**Cognitive Ontology and the Search for Neural Mechanisms: Three Foundational Problems**

Mapping cognitive capacities on neural mechanisms faces three interlocking problems. First, when do different tasks elicit similar cognitive capacities? To address this *operationalization problem*, scientists often assess whether tasks engage the same neural mechanisms. To determine whether mechanisms are of the same kind, however, we need to solve the *abstraction problem* by determining which physical differences we can gloss over, and the *boundary problem* by distinguishing the mechanism from its background conditions. Solving these problems requires understanding how cognitive capacities are elicited in tasks. We argue that this ‘cycle of kinds’ is an iterative process that yields progress only incrementally.

1. **Introduction**

A fundamental goal of cognitive neuroscience is to explain the cognitive capacities that collectively make possible the mental lives of persons. To achieve this goal, we must answer two basic questions: (1) What are these capacities? And (2) How do facts about brains explain them?

In answering (2), a prevailing idea in both science and philosophy is that the brain has these definitive capacities in virtue of containing mechanisms that underlie (or mediate, or implement) those capacities (Craver 2007; Craver and Darden 2001; Bechtel 2008; Piccinini 2020). If capacities are defined in terms of functions relating input to output, the mechanism for a capacity involves the causally organized interactions of entities and activities in virtue of which the input is transformed into output. Such mechanistic explanations frequently span multiple levels of organization: The activities of and interactions composing a neural mechanism are themselves explained by lower-level mechanisms.

Question (1) is the focus of research on “cognitive ontology”, which is dedicated to regimenting the taxonomy of cognitive capacities. Fierce controversies arise both locally and globally with some frequency over how to define cognitive capacities and how to distinguish them one from the other (Janssen, Klein, and Slors 2017; Klein 2012; Anderson 2015; Danziger 1997; Poldrack 2010; Uttal 2001; Poldrack and Yarkoni 2016; Khalidi 2017; Price and Friston 2005). In these discussions, the mechanistic answer to (2) is frequently called into service: The legitimate ontology of cognitive capacities corresponds to the catalogue of distinct neural mechanisms. The correct taxonomy of cognitive capacities, that is, is anchored in the prior delineation of the brain into more or less independent mechanisms.

Here, we characterize three interlocking problems this mechanistic anchoring approach to cognitive ontology must face: The Operationalization Problem (Section 2), the Abstraction Problem (Section 3), and the Boundary Problem (Section 4). Together, they form the “Cycle of Kinds” depicted in Figure 1. The Operationalization Problem concerns a principled uncertainty about how specific experimental tasks correspond to the cognitive phenomena they are used to study. This uncertainty is frequently addressed by assessing whether the tasks engage the same or different neural mechanisms. However, we argue, the mechanistic structure of the world is not simply perceived as such but requires theoretical reconstruction to be discovered. To identify mechanisms as such, we need to abstract away from the buzzing blooming confusing of causal connections (the Abstraction Problem) and distinguish constituents of mechanisms from, e.g. background conditions (the Boundary Problem). These problems can be resolved, we argue, only by having recourse to a prior understanding of what the relevant capacities are and of how they are elicited in cognitive tasks. Thus, we arrive back at the Operationalization Problem (Figure 1). In Section 5, we argue this circularity is neither unique to cognitive (neuro)science nor especially deadly (Chang 2004); instead, we should expect progress at this key interface to be incremental, iterative, and ultimately assessed globally for a system of interrelated concepts.



*Figure 1. The Cycle of Kinds: An overview of the three problems and their interrelation for the internalist, mechanistic anchoring approach to cognitive ontology.*

1. **The Operationalization Problem**

The Operationalization Problem is faced by any project linking cognitive capacities to brain mechanisms. One cannot assess or establish such linkages without the use of tasks designed to elicit behavior reflective of the capacity in question. Such tasks provide the stimulus conditions and behavioral measures that allow one to interpret brain activities (or the absence thereof) in cognitive terms. In lesion studies, for example, subtly different measures of behavior are necessary to interpret the lesion as producing a cognitive deficit. Neurologists use clinical tasks to localize possible lesions via functional loss. Imaging studies use tasks and subtractions to activate (and localize) some capacities and not others. The experimental task engages the subject in a behavior taken to indicate the operation (or absence) of the cognitive capacity. Tasks thus anchor this integrative project.

 But how can we be sure that our task measures the capacity we think it measures? We use tasks because we cannot observe the capacities directly; we infer the capacity from task performance. This crucial choice is routinely taken for granted as a bit of the inherited practice one obtains in graduate training and post-doctoral research: We measure working memory with the n-back task or the complex span test, and spatial memory is tested in a Morris water maze or scene recognition task. But it is open to question whether the task actually measures the capacity and, if so, how well it does so. Failures at this locus produce confounded experiments and conceptual confusion (Francken and Slors 2014). Prior to cognitive neuroscience this question has been addressed through the concept of “construct validity” (see Cronbach and Meehl 1955) yet the success of this strategy has been debated. However, this discussion has been conducted primarily within psychology and is hence beyond the scope of this paper (Borsboom et al. 2009).

How must a task be related to a cognitive capacity for it to be used as an assay or measure of that capacity (Borsboom, Mellenbergh, and Van Heerden 2004; Sullivan 2010)? In adopting a task, the researcher at least tacitly embraces a set of assumptions about how the stimulus elicits the capacity in question and how the capacity drives the task-related behavior. Call the collected set of these assumptions the *model of the task* performance. This model describes in more or less detail how one thinks the task conditions (stimulus and background conditions) are transformed into task outputs (e.g., competent completion), revealing the stages and steps of a causal process, perhaps associated with specific concrete structures and systems, one must traverse if one is to perform the task.

A textbook example is the stop signal task which is routinely used as a measure of “response inhibition.” This family of tasks has been endlessly varied in its use to study this cognitive capacity and disorders thought to involve deficits in inhibitory control (Logan and Cowan 1984; Verbruggen and Logan 2008). It has also been varied to apply the task to adult, infant, and impaired humans, as well as to monkeys (Pani et al. 2018), rats (Eagle et al. 2008), and sheep (Knolle et al. 2017). In this task, a subject is instructed to perform an action (e.g., pressing a key when you see a face) unless a stop signal is presented prior to the moment of action. Researchers can vary the timing of the stop signal, for example, and determine the number of errors (in which the subject acts despite receiving the stop signal). Performance on the task is usually characterized in terms of an estimated value called the *stop signal reaction time* (SSRT), which is taken to reflect the capacity to inhibit responses (but see Bissett et al. 2020). This capacity is thought by some to be involved in any cognitive act that requires volitional control over fleeting desires (e.g., going home after two glasses of wine, or studying versus watching memes). Poor performance on such tasks is taken to indicate impulsivity, and is associated with attention deficit disorder and proneness to risk-taking and addictive behaviors (Dalley and Robbins 2017).

The precise task model for the stop signal family of tasks depends on its particular instantiation. It will include, minimally, the tendency of the subject to perform the dominant act (often acquired through pre-training), the ability to perceive the stop signal and to associate it with not acting (often acquired through training), and the ability to suppress the dominant behavioral tendency in light of that association. By shortening the time between the stop signal and the time for action, the go process finishes before the stop process, and error frequency (as well as SSRT) increases. It increases faster for those who are impaired in response inhibition. The ability to inhibit responses is both affected by the task conditions and influences response time, according to this model. That is why the task can be used to measure this capacity.

In perhaps the simplest kind of task model, the cognitive capacity in question is posited as necessary for task performance (or exhibiting the effect): Successful halting after the stop signal indicates that response inhibition capacity is intact; subjects who require greater time to stop are viewed as having less effective response inhibition. In this example of a task model, response inhibition is necessary to even perform the task, and the strength of that capacity is measured in the subject’s probability of success. More generally, what matters is that the task conditions cause the cognitive capacity to be engaged and that the engagement of the capacity is a cause of the task output in such a way that the task output can be taken as an indicator of the capacity.

 The Operationalization Problem is the problem of justifying the indicative relationship between a behavioral task and a cognitive capacity, of justifying the claim that one or more tasks measures the capacity in question (Figure 2). The Operationalization Problem arises because we cannot observe the cognitive capacity independently of our choice of task. To justify this choice, we have to provide evidence for causal facts about a cognitive capacity we cannot measure independently of the task.

When it is possible to use more than one task to measure a capacity, we might take some comfort in their consilience. But this comfort rests on blinding ourselves to the question of when any one of these tasks is measuring the capacity; we also face the new and thorny question of when two tasks measure the same capacity (Francken and Slors 2014). Typical tasks that are said to elicit response inhibition as well are the Simon task, Stroop task and the Eriksen flanker task. To give one more example, in the Simon task, trials where the spatial location of the stimulus (e.g., right) mismatches the spatial location of the required response (e.g., left) typically yield slower responses, since we usually respond to things on the right with our right hand. In other words, the unusual combination of stimulus and response locations induces a conflict between those two sources of evidence, which requires time to resolve. The task model here assumes that interference occurs in the response selection stage of decision making, and thus requires response inhibition (of the typical response) for successful task performance. So, these different tasks are all supposed to involve more or less the same cognitive capacity, known as response inhibition. Note however that it is open for discussion whether they measure *exactly* the same capacity or rather slightly different, related capacities.

The similarity among these tasks is primarily intuitive and probably based at least partly on the phenomenological sense that these tasks require distinctive mental effort (Francken and Slors 2018). To consider these tasks consilient requires a prior decision that they measure the same thing. A similar problem arises in transferring tasks across species or subjects: Does the same task model for stop signal tasks apply in humans and in sheep? Or might humans and sheep have different mechanisms controlling action? As has been noted for the study of the measurement of time (Tal 2016) and temperature (Chang 2004), we seem to face a circle of reasoning: We (at least often) are asked to justify a judgment of consilience on grounds that cannot be established independently of decisions about whether the tasks measure the same capacity.



*Figure 2. Schematic depiction of the Operationalization Problem. Different tasks elicit “the same” cognitive capacity. To what extent these capacities are sufficiently similar is something that needs to be established independently.*

Historically, operationalists had a radical solution to this problem: They *defined* cognitive capacities in terms of their tasks. On the strictest interpretation (e.g., Bridgman 1927), no two different tasks can possibly measure the same cognitive capacity. Instead, most researchers want to retain the logical distinction between cognitive capacities and task performance, if only to allow for the possibility that one and the same cognitive capacity might be involved in different tasks, such as tasks performed daily outside the laboratory.

A more plausible way of determining if two tasks measure the same capacity is by determining whether or not the two tasks involve overlapping neural mechanisms. After all, cognitive capacities are implemented by neural mechanisms, so similar neural mechanisms should indicate similar cognitive capacities. Such “reverse inference” is widely, if controversially, deployed in cognitive neuroscience (Poldrack 2006). Less controversial is the assumption that differences in neural activity observed during task performance might well indicate differences in cognitive processing. For example, McDermott et al. (2009) show that standard “laboratory tasks” for remembering, such as memorization of word lists, activate partly different regions of the brain than do “ordinary” memory tasks, such as recalling a childhood experience, though both of these tasks are routinely used in different laboratories to assay this capacity. Likewise, differential effects of local damage on performance of different tasks indicates independence of the capacities in question (as has been argued for declarative and non-declarative memory, episodic and semantic memory, etc.). Indeed, these are common strategies for lumping and splitting kinds in mechanistic sciences generally.

These strategies use neural mechanisms as the objective arbiters of cognitive similarity and the basis for revising our cognitive ontology. As we show in the next section, the judgment that two mechanisms are identical or different cannot be simply read off the causal structure of things; any definitive judgment requires definitive solutions to both the Abstraction Problem and the Boundary Problem, which we consider in turn.

1. **The Abstraction Problem**

 In order to *delineate* a cognitive capacity, we must indicate when two particular instances of the capacity are capacities of the same kind, just as to delineate a species one must indicate when two individuals belong to it.

In virtue of what does a given object or capacity belong to scientific kind? Dan Dennett, for example, asks: What makes a given chunk of matter a magnet, or a particular kind of magnet, such as a ferromagnet (1987, 43). He discusses two kinds of answer. One is broadly “externalist”: Two objects are of the same kind when they are disposed to act and interact with other things in the same ways. A magnet is a ferromagnet because it behaves like a magnet: It attracts ferrous materials, repels like poles, orients north to south, etc. On this externalist view, lodestones, ceramic magnets and electromagnets belong to the same kind. He also considers a second, “internalist,” answer: Two objects are of the same kind when they have the same or similar organizations of components. Magnets are magnets in virtue of the fact that the spins on neighboring electrons are aligned in an exchange interaction. But since the way in which this alignment induced is different in lodestones (which are possibly magnetized by lightning (Wasilewski and Kletetschka 1999), ceramic magnets (where heated ferrite powder is compacted in the presence of a magnetic field), and electro magnets (where the magnetic properties of many tiny electric currents are combined in a coil of isolated copper wire), the internalist outlook would not count them as belonging to the same kind. These two answers about magnets offer an intuitive starting point for any effort to regiment our ontology of cognitive capacities.

When are two cognitive capacities instances of the same kind? As with magnets, the externalist—or functionalist—will emphasize how the capacity acts and interacts with its environment. This strategy risks lumping things together that behave similarly but have different underlying explanations. Clever Hans gives the correct answers to math problems (etc.), but not the way mathematically trained humans do. Both humans and cephalopods have eyes and phototransducers, but cephalods have a stunning variety and diversity of photoreceptors relative to us (Kingston et al. 2015). So, there is a temptation to look internally for relevant differences: The mechanisms underlying these processes further distinguish which functionally similar processes belong in the same kind. Hans answers questions by tracking the subtle head movements of the questioner, not from memory and reasoning. Human eyes have different receptors for different colors, but cephalopods have receptors that only track differences in the brightness of light. However, when cephalopod receptors are placed under cells that can change color, so-called chromatophores, they are able to detect differences in color with the same brightness by changing the color of these chromatophores (Godfrey-Smith 2016, 77–79; Desmond Ramirez and Oakley 2015). If (behaviorally) similar processes turn out to have different underlying mechanisms, then the internalist approach to delimiting kinds of capacity enjoins us to split the kind, i.e., to consider each as an instance of a distinct cognitive kind.

The internalist approach is very common and familiar in scientific arguments for revising cognitive ontology (or “reconstituting the phenomenon”, see Bechtel and Richardson 2010). To choose just one example among many, Kok and colleagues study the phenomenon “prediction.” They distinguish between attention and prediction, despite the fact that they have similar facilitatory effects on behavior, on the grounds that they have different effects on neural activation in visual cortex (Kok et al. 2012). Recently, Teufel and Fletcher propose to split the taxonomy even further, distinguishing two types of predictions (constraints and expectations) on the grounds that they have different mechanistic/implementation explanations, despite their computational/functional similarities (Teufel and Fletcher 2020).

This is an important kind of ontological progress. But notice that it is predicated on an understanding of when two mechanisms are mechanisms of different kinds. If we are to follow the rule that we should split higher-level kinds when we discover that the same phenomenon is produced by two distinct kinds of mechanism, we need a further set of rules telling us when two mechanisms belong to the same or different kinds. This is the same type of question we started this section with. When are two mechanisms mechanisms of the same kind? Perhaps when they have the same kinds of parts, activities and organizing relations. But when are parts, activities and organizational relations of the same kind?

 To judge two mechanisms to be similar or different, we have to decide on an appropriate grain of abstraction for describing those mechanisms. No two instances of ‘the same’ cognitive mechanism are physically—cell for cell, atom for atom—identical. There is inevitable biological variation from one person to the next, and even one instance to the next in the same person. To see any two mechanisms as of the same kind is necessarily to abstract away from these internal differences. Further, a physical difference between two mechanisms is also a causal (and so likely also a functional) difference between them. So, when we lump two particular mechanisms under the same kind despite causal differences, we necessarily gloss over causal differences between them. Alternatively, if we assign them to different kinds on the basis of only minor differences, every mechanism instance becomes a kind unto itself, and the concept of a scientific kind ceases to be useful for putting like with like. There appears to be no uniquely correct degree of abstraction for describing any given system. Sometimes the differences matter; sometimes they do not (Figure 3). The unfiltered causal structure of the world lacks the resources required to specify the appropriate degree of abstraction and so to specify on its own when internal differences do and do not warrant splitting the kind. This is what we label the Abstraction Problem.



*Figure 3. Schematic depiction of the Abstraction Problem. Two neural mechanisms are abstracted twice. On the first degree of abstraction, the two mechanisms are not of the same kind. When abstracted further, however, both fall under the same kind of mechanism.*

The practically-minded might not find the abstraction problem all that perplexing; it is simply a reason to be a pluralist about kinds, especially in the special sciences. Pluralists acknowledge that the world contains many overlapping and at times cross-cutting kinds. Pluralism is well suited to the wide range of our actual and possible practical needs and to the character of contemporary science. The pluralist will insist that the boundaries of kinds are not completely arbitrary—as radical constructivists might hold. The legitimate causal kinds have to respect the causal structure of things; but that causal structure can be described in many ways (abstracting more or less, and here rather than there) each yielding a possibly legitimate way of carving the taxonomy of kinds, depending on one’s needs. For instance, instead of splitting, it might make more sense to lump different types of predictive brain mechanisms if our focus is on varieties of behavior serving a similar goal (de Lange, Heilbron, and Kok 2018). But different categorizations may be more or less useful in addressing competing needs and interests (see Chang 2004). This form of principled pluralism, in our view, is unobjectionable.

Once we take this pluralist implication on board, it turns out that two mechanisms will be mechanisms of the same kind when we see no relevant differences between them. But relevant to what? Well, relevant to the phenomenon we hope to control, explain, or predict. But notice that we have now bitten our own tails. The causal structure of the world was supposed to tell us which phenomena to include or exclude in the mechanisms underlying cognitive capacities. But the judgment of which causal structures count as relevant parts of those mechanisms depends on how we have specified the capacity to be explained and the type of explanation required— a conceptual decision made at the beginning rather than discovered within the causal order of things. This, of course, is directly the opposite of the direction of fit the internalist strategy of kind delineation exploits: Whether two mechanisms are of the same kind depends on what phenomenon they are called upon to explain and how that phenomenon is characterized.

 The hope that the Operationalization Problem can be solved by identifying cognitive capacities with neural mechanisms and sorting kinds of mechanisms faces the equally fraught challenge of sorting kinds of mechanism: To decide when two tasks measure the same capacity, we appeal to the sameness of underlying mechanisms, but sameness of underlying mechanisms depends on sameness of capacity, and judgments of the latter depend, as we have argued, upon how capacities are operationalized in tasks. This is one loop in the cycle of kinds.

1. **The Boundary Problem**

In addition to the Abstraction Problem, the internalist strategy for fixing cognitive kinds also faces a Boundary Problem. The Boundary Problem is the problem of saying which parts are in the mechanism and which are not. A solution to the Boundary Problem is required, first, to say where one mechanism ends, and another begins—both in space and in time. This problem might arise in a sequential mechanism, such as memory: Where does the encoding mechanism end and the storage mechanism begin? Secondly, a solution to the boundary problem is required, to distinguish mechanisms from their environments and background conditions. Why are some of the entities, activities, and organizational features in the mechanism (in the relevant causal structure) and others not? While this problem might be understood both in terms of spatial containment (what is “in” the mechanism) and in terms of temporal inclusion (what lies “between” input and output (Prychitko 2019)) the more fundamental matter here is relevance (Craver 2007). To be in the mechanism is to be relevant to how it works.

The Boundary and Abstraction Problems are not completely independent; what counts as a component of a given mechanism may sometimes depend on how far we have abstracted away from certain causal difference between instances of ‘the same’ mechanism. Yet these problems are logically distinct. Return to the magnet example: The Abstraction Problem is the problem of deciding whether lodestones, ceramic magnets and electromagnets all belong to the same kind. The Boundary Problem is the problem of determining whether the power source of an electromagnet is part of the magnet (without power the thing will not attract iron and hence not be a magnet) or merely a background condition for the magnet to function or, perhaps more problematically, whether the ambient temperature is part of the magnet, given that its attractive force changes with temperature.

The problem for the internalist answer to the question of cognitive kinds arises from the fact that the world does not come neatly packaged into mechanisms. The unfiltered causal structure of the world is busy, buzzing confusion. It takes considerable insight to carve away enough of the irregularities that allow one to see the kind of orderly mechanistic structure depicted in the call-out boxes of our biology textbooks (which always requires abstracting away from the unfiltered causal structure of the world—hence the connection between the Abstraction Problem and the Boundary Problem). This is why mechanism-discovery is an achievement and not merely a matter of reading off what causes what. When we consider mechanical effects, diffusion of molecules, heat transfer, metabolic exchange, waste production, electrical effects, etc., the causal structure of our bodies bewilderingly complex and interwoven. And these are simply the occurrent mechanisms. One might appeal, in addition to physiological mechanisms at work in the here and now, to regulatory mechanisms that keep these first-level mechanisms operating, developmental mechanisms that put them in place, and evolutionary mechanisms that brought them out in the first place. Which of these ways of situating a biological trait in the causal structure of the world is uniquely correct for determining the appropriate cognitive ontology? Which of these mechanisms do we foreground and background in our search for kinds (Figure 4)?



*Figure 4. Schematic depiction of the Boundary Problem. Different ways of distinguishing between mechanism and background conditions imply different mechanisms.*

The problem is that one can be led to lump or split differently depending on which entanglements one decides to include in the mechanism. To use prediction as an example again, does the mechanism of a prediction include both the “sending” brain area (sending a representation of predicted input to another area) as well as the “receiving” brain area (where the prediction has its consequences)? What if there is no sending area, such as in the case of the constraints-variant of predictions discussed above (Teufel and Fletcher 2020)?

This, again, emphasizes a degree of liberty in this mechanistic endeavor (allowing for many possible, equally correct reasons for lumping and splitting depending on which mechanism one attends to or how one attends to it). And it again illustrates the fact that which of these mechanisms is relevant for the purposes of building our taxonomy depends on what we want our taxonomy to do. And here is the key point: The direction of fit is opposite of what we were looking for when we wanted the mechanistic structure of the world to make our taxonomic decisions definitive.

In practice in cognitive neuroscience, the Boundary Problem, while not often explicitly recognized as such, is reduced by the methodological practice of “subtraction”. In order to find the neural activity associated with a particular cognitive capacity, one has to subtract (or contrast) the neural activity in a control condition of the task (not involving the cognitive capacity) from the neural activity in the experimental condition of the task (involving the cognitive capacity; see also section 2). In this way, the boundaries of a neural mechanism are constrained by the task design, since irrelevant factors that could cause the neural activity (or confounds) are supposed to be left out of the equation.

For example, when one aims to identify the neural mechanism of motion perception, two task conditions are created: In the experimental condition, coherently moving dots are presented, while in the control condition static dots or randomly moving dots are presented. When such a task is combined with fMRI measurements, part of the neural mechanism of motion perception—occipital area V5 (human MT+)—can be functionally delineated upon contrasting experimental and control conditions. In this way, the boundary between V5 and e.g., neighboring area V4 (which is color but not motion-sensitive) can be found. Note though that these boundaries are still dependent on statistical choices related to brain measurement (e.g., BOLD activation threshold for voxels to be included) and methodological choices (e.g., choosing the appropriate control condition) and assumptions (e.g., pure insertion, see Friston et al. 1996).

So we have again completed the Cycle of Kinds: Whether we have one or two kinds of mechanism will depend on which task conditions we have taken as paradigmatic for the purposes of studying it and which control conditions properly exclude the appropriate confounds.

1. **Progress by moving through the Cycle of Kinds**

In the above sections, we argue that the cognitive ontology project, i.e., the search for the mapping between cognitive capacities and the brain parts and processes, faces three interlocking problems. First, we lack an independent ground for associating a given cognitive task with a given cognitive capacity. To avoid this Operationalization Problem, we might look at the neural realizers of these capacities and compare the neural activity in different tasks that are supposed to invoke the same capacity. However, this move leads us to two further problems. First, we cannot determine whether two cognitive capacities are of the same kind by looking at their neural mechanisms only. For we cannot decide when two mechanisms are mechanisms of the same kind by merely looking at their internal causal structure. Looking more or less abstractly at the mechanisms, we might lump and split one and the same mechanism differently. Furthermore, since neural mechanisms do not come neatly packaged for us, we have to decide which components and activities are part of the mechanism and which are instead e.g., background conditions that do not fall within the mechanistic boundaries. Indeed, although cognitive neuroscientists try to escape the Boundary Problem with their experimental controls and subtractions, in fact this only leads us through the Cycle of Kinds, back to the Operationalization Problem.

If the above analysis is correct, there is something circular in the search for a correct cognitive ontology—at least when look to the neural mechanisms underlying cognitive capacities to make our ontological decisions for us. But is this approach *viciously* circular? It would be if the Operationalization Problem we end up with when trying to solve the Boundary Problem and/or the Abstraction Problem is exactly the same problem we wanted to solve by letting neural mechanisms make taxonomic decisions for us. But that need not be the case. Although deciding on the boundaries and abstraction level of mechanisms cannot be done without having recourse to functional, task-related considerations about cognitive capacities, this does not mean that the tasks involved in these considerations are the same original tasks we started out with. The part of the Boundary and Abstraction Problem that needs to be settled by going back to functional, task-related considerations about cognitive capacities, need not overlap completely with the functional-level question we started out with when deciding to let neural mechanisms make taxonomic decisions. The cycle can be iterative, fine-tuning answers to each problem under the constraint of the others, without being viciously circular (see Chang 2004).

Progress in this iterative practice involves reducing the incongruities among tasks, cognitive ontologies, and our understanding of mechanisms. We cycle from a functional description of capacities, to the neural implementation, to the task choice and task model, and back again. Repetition of this circular process need not involve stagnation but may yield increasingly more refined functional concepts and informed decisions about what to count as and in a mechanism. It would, after all, be a tremendous achievement to bring our solutions to these three component problems in the cycle of kinds (tasks, capacities/functional roles, and mechanisms) into alignment for any given practical project. Such alignment is the stop signal in the search for cognitive kinds, halting the cycle of accommodation.

Our proposal aligns with key features of Boyd’s “homeostatic property cluster” theory of natural kinds (Boyd 1989). According to Boyd, externalist (functionalist) and internalist (mechanistic) approaches to kind delineation can be conjoined: An object or capacity is a member of a kind in virtue of the similarities both in how they regularly behave (the externalist/functionalist answer) and in the mechanisms that explain why they regularly behave that way. Boyd’s naturalistic “principle of accommodation” enjoins us to split kinds whenever we find inductively relevant differences; and these differences might be found either in the cognitive function or in the mechanism. Boyd’s view expresses concisely the idea that scientists should populate their models and theories with kinds that best systematize our knowledge of the world’s causal structure (Salmon 1984) and that therefore offer the most “bang for the buck,” maximizing predictive leverage and instrumental control in the most economical way (Strevens 2008). If one fails to recognize real distinctions between kinds, one’s model or theory necessarily suffers in some prediction, instrumental application, and explanatory task (reducing the bang). On the other hand, if one distinguishes in one’s models and theories functionally and mechanistically identical kinds of objects and capacities, one necessarily introduces predictively and instrumentally irrelevant and otiose detail (increasing buck). Boyd is not explicitly concerned with the experimental tasks used to operationalize the capacities in question. But if we are to understand the process by which our understanding of cognitive ontology will be sharpened, we must include the tools and methods by which we induce and measure these capacities into account. They cannot simply be taken for granted but are themselves part of the intellectual background that embodies are ontology in our material practices. These practices are especially worth of attention in the cognitive sciences in part because there is as much difference of opinion (if not open controversy) over the adequacy of different tasks as there is about the ontology. The ability to construct any coherent and economical picture of how tasks, capacities, and mechanisms relate, is itself a scientific achievement of that deserves to be taken seriously.

The iterative cycle we describe here is not unique to cognitive neuroscience but has analogues in even more “basic” sciences, especially in the early days of concept formation. For example, Hasok Chang describes a similar process of “epistemic iteration” for the concept of temperature, its mechanisms, and its measures in the study of temperature in the Eighteenth and Nineteenth Centuries (Chang 2004). Knowledge accumulation is possible even where we can point to no secure and indubitable foundations, but it is measured in its coherence, its predictive adequacy, and its stability. The feeling of hotness acts as the first, intuitive and roughly-hewn touchstone guiding the search for the thing “heat” that might explain it, even when it is utterly unclear what, if anything, the feeling of hotness detects in things. Likewise, our intuitive interactions with remembrance serve as the anchor point for scientific investigation, which involves the development of tasks, controls, and ontologies that can lead us productively away from that intuitive home. It is therefore not a conceptual failing of contemporary cognitive science that its taxonomy of kinds is in flux; this is in keeping both with a healthy pluralism, as described above, and with the way that other sciences have developed. A certain looseness in kind definitions and matters of ontology, especially when such matters are (most would agree) far from settled, provides space for scientific research programs to live and breathe. Yet their work, if it is to make the iterative progress seen in other sciences, must be guided by an underlying recognition of the task: To make progress by moving through the cycle of kinds, bringing our tasks, concepts, and explanations into stable equilibrium.

So here is a practical take-home message of these philosophical investigations for cognitive neuroscience: We have argued that progress in cognitive ontology will be iterative and cyclic, even in the best of conditions. In practice, this means that there is an important, additional stage after data analysis and interpretation: Going back to, and possibly correcting or rejecting, the cognitive ontology, the external (functional) description, one’s understanding of the mechanisms, and one’s model of the task models (see also Krakauer et al. 2017). The current emphasis on neural mechanisms should therefore be balanced by a renewed interest in the mind and behavior and how we study them experimentally. The cycle of kinds is deep and unbroken, but it is the ineliminable tie that binds our experimental practices to our ontological categories for parsing mind and brain alike.

**References**

Anderson, Michael L. 2015. “Mining the Brain for a New Taxonomy of the Mind.” *Philosophy Compass* 10 (1): 68–77.

Bechtel, W. 2008. *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. Taylor & Francis Group.

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

Bissett, P.G., H.M. Jones, R.A. Poldrack, and G.D. Logan. 2020. “Severe and Pervasive Violations of Independence in Response Inhibition Tasks.” *PsyArXiv*.

Borsboom, D., A.O.J. Cramer, R.A. Kievit, A. Zand Scholten, and S. Franic. 2009. “The End of Construct Validity.” In *The Concept of Validity*, 135–70.

Borsboom, D., G.J. Mellenbergh, and J. Van Heerden. 2004. “The Concept of Validity.” *Psychological Review*.

Boyd, Richard. 1989. “What Realism Implies and What It Does Not.” *Dialectica* 43 (1–2): 5–29.

Bridgman, P.W. 1927. *The Logic of Modern Physics*. New York: Macmillan.

Chang, H. 2004. *Inventing Temperature: Measurement and Scientific Progress*. New York: Oxford University Press.

Craver, C.F. 2007. *Explaining the Brain*. Oxford: Oxford University Press.

Craver, C.F., and L. Darden. 2001. “Discovering Mechanisms in Neurobiology: The Case of Spatial Memory.” In *Theory and Method in the Neurosciences*, edited by P.K. Machamer, R. Grush, and P. McLaughlin, 112–37. Pittsburgh: University of Pitt Press.

Cronbach, Lee J., and Paul E. Meehl. 1955. “Construct Validity in Psychological Tests.” *Psychological Bulletin* 52 (4): 281–302.

Dalley, Jeffrey W., and Trevor W. Robbins. 2017. “Fractionating Impulsivity: Neuropsychiatric Implications.” *Nature Reviews Neuroscience* 18 (3): 158–71.

Danziger, Kurt. 1997. *Naming the Mind: How Psychology Found Its Language*. London: SAGE Publications.

Dennett, Daniel C. 1987. *The Intentional Stance*. Cambridge: MIT Press.

Desmond Ramirez, M., and Todd H. Oakley. 2015. “Eye-Independent, Light-Activated Chromatophore Expansion (LACE) and Expression of Phototransduction Genes in the Skin of Octopus Bimaculoides.” *Journal of Experimental Biology* 218 (10): 1513–20.

Eagle, Dawn M., Christelle Baunez, Daniel M. Hutcheson, Olivia Lehmann, Aarti P. Shah, and Trevor W. Robbins. 2008. “Stop-Signal Reaction-Time Task Performance: Role of Prefrontal Cortex and Subthalamic Nucleus.” *Cerebral Cortex* 18 (1): 178–88.

Francken, Jolien C., and Marc Slors. 2014. “From Commonsense to Science, and Back: The Use of Cognitive Concepts in Neuroscience.” *Consciousness and Cognition* 29 (October): 248–58.

———. 2018. “Neuroscience and Everyday Life: Facing the Translation Problem.” *Brain and Cognition* 120: 67–74.

Friston, K. J., C. J. Price, P. Fletcher, C. Moore, R. S.J. Frackowiak, and R. J. Dolan. 1996. “The Trouble with Cognitive Subtraction.” *NeuroImage* 4 (2): 97–104.

Godfrey-Smith, P. 2016. *Other Minds. The Octopus, the Sea, and the Deep Origins of Consciousness*. New York: Earrar, Strauss and Giroux.

Janssen, Annelli, Colin Klein, and Marc Slors. 2017. “What Is a Cognitive Ontology, Anyway?” *Philosophical Explorations* 20 (2): 123–28.

Khalidi, Muhammad Ali. 2017. “Crosscutting Psycho-Neural Taxonomies: The Case of Episodic Memory.” *Philosophical Explorations* 20 (2): 191–208.

Kingston, Alexandra C.N., Trevor J. Wardill, Roger T. Hanlon, and Thomas W. Cronin. 2015. “An Unexpected Diversity of Photoreceptor Classes in the Longfin Squid, Doryteuthis Pealeii.” *PLoS ONE* 10 (9): e0135381.

Klein, Colin. 2012. “Cognitive Ontology and Region- versus Network-Oriented Analyses.” *Philosophy of Science* 79 (5): 952–60.

Knolle, Franziska, Sebastian D. McBride, James E. Stewart, Rita P. Goncalves, and A. Jennifer Morton. 2017. “A Stop-Signal Task for Sheep: Introduction and Validation of a Direct Measure for the Stop-Signal Reaction Time.” *Animal Cognition* 20 (4): 615–26.

Kok, Peter, Dobromir Rahnev, Janneke F.M. M Jehee, Hakwan C. Lau, and Floris P. De Lange. 2012. “Attention Reverses the Effect of Prediction in Silencing Sensory Signals.” *Cerebral Cortex* 22 (9): 2197–2206.

Krakauer, John W., Asif A. Ghazanfar, Alex Gomez-Marin, Malcolm A. MacIver, and David Poeppel. 2017. “Neuroscience Needs Behavior: Correcting a Reductionist Bias.” *Neuron* 93 (3): 480–90.

Lange, Floris P. de, Micha Heilbron, and Peter Kok. 2018. “How Do Expectations Shape Perception?” *Trends in Cognitive Sciences* 22 (9): 764–79.

Logan, Gordon D., and William B. Cowan. 1984. “On the Ability to Inhibit Thought and Action: A Theory of an Act of Control.” *Psychological Review* 91 (3): 295–327.

McDermott, Kathleen B., Karl K. Szpunar, and Shawn E. Christ. 2009. “Laboratory-Based and Autobiographical Retrieval Tasks Differ Substantially in Their Neural Substrates.” *Neuropsychologia* 47 (11): 2290–98.

Pani, Pierpaolo, Franco Giarrocco, Margherita Giamundo, Roberto Montanari, Emiliano Brunamonti, and Stefano Ferraina. 2018. “Visual Salience of the Stop Signal Affects the Neuronal Dynamics of Controlled Inhibition.” *Scientific Reports* 8: 14265.

Piccinini, G. 2020. *Neurocognitive Mechanisms*. Oxford University Press.

Poldrack, Russell A. 2010. “Mapping Mental Function to Brain Structure: How Can Cognitive Neuroimaging Succeed?” *Perspectives on Psychological Science* 5 (6): 753–61.

Poldrack, Russell A., and Tal Yarkoni. 2016. “From Brain Maps to Cognitive Ontologies: Informatics and the Search for Mental Structure.” *Annual Review of Psychology* 67: 587–612.

Poldrack, Russell A. 2006. “Can Cognitive Processes Be Inferred from Neuroimaging Data?” *Trends in Cognitive Sciences* 10 (2): 59–63.

Price, Cathy J., and Karl J. Friston. 2005. “Functional Ontologies for Cognition: The Systematic Definition of Structure and Function.” *Cognitive Neuropsychology* 22 (3–4): 262–75.

Prychitko, Emily. 2019. “The Causal Situationist Account of Constitutive Relevance.” *Synthese*.

Salmon, W. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.

Strevens, M. 2008. *Depth: An Account of Scientific Explanation*. Cambridge, MA: Harvard University Press.

Sullivan, J.A. 2010. “Reconsidering ‘spatial Memory’ and the Morris Water Maze.” *Synthese* 177 (2): 261–83.

Tal, Eran. 2016. “Making Time: A Study in the Epistemology of Measurement.” *British Journal for the Philosophy of Science* 67 (1): 297–335.

Teufel, Christoph, and Paul C. Fletcher. 2020. “Forms of Prediction in the Nervous System.” *Nature Reviews Neuroscience* 21: 231–242.

Uttal, William R. 2001. *The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain.* Cambridge: MIT Press.

Verbruggen, Frederick, and Gordon D. Logan. 2008. “Response Inhibition in the Stop-Signal Paradigm.” *Trends in Cognitive Sciences* 12 (11): 418–24.

Wasilewski, Peter, and Günther Kletetschka. 1999. “Lodestone: Natures Only Permanent Magnet-What It Is and How It Gets Charged.” *Geophysical Research Letters* 26 (15): 2275–78.