frequently co-contributed), the converse conditional probability will sink. After the neural shift, the conditional probability that *m1*is being used, given that *m3* is, will head toward zero; for at the later time, whenever *m3* contributes, *m1* is *not* also contributing. And this kind of statistical washing out of past contributions is highly likely to occur in the context of human cognition, given the brain's constant activity and the range of tasks to which individual mechanisms contribute (Anderson 2010, 2014). So, when one mechanism falls out of use but another, previously co-contributing

one does not, CPC's measure is likely to eliminate quickly the former from the cognitive system without eliminating the latter (*modulo* complications to do with the former mechanism's possible contribution to other forms of intelligent behavior, which might be enough to keep it in the cognitive system, and for good reason).

Nevertheless, the vagaries of actual history seem likely to lay pitfalls. Is there a way to avoid them? Bear in mind that CPC was originally proposed as a diagnostic measure,[6] specifically because past relative frequencies are subject to noisy fluctuation. They can serve as a relatively reliable measure of the causal dispositions of the properties or structures of interest, but they serve only as proxy. Is there a more accurate way to home in on the kind of relation that holds among components of an integrated cognitive system?

I propose to amend CPC by invoking counterfactuals, in particular, counterfactual relative frequencies of the causal contribution of the mechanisms in question to the performance of various types of intelligent behavior. This kind of amendment invites elaborate formal specification, replacing Section III.'s talk of actual past causal contributions – as determinants of the conditional probabilities in question – with talk of conditional probabilities determined by carefully specified counterfactual deployments. For present purposes, though, it will suffice to indicate how this general idea handles de Brigard's objection. Consider PASA again. Whether a posterior mechanism should be part of the cognitive system at the time after an anterior shift depends on whether it continues to be *disposed* to contribute to the production of various forms of intelligent behavior. And, that is captured by asking what causal contributions it would make under various counterfactual conditions. If it is still "wired up and ready to go" in realistic situations, counterfactual considerations will reveal this, and CPC will render the appropriate

[6] "A mechanism's measuring up in the way I have described may not constitute its status as part of an integrated cognitive system. Nevertheless, so long as a subject has a fair amount of experience in the world, this measure is, I submit, highly correlated with, and thus at least diagnostic of, integration." (Rupert 2009, 43)

verdict. If it is no longer situated so as to contribute, then the counterfactuals in question will wash it out of the cognitive system. The reasoning applied two paragraphs back will now apply with even more force; for any given mechanism *m* in the cognitive system, the probability of the use of the no-longer-efficacious posterior mechanism, given the use of *m*, will sink toward zero, drastically reducing the number of times the no-longer-efficacious mechanism in question appears on the list associated with Step 4 (the first of the two "winners" lists), which has the effect of making them less likely to make the cut at Step 5. In the terms of the example schematized above, the now out-of-use and incapacitated posterior mechanism will sink, not from 24 to 12, but from 24 to zero, on the portion of Step 4.'s list that comprises sets with high conditional probabilities.

Before moving on, consider an additional way in which CPC's treatment of conditional probabilities might be modified so as to marginalize odd histories. Generally speaking, small sample sizes are to be discounted in statistical analyses, given how likely they are to be products of noise. This provides a principled way to handle cases of conditional probabilities that are flukishly high or low simply because the types of events in question have rarely occurred. In this way, we might eliminate, early in the process, consideration of certain high conditional probabilities (leaving them off the ranked order constructed at Step 3).

IV.C Klein's critique

In his critical notice of *Cognitive Systems and the Extended Mind*, Colin Klein raises two objections to CPC. The first of Klein's objections pertains to "inhibitory mechanisms like executive control; while they may contribute to many tasks, they need not be frequently co-active with any *particular* set of cognitive mechanisms" (Klein 2010, 254). The relevant kind of case in which inhibition plays a role in the human cognitive architecture has, I take it, the

following feature. If the inhibitory effect hadn't occurred, the behavior at issue wouldn't have been produced. In some cases, this is because, if the inhibitory mechanism hadn't acted, a different form of behavior would have ensued, as, for instance, a result of competitive processing (Botvinick and Cohen 2014). Assume causation is contrastive. If the query concerns, for example, the reason the subject reported on the color of the shape on the monitor rather than the size of that shape, the inhibitory mechanism can clearly make a causal contribution. Similar comments about causality apply when inhibitory mechanisms reduce the force of a response or when an inhibitory mechanism inhibits an inhibitor (so as to allow the occurrence of the process that the latter mechanism had been inhibiting). So, I don't see any special problem with inhibitory mechanisms, so long as a case can be made for their causal contribution to the production of behavior.

But, why does Klein ask about the frequency of co-activation of an inhibitory mechanism with any *particular* further mechanism? It's not the case that CPC requires of any particular mechanism that it frequently be co-active with any other, for the reasons articulated above. Setting aside questions about the absolute frequency of co-activity, though, we might nevertheless ask specifically about the status of mechanisms of executive control. So far as I can tell, such mechanisms are likely to make CPC's cut. For just about *any* form of intelligent behavior, it's likely that for any mechanism(s) other than the executive control mechanisms that help to produce that form of behavior, executive control mechanisms causally contribute alongside those other mechanisms; thus, for most mechanisms and for most forms of intelligent behavior, the probability of a given executive mechanism contributing to the performance that form of behavior will be high if the nonexecutive mechanism in question is contributing. For instance, whenever a mechanism contributes to the production of writing, it's likely that

mechanisms of executive control will also contribute (can one stay focused on the task of writing without a contribution from executive control?). Admittedly the converse conditional probabilities are most likely not high; for any given nonexecutive mechanism, the probability is low that it's being used given that a given executive control mechanism is being used. But, because so many forms of intelligent behavior are like writing – unlikely to be performed if not modulated in some way by executive mechanisms – executive mechanisms will show up many times in sets with high conditional probabilities.

The second of Klein's concerns constitutes what I'll call the 'problem of ambient light' (Rupert 2009, 43). As a run-up, consider CPC's treatment of background conditions. In some sense, oxygen contributes causally to the performance of every instance of a subject's intelligent behavior. Thus, relative to any subset of other mechanisms that have causally contributed to the production of intelligent behavior, the probability of oxygen's contributing given that that subset is, is unity. The converse probabilities are not high (they will simply be the marginal probabilities, given oxygen's status as an omni-contributor). Regardless, given how many high conditional-probability subsets oxygen is a part of, CPC would seem sure to include it in the cognitive system.

There is a straightforward way to deal with this potential problem: the limitation of CPC's application to mechanisms that *distinctively* causally contribute to the production of intelligent behavior (Rupert 2009, 42). Given that oxygen contributes causally to *every* process in the organism, regardless of whether it has anything to do with the production of intelligent behavior, the case can easily be made that oxygen's contribution to intelligent behavior is not distinctive. Notice, too, that oxygen does not appears as an element in typical models in

cognitive science; this alone might justify setting it aside as irrelevant (possibly because this provides evidence that oxygen does not make a distinctive contribution to cognition).

The problem of ambient light arises in the cases in which a pervasive quantity does, in fact, make a distinctive contribution to the production of intelligent behavior. The visual system is clearly part of the cognitive system and it uses light in a distinctive way; light makes a contribution to visual processing that it does not make to metabolism, respiration, digestion (etc.), in contrast to the case of oxygen. Light is, however, everywhere. Something is deeply wrong with CPC, if it includes light in the cognitive system.[7]

It's important to distinguish, however, between mechanisms and quantities and between token causal contributors and types of causal contributors. A given subject's cognitive system is a relatively persisting collection of token mechanisms that themselves contribute repeatedly over time (or are in a position to contribute repeatedly). The light that causally contributes to visually guided behavior isn't a mechanism,[8] and its contribution is one-off. Light is a pervasive quantity in nature, small amounts of which interact with a subject and go on their way. There is no single thing – this photon here – that causally contributes repeatedly to the production of intelligent behavior. And this is not a peculiarity about light. It would seem to hold for all of the senses – taste, smell, hearing – that interact with a stuff-like quantity in the environment, and thus provides the basis for excluding all such quantities from the cognitive system. If a problem remains, it reduces to the problem of infrequency discussed in the preceding subsection.[9]

---

[7] In Rupert (2009, 43), I attempted to solve the problem by appealing to a perceptual bottleneck. Klein (2010, 254) worries, quite reasonably, that this proposal begs the question against the extended-mind theorist. I hope here to solve the problem of ambient light in a way that is independently motivated and that clearly does not beg the question against the extended-mind theorist.

[8] For the exploration of a closely related distinction – between mechanisms and resources – as they play a role in mechanistic explanations, see Klein (2018).

[9] It would not be unreasonable to set aside light, sound waves, and the like as candidate mechanisms for inclusion in the cognitive system on grounds of their being at a different level in nature from cognition. Cognitive systems aren't causally isolated, and nor do they permeate nature (interacting only with each other); thus, some account must be

V. Conclusion

A theory of cognitive systems individuation has many potential implications. It may help to resolve debates about extended cognition, embodied cognition, and group cognition. Perhaps even more importantly, it might shed light on the nature of cognition itself – or at least on the nature of the sort of cognition to which human scientists have access (Rupert 2013). If our theory of cognitive systems individuation rests on a principle that captures something deep about cognition – for instance, by identification of the structure that facilitates flexible intelligence – some enlightenment has been delivered. In the preceding, I have argued for the conditional probability of co-contribution account of cognitive systems individuation, partly by defending it against criticisms, but also on the grounds that it reflects a deep truth about how flexibility is achieved, and thus how (at least the local kind of) intelligence is possible.

Looking ahead, CPC's prospects depend partly on how well it stacks up against other formally specified contenders. Integrated information theory, for example, might be pressed into service as a way of individuating cognitive systems (setting questions about consciousness aside). Graph-theoretic approaches also seem promising. Additionally, given the number of theorists who view predictive processing as the unifying principle of all cognition, perhaps the integrated nature of the cognitive system can be characterized in predictive processing terms. And, it will be worth asking, too, whether some of these various approaches are extensionally equivalent.

References

given of how a cognitive system interacts with the noncognitive natural world. The impinging of quantities behaving in the everyday, nondistinctive ways that physics describes provides a motivated boundary, though there is some risk here of begging the question against the extended mind theorist.

Anderson, Michael L. 2010. "Neural Reuse: A Fundamental Organizational Principle of the Brain." *Behavioral and Brain Sciences* 33: 245–313.

Anderson, Michael L. (2014). *After Phrenology: Neural Reuse and the Interactive Brain*. Cambridge, MA: MIT Press

Botvinick, Matthew M., and Jonathan D. Cohen. (2014). The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science* 38: 1249–1285.

Brooks, Rodney. (1999). *Cambrian Intelligence: The Early History of the New AI.* Cambridge, MA: MIT Press.

Chalmers, David J. (2008). Foreword to Andy Clark, *Supersizing the Mind* (in Clark 2008).

Clark, Andy. (2007). Curing cognitive hiccups: A defense of the extended mind. *Journal of Philosophy* 104, 4: 163–192.

Clark, Andy. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. New York: Oxford University Press.

Clark, Andy (2011). Finding the mind. *Philosophical Studies* 152: 447–461.

Clark, Andy, and David J. Chalmers (1998). The extended mind. *Analysis* 58: 7–19.

Cole, Michael W., Jeremy R. Reynolds, Jonathan D. Power, Greg Repovs, Alan Anticevic, & Todd S. Braver. (2013). Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature Neuroscience* 16, 9: 1348–1355.

De Brigard, Felipe. (2017). Cognitive systems and the changing brain. *Philosophical Explorations* 20 (2): 224–241.

Dechêne, Alice, Christoph Stahl, Jochim Hansen, and Michaela Wänke. (2010). The truth about the truth: A meta-analytic review of the Truth Effect. *Personality and Social Psychology Review* 14, 2: 238–257.

Dennett, Daniel C. (1991). *Consciousness Explained*. Boston, MA: Little, Brown and Company.

Dretske, Fred (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Driver, Jon, and Jason B. Mattingly. (1998). Parietal neglect and visual awareness. *Nature Neuroscience* 1: 17–22.

Fodor, Jerry A. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.

Friston, Karl. (2013). Life as we know it. *Journal of the Royal Society, Interface* 10, 20130475: 1–12.

Gernsbacher, Morton Ann, and David A. Robertson. (2005). Watching the brain comprehend discourse. In A. Healy (ed.) *Experimental Cognitive Psychology and Its Applications*, pp. 157–167. Washington, D.C.: American Psychological Association.

Hohwy, Jacob. (2016). The self-evidencing brain. *Noûs* 50, 2: 259–285.

Huebner, Bryce. (2013). *Macrocognition: A Theory of Distributed Minds and Collective Intentionality*. Oxford: Oxford University Press.

Hutchins, Edwin. (1995). *Cognition in the Wild*. Cambridge, MA: MIT Press.

Klein, Colin. (2010). Critical Notice: *Cognitive Systems and the Extended Mind* by Robert Rupert. *The Journal of Mind and Behavior* 31 (3&4): 253–264.

Klein, Colin. (2018). Mechanisms, resources, and background conditions. *Biology and Philosophy* 33:36 https://doi.org/10.1007/s10539-018-9646-y

Mandelblit, Nili, and Oron Zachar. (1998). The notion of dynamic unit: Conceptual developments in cognitive science. *Cognitive Science* 22, 2: 229–268.

Marr, David. (1982). *Vision*. New York: W. H. Freeman and Company.

Rowlands, Mark. (2010). *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*. Cambridge, MA: MIT Press.

Rupert, Robert D. (2004). Challenges to the hypothesis of extended cognition. *Journal of Philosophy* 101, 389–428.

Rupert, Robert D. (2009). *Cognitive systems and the Extended Mind*. Oxford: Oxford University Press.

Rupert, Robert D. (2010). Extended cognition and the priority of cognitive systems. *Cognitive Systems Research* 11: 343–56.

Rupert, Robert D. (2011). Cognitive systems and the supersized mind. *Philosophical Studies* 152: 427–436.

Rupert, Robert D. (2013). Memory, natural kinds, and cognitive extension; or, Martians don't remember, and cognitive science is not about cognition. *Review of Philosophy and Psychology* 4: 25–47.

Sporns, Olaf, Dante R. Chialvo, Marcus Kaiser, & Claus C. Hilgetag. (2004). Organization, development and function of complex brain networks. *Trends in Cognitive Sciences* 8, 9: 418–425.

Talmy, Leonard. (2015). Relating language to other cognitive systems: An overview. *Cognitive Semantics* 1: 1–44.

Tononi, Giulio. (2008). Consciousness as integrated information: A provisional manifesto. *Biol. Bull.* 215: 216–242.

Tosun, Sümeyra, Jyotsna Vaid, and Lisa Geraci. (2013). Does obligatory linguistic marking of source of evidence affect source memory? A Turkish/English investigation. *Journal of Memory and Language* 69: 121–134.

Weiskopf, Daniel A. (2010). The Goldilocks Problem and extended cognition. *Cognitive Systems Research* 11: 313–323.

Wellman, Henry M., David Cross, and Julanne Watson. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development* 72: 655–84.

Wilson, Margaret. (2002). Six views of embodied cognition. *Psychonomic Bulletin and Review* 9: 625–636.

Wilson, Robert. (2004). *Boundaries of the Mind: The Individual in the Fragile Sciences.* Cambridge: Cambridge University Press.

Zerilli, John (2019). Neural redundancy and its relation to neural reuse. *Philosophy of Science* 86(5).