12 Implications from the Philosophy of Concepts for the Neuroscience of Memory Systems

Anna Leshinskaya and Enoch Lambert

12.1 Introduction

It is a common assumption in cognitive science that the mind is furnished with elements called concepts. These constituents allow us to track observations and form thoughts about the same thing: if one believes that apples are tasty and observes that that's an apple, one can relate those propositions appropriately (Carey, 2009; Margolis & Laurence, 1999; Murphy, 2002; Rey, 1983; Smith & Medin, 1981). Psychologists seek to explain behavior hypothesized to rely on concepts, such as categorization, inference, and language comprehension. Cognitive neuroscientists seek to identify functionally distinct components of the mind and brain via what those components represent and compute (their characteristic features). Here, we ask how cognitive neuroscientists could probe conceptual representation as such a putative functional component.

Concepts are treated as a distinct topic of study within cognitive neuroscience (Hoffman, McClelland, & Lambon-Ralph, 2017; Mahon, 2015; A. Martin, 2007; Patterson, Nestor, & Rogers, 2007), but this literature is mostly silent on the question of what distinct cognitive or functional characteristic features concepts are thought to have. To be a well-defined empirical entity, concepts or conceptual memory must have operational criteria sufficiently precise to allow one to identify and measure them distinctly from other kinds of mental representations—notably, from other things stored in long-term memory.[1] Here, we seek out such criteria. Ideally, these criteria remain aligned with the findings and theoretical frameworks in psychology of concepts. Failing to find such criteria might lead one to reconsider what it means to have a cognitive neuroscience of concepts and whether this term is useful. This is the issue at stake here.

In the first half of this chapter, we review existing approaches in cognitive neuroscience for differentiating among kinds of long-term memories by virtue of their characteristic and measurable features, some of which could serve as candidates for operational criteria for concepts. For example, we discuss the difference between memories that aggregate across experience and those that pick out unique experiences,[2] and the difference between memories that are specific to a certain sensory channel and those that are channel invariant. However, we encounter a repeated problem: these distinctions do not neatly map onto the notion of concepts as it is often used in the psychology and neuroscience literature, and these criteria permit cases that would only reluctantly be called concepts by most researchers and laypersons (e.g., participants in psychology experiments). In our analysis of these cases, we propose that to account for both lay and researcher intuitions about concepts, we must appeal to a notion of sharedness: whether a memory refers to information that is also likely to be represented by others.

In the second half, we grapple with whether sharedness should be considered important in the scientific treatment of concepts rather than only describing a folk psychological intuition. In so doing, we raise the possibility that folk-psychological intuitions about memory may in fact play a role in a scientific theory of the brain, to the extent that they themselves may influence how memories are stored. Lastly, we propose an empirical path forward to settling this issue.

12.2 Preliminaries: Concepts as Representational Types

We take as a premise that concepts are (at least) one type of mental representation and are distinct from certain other distinguishable types. This view is a pillar for work in developmental psychology, which uses it to distinguish among hypotheses regarding infant cognition (Carey, 2009; Mandler, 2004) and in philosophical treatments, which discuss features distinguishing concepts from other kinds of representations (Margolis & Laurence, 1999; Millikan, 2017; Rey, 1983). It has also been a fundamental guiding framework in cognitive neuropsychology, which proposes distinct memory systems to account for patterns of impairment following neural damage (Caramazza et al., 1990; Eichenbaum & Cohen, 2001; Schacter & Tulving, 1994; Tulving, 1972, 1984; Warrington, 1975; Warrington & Taylor, 1978). For the

distinctions among memory types that we delineate below, we draw on both theoretical and empirical justifications.

However, it must be noted that it is not a universally accepted premise in cognitive neuroscience that concepts are a type of mental representation. A different view is that concepts are a type of stimulus or task. This tradition studies representations of concepts, that is, mental representations about things with certain properties, rather than conceptual representations, that is, mental representations with particular properties (for reviews: Binder et al., 2009; Hoffman et al., 2017; A. Martin, 2007; McRae & Jones, 2013). In this representation of concepts view, there is no meaningful distinction among different representations evoked by the same "conceptual" stimulus. Thus, studying conceptual processing involves measuring the brain as it engages in processing the meaning of a word or picture relative to not accessing its meaning or relative to meaningless stimuli. For example, neural activity is measured as participants read and understand words relative to counting how many times the letter E appears in those words or reading pseudowords (Binder et al., 2009; Renoult, Irish, et al., 2019). Relatedly, different categories of words or images are contrasted (e.g., animals vs. tools), and the result is taken as a reflection of how animal and tools concepts are represented (A. Martin et al., 1996; Watson et al., 2013). All of the mental machinery engaged is relevant to the question of how we represent concepts, since the interest is in the set of processing mechanisms associated with certain classes of stimuli, not representations with certain qualities or of certain kinds.

In our view, a single stimulus—such as the word "apple"—can (at least conceivably) engage a multitude of different kinds of representations: we might retrieve a mental image of a bright red apple, knowledge of how apples grow, the lexical knowledge that its plural is "apples," a vivid childhood memory of picking apples with grandma, and even a salivation response in anticipation of eating one. Only some of these are conceptual. Simultaneously, a similar representation or process can be evoked by different stimuli: both the word apple and a classically conditioned auditory tone could evoke the anticipation of a reward or a salivation response. In our view, stimulus type cannot be assumed to provide a good classification scheme for the functional components of the mind and brain. Rather, it is an empirical matter to determine how mental representations are different or alike, which can be resolved by measuring properties of those mental

representations. As we hope to make clear throughout the chapter, there are theoretical and empirical justifications for our view. Thus, we take as a starting point that there exists an important question regarding the different kinds of things stored in long-term memory, even if evoked by the same meaningful stimulus, and the questions we raise are how we should distinguish among them using the methods of cognitive neuroscience, and which ones best map on to the psychological notion of concept.

12.3 Approaches to Distinguishing among Types of Memory

Many cognitive neuroscientists have indeed taken up the task of distinguishing among types of long-term memory by virtue of their characteristic properties as measurable with the tools of neural recording or imaging. It is based on this work that we seek candidates for conceptual aspects of memory. We take the term "semantic memory" to be roughly equivalent to conceptual memory, and take concepts to be elements within semantic or conceptual memory.

We begin by reviewing three major well-operationalized criteria that have been used to distinguish among types of long-term memories: (1) experiential scope, or the extent to which a representation aggregates over specific experiences; (2) channel specificity, that is, whether a representation is specific to a sensory channel of input; and (3) its similarity profile, that is, whether a representation tracks closely with physical stimulus similarity or departs from it. We consider the idea that these distinctions can serve to operationalize the idea of conceptual or semantic representations, supplying the kind of operational definition cognitive neuroscience needs to study concepts as one or more functional neural components. Specifically, we consider the idea that conceptual representations could be ones that aggregate over more experience, are not specific to a sensory channel, and can depart from physical similarity space.

However, we also find that cognitive neuroscientists do not always use the terms "concept" or "semantic memory" to describe the components of memory that these approaches pick out, and that these operational definitions do not align with the way these terms are used in cognitive neuroscience or in cognitive science more broadly. Indeed, these criteria, alone and in combination, allow cases that are not typically considered concepts. While this might seem like only a terminological issue, it is in fact a deeper ontological one about what entities exist in the mind and how we should

catalog mental phenomena. In the final section, we consider the latent intuition likely underlying these terminological choices, and whether it merits a revision to how memory systems are studied.

12.3.1 Experiential Scope

Suppose I believe that Szechuan peppers are spicy. How do I know that? Perhaps someone told me this once at a Chinese restaurant, pointing out the peppers on a plate, and I recall this specific episode vividly from a first-person perspective. Or perhaps I sampled them on my own and came to the conclusion that this must be generally true. Or perhaps it is a fact I know from repeated experiences, without recalling any individual experience. The difference between recalling unique personally experienced episodes and representations that are formed by aggregating over many such episodes has been an important distinction in cognitive neuroscience and neuropsychology.

Researchers have used a number of approaches to isolate memory of unique episodes (often termed "episodic memory") from our more general knowledge. Neuropsychological studies find that hippocampal damage disproportionately impairs patients' ability to recall individual events vividly, but that such patients typically retain general factual knowledge (Manns, Hopkins, & Squire, 2003; Mishkin, 1997; Nadel & Moscovitch, 1997; Schacter & Tulving, 1994; Sekeres, Winocur, & Moscovitch, 2018; Squire, Knowlton, & Musen, 1993; Tulving, 2002; Winocur et al., 2010). However, for recall of naturalistic lived experiences, these two kinds of memory can be difficult to isolate experimentally, as the opening vignette attests: when we don't actually know how an individual came to form their belief or what information allows them to attest it, it is difficult to know to what extent it encodes a single episode or aggregates across repeated experiences. To address this issue, one can distinguish individual from aggregated aspects of memory by probing learning in a controlled lab setting and recording neural activity in response to retrieving it. One can define episodic memory as those aspects of a memory that refer to unique aspects of a specific episode, and other kinds of memory as referring to regularities across many of them.

For example, suppose that one encountered Szechuan peppers on ten distinct occasions, and each of those times they were spicy but also each time served alongside different side dishes. Once, they were served with a yogurt dish—a memorable experience for the relief it gave to the palate afterwards. Recalling that particular individual pairing of Szechuan peppers

with yogurt would be considered episodic because it is a unique encounter requiring individuating that experience from others, which did not feature yogurt. Recalling the regularity that Szechuan peppers are spicy without retrieving any individual episode but only by recalling aggregated experience is considered something else—something with a larger "experiential scope" (in the terminology used here—that is, the extent to which a representation aggregates over specific experiences).

There are both theoretical and empirical motivations for seeing the size of experiential scope as an important distinction. Some research finds that patients with episodic deficits appear able to learn some kinds of new representations that aggregate across repeated presentations of stimuli, though possibly in different ways than controls (Knowlton, Ramus, & Squire, 1992; Knowlton, Squire, & Gluck, 1994; Myers et al., 2003; Tulving, Hayman, & Macdonald, 1991; Verfaellie, Koseff, & Alexander, 2000; cf. Manns et al., 2003). Behavioral studies with nonhuman animals also support such distinctions. In an elegant approach, mice were given numerous trials of a water-maze task, in which they were placed in a pool too deep to stand and had to find a platform on which to rest (Richards et al., 2014). On each individual trial, the platform was in a different location, centered on a coordinate mean. On test trials, in which there is no platform, search behavior could be measured as reflecting the search for specific locations previously experienced or to the coordinate mean (which did not characterize any specific prior platform but rather the central tendency summarizing all of their prior experiences). On day 1 after learning, behavior was best characterized as reflecting specific locations experienced during learning, but after thirty days, mice went to their Cartesian average location, suggesting that memories might be transformed over time from unique to aggregated. Furthermore, neuroimaging work with humans finds that the hippocampus tends to represent episode-unique information, while other areas are more likely to aggregate (Dimsdale-Zucker et al., 2018; Tompary & Davachi, 2017; Wiggs, Weisberg, & Martin, 1999; but cf. Schapiro et al., 2016).

Finally, there are reasons to understand these as theoretically divergent functions of memory. One allows you to recall exactly what happened and when, while the other allows you to acquire a generalizable model of the world. There are compelling computational demonstrations indicating that the functions of storing episode-unique versus aggregated information inherently trade off in the kind of learning algorithms they require, and

are thus best handled by at least partially distinct learning and memory systems (Eichenbaum, 2004; McClelland, McNaughton, & O'Reilly, 1995; O'Reilly & Rudy, 2001; Winocur et al., 2010). This is because one must use a slow learning rate to capture generalities that hold across experiences (not take any individual sample too seriously), which is inherently at odds with the fast learning rate that is required to recall the distinguishing features of a single episode. Having distinct systems with different learning rates allows us to learn both the rule and the exceptions governing experience.

The question we return to is whether aggregated memories are the right operationalization for concepts. Does the notion of concept or semantic memory from cognitive science map neatly onto the aggregated versus unique distinction used in neuroscience? On one hand, the idea of aggregation dovetails with the psychological literature on categorization, which shows that we tend to categorize entities on the basis of the typical features or characteristics of their category, being highly sensitivity to probabilities of a certain feature across encounters with that type (Rips, Shoben, & Smith, 1973; Rosch, 1975; for reviews, see Smith & Medin, 1981, and Murphy, 2002). For example, we readily judge that if something has wings, it probably flies but doesn't swim (even though penguins are an exception). Feature correlations across categories predict patterns of data on development and deterioration in semantic dementia. For example, categories that have more shared properties with their semantic neighbors are more robust to deterioration and yield different kinds of naming errors (Hodges & McCarthy, 1995; Rogers et al., 2004).

On the other hand, there are reasons to think that concepts are a more specific kind of representation than just any aggregate memory. Classically, Smith and Medin (1981) argued that psychological theories should distinguish between heuristics we use to identify members of a category ("identification procedures") and those we use to define inclusion in the category ("conceptual cores"). For example, we might use hairstyles and clothes to judge gender quickly, but we do not believe that these are defining or essential—even if they are extremely prevalent and diagnostic. The fact that participants can represent both of these facts simultaneously about a single category was elegantly demonstrated by Armstrong, Gleitman and Gleitman (1983), who showed that participants judge that some instances of the category "even number" are somehow better examples than others (e.g., 2 is better than 34), despite knowing the definition. Likewise, we might use

appearance in Chinese restaurants to decide if an ambiguous vegetable is a Szechuan pepper, but resist the idea that being in a Chinese restaurant is important for it being one (Prasada & Dillingham, 2009). We may not even know what those defining features are, and yet we still believe they exist (Carey, 2009; Rey, 1983). In short, there is a disconnect between how we judge things to really be (definitional judgment) versus how we categorize them. While the latter might rely on aggregate representations, the former does something more—perhaps setting logical bounds of a concept—and is argued to be more important to concept meaning.

This distinction between categorization versus definition appears to inform not only everyday intuitions but also researchers' use of the term "concept." Representations simply summarizing experiences appear in a broad swath of literature not using the terms "concepts" or "semantic memory." For example, the literature on ensemble perception shows that we spontaneously average the features of sets of items, whether the sizes of circles or orientation of Gabor patches, and can report them after a delay with remarkable accuracy (Alvarez & Oliva, 2009; Ariely, 2001; Haberman & Whitney, 2009; Parkes et al., 2001). Aggregation over experience is prominent in reinforcement learning (Shohamy & Daw, 2015), speech segmentation and visual statistical learning (Chun & Turk-Browne, 2008; Saffran, Aslin, & Newport, 1996), natural scene statistics (Simoncelli & Olshausen, 2001), and a huge range of other cognitive domains. In few cases is the term "concept" invoked to describe these phenomena. Conversely, researchers in semantic memory do not take just any evidence about aggregated memory as relevant to their phenomena of study: their reviews do not cite learning experiments with Gabor patches, water maze platforms, or the distribution of phonemes in spoken language. A related fact was noted in a highly extensive review of the concepts literature in cognitive psychology, arguing that many of these omissions seem arbitrary (Murphy, 2002).

Thus, it would seem that the average location of water-maze platforms is not considered part of conceptual memory in the cognitive neuroscience of concepts in the same way as "apples are a fruit that tends to grow on trees", but there is no formal definition that clearly distinguishes these examples. It thus seems like there is a latent intuition driving these terminological uses. Before we probe this intuition further, and consider philosophical proposals to answering it, we evaluate two other criteria that may be important to characterizing concepts: channel specificity and stimulus similarity.

12.3.2 Channel Specificity

Neuropsychologists have long used specific criteria to probe impairments to semantic memory following degenerative disease (such as semantic dementia) or stroke (Caramazza, Berndt, & Brownell, 1982; Hodges, 1992; Shallice, 1988; Warrington, 1975). Such criteria have identified impairments that are consistent regardless of the sensory channel or stimulus modality (pictures, words, sounds) with which the patient is tested. Indeed, there are many cases of patients showing impairment across various tasks probing their knowledge, including matching pictures to their names, selecting sets of more versus less related words or pictures, or judging the typical attributes of categories. Such patients are relatively spared on measures of episodic memory, such as recalling autobiographical details of their lives (Hodges, 1992). Additionally, such patients can be contrasted with others who have deficits specific to representations proprietary to certain sensory channels. For example, damage to a structural description system is thought to be responsible for impairments to the ability to recognize unconventional views of objects, relative to recognizing them from conventional views, which attest to the sparing of the concept itself (Schacter & Tulving, 1994; Warrington & Taylor, 1978). Instead, the mapping between visual form and the conceptual system is what is affected. It is also thought to be implicated in modality-specific naming disorders, such as optic aphasia, in which patients cannot name objects when they are presented visually but can otherwise. Because they can name them when presented in tactile form or verbal description, it is not a deficit at the linguistic or semantic level. Simultaneously, it is not an impairment to basic visual processing, since they can perform appropriate actions with those objects (Beauvois, 1982; Caramazza et al., 1990; Riddoch & Humphreys, 1987). Therefore, the idea of a distinct neural system responsible for modality-invariant, non-autobiographical memory, and their double dissociates, is well grounded in neuropsychological data.

Neuroimaging approaches to semantic memory guided by the same principle have tracked neural responses that are common to different modalities of presentation of the same concept, such as the word “rabbit” and pictures of rabbits (Devereux et al., 2013; Fairhall & Caramazza, 2013; Simanova et al., 2012). However, these approaches can still reflect channel-specific representations because even representations proprietary to a sensory modality can be associatively retrieved from another cue. For example, a verbal cue can prime motor plans and vice versa (Fischer & Zwaan, 2008; Yee et al.,

2013), and likewise, whatever knowledge structure that allows one to map the visually specific, structural description of rabbit to the word "rabbit" can also be retrieved in reverse fashion when reading "rabbit." Another approach has been to study whether representations are preserved in cases where a sensory channel is missing from birth, such as congenital blindness (Bedny, Caramazza, Grossman, et al., 2008; Bedny, Caramazza, Pascual-Leone, et al., 2011; Peelen et al., 2013; Peelen, Romagno, & Caramazza, 2012; Striem-Amit et al., 2018; Wang et al., 2015). If a neural response to the word "rabbit" in the congenitally blind is similar to the responses in the sighted, it cannot be a representation that belongs only to the visual channel. While none of these approaches are impervious to criticism, the issue we focus on here is whether in ideal circumstances these characteristics are even sufficient. Are cross-modally accessible representations always conceptual? Are such criteria enough to distinguish conceptual representations from others?

One reason to think they are not is to consider the kinds of cases we raised in the earlier section on aggregation. We noted that representations such as the average location of water-maze platforms is not considered a concept by many researchers, and it does not seem that making this representation cross-modal helps. Suppose that the mice could find the platforms haptically when blindfolded. Even if their representation of location is not tied to vision specifically, the example does not seem intuitively more conceptual. It is also relevant to note that literature on cross-modal integration is not normally considered to be investigating concepts. For instance, there is a rich body of work reporting common responses to dynamic faces and vocal sounds in the posterior superior temporal sulcus (Deen et al., 2015). This area is not considered to represent concepts, and is instead described as an integrative audiovisual area. In sum, literatures on semantic memory and conceptual knowledge do not use just any cases of cross-modally accessible representations as part of the relevant phenomena. We suspend discussion of why, and why this should matter, until section 12.3.4.

12.3.3 Stimulus Similarity

A third approach that we consider measures different kinds of similarity relations among stimuli as they are represented in various neural areas. A cardinal feature of concepts, according to the psychological literature, is that similarity relations among concepts need not follow the surface or physical feature similarity of their referents (Carey, 2009; Gopnik & Meltzoff,

1997; Keil et al., 1998; Mandler, 2004). In a classic example, children will judge that an animal born a raccoon but transformed to look like a skunk is fundamentally still a raccoon despite having the superficial features of a skunk (Keil et al., 1998). This idea has been crucial in research on conceptual development and beyond (Murphy, 2002).

Departure from physical feature similarity has sometimes, but relatively rarely, been adopted in cognitive neuroscience. While it has been used in our work on semantic memory (Leshinskaya & Caramazza, 2015; Leshinskaya et al., 2017), its most extensive use has been in work on visual recognition and categorization where the term "concept" is only sometimes invoked (Bracci & Op de Beeck, 2016; Freedman et al., 2001; Haushofer, Livingstone, & Kanwisher, 2008; C. B. Martin & Barense, 2018; Miller et al., 2003; Proklova, Kaiser, & Peelen, 2016; Roy et al., 2010). Consider, for example, Freedman and colleagues (2001), who specifically probed representations that discretize a continuous feature space of visual stimuli, leading to a categorical structure. Macaques were taught category boundaries among artificial images, which were a continuously interpolated set of morphs between two reference images: a cat and a dog (Freedman et al., 2001). The category boundary in this parametric stimulus space was defined arbitrarily, designating everything on one side as cats and everything on the other as dogs. This made it possible to dissociate distance in the feature space from category membership by finding pairs of items that are equidistant in the feature space that are either within a boundary (are both cats) or straddle it (one is a cat; the other is a dog). Neural responses could then be analyzed as reflecting distance in the stimulus parameter space or distance according to the category boundary (crossing it or not). It was found that some neurons, specifically in the lateral prefrontal cortex (lPFC), exhibited categorical responses, while others (in the inferior temporal cortex) reflected the overall visual distance but were insensitive to the boundaries.

The above is an important finding in its own right, irrespective of whether one calls the categorical representation "conceptual." For the present discussion, however, we raise the question of whether it should be. Should this finding be considered as demonstrating that the lPFC is a semantic area, or something else? In spite of its use of an important characteristics of concepts from the psychological literature, work of this sort is not incorporated in many reviews of semantic memory and not used as criteria in an operationalization of concepts in cognitive neuroscience (Binder & Desai,

2011; Binder et al., 2009; A. Martin, 2007; Patterson et al., 2007; Renoult, Irish, et al., 2019). While this could be simply a methodological oversight, we consider below that a separate intuition guides researchers' use of this term, and ask whether this intuition should be followed or discarded.

12.4 Probing Intuitions about Concepts

In this section, we offer preliminary evidence that the operational characteristics examined thus far do not always guide cognitive neuroscientists' judgments about the notion of a concept. We suggest that judgments of the extent to which information is shared capture usage not covered by the operational characteristics.

We have considered three well-operationalized characteristics for distinguishing among kinds of representations in long-term memory that can and have been used in cognitive neuroscience: (1) whether a representation aggregates across experience or individuates specific episodes, (2) whether it is proprietary to a certain sensory input channel or is channel invariant, and (3) whether it can depart from physical stimulus similarity. These distinctions are well defined, cognitively meaningful, and work to distinguish distinct neural systems reliably. Many of them cohere with criteria set out in cognitive psychology for conceptual memory, even though each alone seemed not to capture everything the term "concept" is expected to capture. Below, we consider whether a combination of all of these criteria might serve as a good candidate in our search for an operational definition of "concept" for cognitive neuroscience. However, we argue that even as a combination, this set of criteria does not always guide cognitive neuroscientists' usage of the term "concept," and we further probe what intuition or theoretical commitment does guide it.

12.4.1 Combining the Operationalized Characteristics

Suppose a mouse is tasked with learning the location of water-maze platforms across repeated experiences, aggregating them into an average. The mouse can access this knowledge visually or haptically. Moreover, the mouse acquires two categories of water-maze platforms—ones that fall on the east side of the pool and ones that fall on the west side—and we find representations of categorical cross-modal representations of water-maze positions in their brains. Have we identified a neural substrate supporting conceptual memory, at least for a specific concept?

The participant in this hypothetical experiment is a mouse, but there are analogous examples from everyday human experience. Let's consider your long-term memory of the average route you take to work, the typical location of gas stations on the way there, the usual color of bicycles you own, or the typical contents of your grandmother's attic. All of these have (or can easily have) the set of characteristic features we just outlined but are not canonically conceptual. If our operational criteria for concepts include knowledge such as categories of water-maze platform locations and the typical content of one's grandmother's attic, then it would behoove concepts researchers to incorporate findings regarding such representations into their theoretical frameworks and reviews, and expand the scope of their experiments to include them. However, this does not appear to be the case. Despite meeting the representational criteria, the term "conceptual" is not typically used to denote these kinds of representations across the literature in cognitive neuroscience.

To substantiate our observations further regarding the use of the terms "concept" and "semantic memory" in the literature, we additionally queried cognitive neuroscientists of semantic memory directly. We sent a small survey to thirty-three researchers in the cognitive neuroscience of semantic memory. Of these, eleven completed the survey. Ten of these were either postdoctoral or professorial level scholars, and one was a graduate student. Although this is a small sample, we believe it is unbiased because the participants did not know the question under investigation or the position of the authors on it. Indeed, we had not discussed this topic with any of the survey participants.

Participants provided electronic consent in accordance with procedures approved by the Institutional Review Board of the University of California, Davis. Participation was entirely voluntary, and responses were anonymized; no compensation was provided. We first asked participants to indicate their background, and then presented the following instructions:

> In the following, each question will describe a mental representation or belief that a person might have. For each one, indicate the degree to which this representation would be typically studied under the domain of "semantic memory" or "conceptual knowledge." We are interested in your opinion about how these terms are used in your field.

Participants were then shown one phrase at a time describing a belief, and a slider response ranging from 1 ("definitely no") to 5 ("definitely yes") indicating the extent to which they agreed that this belief would be considered "semantic memory," "conceptual knowledge," or "something else." Phrases

were of three kinds: examples of highly canonical conceptual representations (such as "Dressers are a kind of furniture used to store clothes" and "Grandmothers are women whose offspring also have offspring"), examples of canonical episodic memories that are specific to a time and place ("One time, I rode my bike along the river in town and found an abandoned mill twenty miles away" and "The gas price today is $3.45/gallon"), and examples that aggregate over experiences but are not canonically semantic ("The dresser in my bedroom is about two feet from the bed in the northwest corner of the room, and I can usually navigate around it in the dark," "Most of the bicycles I've owned have been yellow," and "My grandmother tends to store old photo albums in her attic but not heirlooms").

We present the results in table 12.1. Responses to "semantic memory" and "conceptual knowledge" questions were highly correlated ($r = 0.99$). So, these were collapsed into a combined "semantic/conceptual" rating. We found that the mean response on the semantic rating scale for canonical semantic examples ($M = 4.67$) was significantly higher than for both the canonical episodic examples ($M = 2.66$, $t[10] = 7.02$, $p < 0.001$) and the aggregate personal examples ($M = 3.32$, $t[10] = 5.63$, $p < 0.001$). Correspondingly, responses on the "something else" rating scale were lower for the canonical semantic items ($M = 2.18$) than both the canonical episodic items ($M = 4.27$, $t[10] = 7.39$, $p < 0.001$) and the aggregate non-canonical items ($M = 3.25$, $t[10] = 2.63$, $p = 0.025$). The pattern of responses also reveals that ratings for the aggregate non-canonical items were intermediate between the canonical semantic and episodic items. Thus, the semantic rating was indeed higher for the aggregate non-canonical items than for the canonical episodic items ($t[10] = 2.18$, $p = 0.046$) and the rating for "something else"

Table 12.1
Results from a survey of eleven researchers in the cognitive neuroscience of semantic memory

	Rating Type	
Item Type	Mean (*SD*) Rating on Semantic/Conceptual Scale (1–5)	Mean (*SD*) Rating on Something Else Scale (1–5)
Canonical semantic	4.67 (0.42)	2.18 (0.67)
Canonical episodic	2.66 (0.88)	4.27 (0.79)
Aggregate non-canonical	3.32 (0.91)	3.25 (1.28)

was lower ($t[10] = 2.82$, $p = 0.018$). Overall, this suggests that cognitive neuroscientists of semantic memory—at least, the ones queried here—classify these aggregated experiences differently from both episodic and semantic memory. They do not believe they are considered conceptual in the same way as the canonical examples. Furthermore, they are less confident that they are studied as "something else" relative to episodic examples. Overall, this suggests that there is genuine uncertainty in cognitive neuroscience of semantic memory about how such examples should be classified with respect to existing ideas about memory systems. Although the small sample may fail to capture intuitions of all cognitive neuroscientists, it reinforces our observations from the literature regarding how these terms are used. Altogether, we believe that something is missing from the definition of semantic memory as it is currently actually used in the field.

There are two paths forward. One path is to look for additional criteria that will help us classify the non-canonical examples and incorporate these additional criteria into the operational definition of semantic memory. The other path is to suppose that these in-practice use patterns of the term "semantic memory" are not meaningful, or are erroneous, and should be replaced with formal, operational definitions that make empirical predictions about brain and behavior. If intuitions about concepts do not coincide with relevant scientific notions in cognitive neuroscience, then these intuitions have no place guiding terminology in the field. To adjudicate between these paths, we consider what additional criteria might guide the latent intuition behind the usage of the term "concept" and whether these criteria may be scientifically useful.

12.4.2 Probing the Intuition

It is possible that there is no coherent way to distinguish between bona fide examples of concepts, such as the knowledge that apples are a type of fruit that grow on trees, and these seemingly odd, non-canonical examples, such as the knowledge that my grandmother liked to store photo albums in her attic. However, we propose that a coherent but latent intuition does guide this distinction, and that the basis for this intuition turns on notions of sharedness or intersubjectivity.

Historically, Tulving (1972) first proposed to distinguish semantic memory, which stores concepts, from episodic memory, which stores personal experiences specific to a time and place. Episodic memory was thus characterized

by at least two features: specificity to time and place, and a personal or autobiographical nature (Tulving, 2002). Both features have been used (sometimes separately) in the operational definition of episodic memory since then (Nadel & Moscovitch, 1997; Renoult, Davidson, et al., 2012; Westmacott et al., 2004; Winocur et al., 2010). On the other hand, when semantic memory has been operationalized, only the characteristic of generality of time and place have been used explicitly. The inverse of autobiographicality has not been used to operationalize it (though cf. Renoult, Davidson, et al., 2012).

We propose that this inverse could be called "sharedness" or "intersubjectivity" and that it serves as one endpoint of a continuum opposite to autobiographical memory. This continuum would reflect the extent to which information describes one's own idiosyncratic and personal experience or whether it is seemingly something that many others have also observed and know about. This idea is distinct from (though related to) generality or aggregation because autobiographical knowledge can span beyond a singular episode. For example, the knowledge of what I *typically* have for breakfast, or what my grandmother *tends* to store in her attic, are exactly the kinds of cases that this accommodates.

Figure 12.1 illustrates several examples of less and more autobiographical information. At one extreme is the particular breakfast I had this morning, alone at home. At the other is the idea of food eaten in the morning, which leaves open what exactly is eaten but increases its universality in human experience. No matter the country or culture, one could understand "breakfast" in this way. A simple litmus test for locating information along this dimension is by imagining what content is reliably conveyed by the

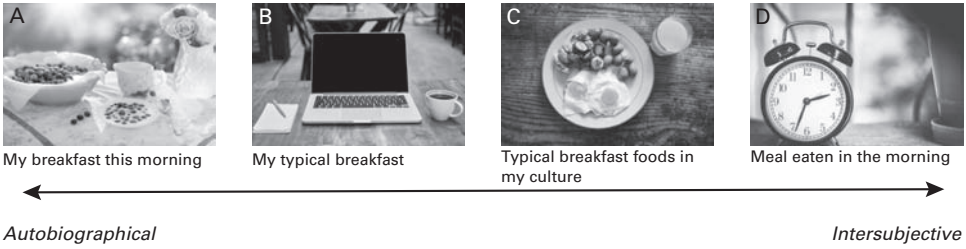


Figure 12.1
Hypothesized placement of examples of different kinds of memories in terms of their level of autobiographicality versus intersubjectivity/sharedness. (Images are open-source with free license). See color plate 3.

use of a word: when I say "breakfast" in conversation, I don't expect the word itself to convey anything about *my* breakfast that morning, or that I usually have breakfast in a sunny nook in my kitchen. I might expect others to understand that it is a meal eaten at the start of the day; and maybe that it contains typical breakfast foods in North America. Of course, this inference requires an understanding of what is common knowledge among speakers (like much of conversation).

This distinction between personal, autobiographical versus shared, intersubjective information seems to capture the intuitions and usage surrounding the word "concept" in cognitive neuroscience. For example, Hodges (1992) described semantic memory as knowledge that is "culturally shared," and Tulving (1972) described it as referring to the world rather than the self. The usage from these leading neuropsychologists has been influential; it is taken as the inherited view in modern work (e.g., Yee & Thompson-Schill, 2016).[3] It may be for this reason that cat–dog categories, locations of water-maze platforms, the contents of one's grandmother's attic, and other aggregated, categorical, cross-modal representations are not universally treated as conceptual. We argue that this is because they are idiosyncratic, personal, and lacking a sense of sharedness.

From the perspective of the observer or knower, a concern with sharedness is ultimately a concern with representing facts about the world that are independent of one's fallible view. A given observation may be a chance occurrence, and glitch of the senses, or an internally generated hallucination; it may be highly local and context dependent; or it may be simply incidental. Perhaps, the spiciness of a specific Szechuan pepper is because of my overly sensitive taste buds, a poor choice of specimen, contamination from hot sauce, or the peculiarities of my local Chinese restaurant. In these cases, storing the notion that Szechuan peppers are spicy as a fact in semantic memory would be a bad idea.

A similar principle might guide the intuitions documented in psychology experiments that clothing style is not part of the concept "boy" or that looking like a raccoon is not part of the concept "raccoon," as we described earlier. Judgments about more defining or essential properties are similarly concerned with representing what things really are. Even if we never obtain the ground truth, true properties are more likely to be shared across viewers, viewpoints, and incidentals of the environment. So, a concern with sharedness dovetails with these concerns also.

In the water-maze example, the tendency of platforms to be on the west side of the pool might be an accidental feature of the maze, itself an artificial setup specific to the lab. Whether other mazes have those features and whether other mice ever experience them is unknown to the mice. Yet, if the mice had evidence that they are not idiosyncratic, individual experiences, perhaps these representations would start to resemble more classic examples of concepts. One relevant cue would be if mice had ways of reaching the understanding that these mazes are a ubiquitous feature of mouse life. Suppose mouse A leaves the experiment chamber and returns to the cage, signaling to mouse B that it's a category 1 type day—the platforms will be mostly on the west side! They could come to believe these were real objects in the world, no longer purely autobiographical. We conjecture that the use of the term "concept" in cognitive neuroscience is implicitly guided by these considerations, similarly to how it guides laypersons in their judgments in psychology experiments.

The critical question we now raise is whether such considerations about sharedness versus autobiographicality should form part of the formal operational definition of the scientific term "concept" in cognitive neuroscience—that is, be used explicitly as a way to distinguish memory systems in neuroscientific experiments (which they have not been).

12.5 Proposals for Sharedness as Operational Criteria for Concepts

To address the question of whether sharedness should be part of the operational definition of conceptual memory, we turn to discussions in the philosophy and psychology of concepts regarding the importance of sharedness in a formal theory of concepts.

Two recent works on concepts, Ruth Millikan's *Beyond Concepts* (2017) and Susan Carey's *The Origin of Concepts* (2009), take opposing views on the extent to which cognitive science should treat sharedness as a critical property of conceptual representation. Their concern is in the objective fact of sharedness: whether concepts are, or cannot be, mental representations that are truly common among people, truly distinct from idiosyncratic bits of belief. This would seem relevant for the question of whether the scientific notion of "concept" should incorporate sharedness in its definition. Accepting Milikan's view would entail discarding this distinction and thus suggest a major ontology revision to cognitive neuroscience. Accepting Carey's view

would require justifying an objective basis for distinguishing shared versus autobiographical content. We resolve this tension by offering a third view.

12.5.1 What Is Shared among Minds?

An influential philosophical position on concepts is that meanings, whether of words or mental representations such as concepts, are largely not in the head (Kripke, 1972; Putnam, 1973). A concept is a mental representation that points to a referent, but it need not have any special content apart from this referential relation. This is illustrated by a thought experiment that asks us to suppose that there exists a twin Earth in which you have a perfectly identical twin who has had all of the same experience as you (Putnam, 1973). The only difference in the twin world is that the molecular structure of water is XYZ rather than H_2O. Yet, to all available observations, the properties of twin water are identical to those of Earth water. The key intuition is that your thoughts about water and your twin's thoughts about water refer to different things, even though your experiences and knowledge about water are the same. This divergence in meaning is due to being embedded in environments in which these concepts point to different elements. By this argument, any property of a concept can be revised without changing its meaning. For example, it may turn out that all cats are really demons from Mars rather than mammals, and yet our concept "cat" would still point to the same things in the world. Thus, it is not by virtue of anything we believe about water or cats that these concepts point to their referents. Rather, reference or meaning obtains by virtue of a special kind of relation between a thing in the world and your mental representation of it.

If one accepts this argument, it is not clear what exactly in a mental representation is or need be shared across minds for concepts to mean the same thing—apart from their pointing to the same thing (where things may be physical, social, abstract, or otherwise). It does not, for purposes of meaning, matter what things we attribute to the concept "Szechuan pepper"—where one finds it, how it tastes, what our favorite dish is to put it in. The critical implication is that there is no objective fact of the matter as to which properties of Szechuan peppers are more or less important for the meaning of the concept "Szechuan pepper." This leaves us without an objective dividing line among the things in the head with respect to what something really is or means and, consequently, between personal belief and shared knowledge.

One response, then, is to throw in the towel—largely to accept that everything that is in the head is equally irrelevant to concept meaning, and to accept that there is no objective dividing line between meaning-relevant information and the rest of our seemingly local, subjective, and idiosyncratic memory: the average location of water-maze platforms and the typical contents of grandma's attic. We take this to be the implication of Millikan's (2017) arguments, as we elaborate below. The outcome of this view is that our intuitions about concepts as "shared" bits of knowledge are a relic of our folk psychology, and that they maybe should not be taken as defining the division between semantic memory and other things in long-term memory.

The alternative response is to defend the line between meaning-relevant, definitionally important, shared knowledge versus irrelevant, idiosyncratic, personal beliefs. This requires a concrete proposal for how to do so. We follow Carey's (2009) arguments for how and why we should attempt to draw this line with objective and formal criteria. If there really are such shared contents, then the scientific attempt to figure out what those are seems well motivated. We discuss whether these criteria can be practically applied.

A third, intermediate, view is to accept that there is no fact of the matter as to what mental content is central to conceptual meaning, but to propose that our meta-cognitive, intuitive concern with this makes the distinction psychologically real. This view, however, would suggest a significant departure for how cognitive neuroscience delineates memory systems, and we elaborate these implications in section 12.5.4.

12.5.2 Throw in the Towel: Unicepts and Non-Shared Content

Millikan (2017) suggests a revision to how we understand the mental representations typically called concepts; she offers instead the notions of unicepts and unitrackers, which, unlike concepts, have no mandate to be shared. While Millikan grants that mental representations should refer to the world, their function for the user is to track objects in the distal world, under the various sensory conditions in which they present themselves: to see the same thing as the same and different things as different. A unicept is a mental representation that binds together the various bits of information we might have about the same thing, and unitrackers flexibly connect unicepts to observations. For example, a unitracker for Szechuan pepper might be sensitive to the kind of restaurant or dish in which an observer has tended to find those peppers, to the extent that this information is

useful. For unicepts and unitrackers to perform this role, they absolutely should and do reflect the idiosyncrasies, localities, and subject specificities of personal experience.

Despite having a highly experience-specific origin, these mental representations pick out enough commonality across individuals because there is systematicity to the information in the world. Millikan (2017) asks us to imagine the world as a many-dimensional space, where each dimension is a feature or property and each object is a point in this space. The result is clumpy, with dense areas separated by sparse gaps, because while some feature combinations are common, others are rare. Being a pepper and being served in a chicken dish are related. The same distal object thus has a "univocal" reason for creating its diverse sensory impressions, creating "many quite reliable ways to identify a human or a cat, an oak tree or an automobile or a laptop or a piano, as such" (p. 12)—a fact that helps us accumulate knowledge in constantly varying circumstances. When we talk to each other, we make pointers to the same clump, even if each of us has a slightly different view of it. What matters is that our mental representations share a referent to the same clump, and communication can work.

Millikan (2017) thus rejects the intuition that concepts must be the same across minds. On the basis of "meaning rationalism," she argues that communication, meaning, and reference need not be perfect but rather as evolved phenomena, and need only be as good as needed to be selected—and thus can fail even the vast majority of the time (e.g., mating acts in many species). Without speculating about failure rates in the use of concepts, this fact should already make us suspicious about intuitions concerning what must be present in the head for, say, humans to communicate accumulated knowledge across generations. In short, Millikan has developed and refined an elaborate and comprehensive theory about linguistic and mental intentionality and the lack of need for classical philosophical concepts.[4]

The upshot is a theory of mental content that does not posit a distinction between conceptual content and most of the rest of the information in memory.[5] "Information used to same-track is all of a kind, none of it more important, more defining, or more conceptual than the rest" (Milikan, 2017, p. 49). The stuff formerly taken to be concepts can be personal, local, and even temporary. In her example, "my glass at this party" is as much a concept as any other; a point we see as extending to today's water-maze platforms, and to imply that "typically found in Chinese restaurants" is

part of the concept "Szechuan pepper," just as much as any other fact. If this does not conform to our intuitions, our intuitions are irrelevant. The upshot of this view for cognitive science is that trying to establish a conceptual versus non-conceptual boundary in memory may not be meaningful.

Millikan's (2017) take on mental representation is aligned with a number of approaches and observations in modern cognitive neuroscience. The idea of a clumpy correlated feature space stemming from the statistics of the world is an influential one, and can explain important phenomena in semantic system impairment (Capitani et al., 2009; McClelland & Rogers, 2003; Rogers et al., 2004, but cf. Caramazza et al., 1990, for arguments that it might not be the most parsimonious explanation). Recent neuroimaging work has suggested that the very organization of the cortex might be structured to facilitate the readout of correlated but diverse information. Konkle (2019) argues that having "content-channels" that represent a full diversity of features about the same kind of thing, from their typical basic forms to structural descriptions to more semantic knowledge, can facilitate their readout and extraction from observation by making their alignment explicit in cortical space. For example, animate things tend to have broadly curvy forms and tend to draw on foveal representations, while inanimate things tend to be boxy and extend more in to the periphery. All of these distinctions seem to be captured in aligned ways in overlapping parts of the ventral visual stream, so that curvy things, foveated things, and animate things all draw on overlapping cortex (Long, Yu, & Konkle, 2018). This can facilitate the use of this mutual information for recognition or, as Millikan would put it, "same-tracking." Perhaps, in neural space, representational types are too dense a mixture to cleave into concepts and the rest of knowledge cleanly.

Finally, some cognitive neuroscientists argue that because concepts are accumulated from one's potentially idiosyncratic personal experience, there is no reasonable way to separate a shared aspect from the rest, and the goal should not be to search for it (Casasanto & Lupyan, 2015; Yee & Thompson-Schill, 2016). They marshal substantial evidence that the information participants retrieve about the same referent varies to a large degree, influenced by current task, recent experience, accumulated experience, and individual cognitive differences, suggesting that the bulk of information we retrieve about the same thing is more variable than shared (Yee & Thompson-Schill, 2016). We have taken for granted that many different properties of the same referent are available, and the psychological evidence is compelling

that their retrieval varies by situation: the spiciness of Szechuan peppers if anticipating eating it, the cuisine that uses it if seeking it out. However, the question remains whether we can systematically sort that information into aspects that are more idiosyncratic and more shared. Below, we review one additional proposal for how to do so.

12.5.3 Drawing the Line: Objective Criteria for Conceptual Content

Although not a direct response to Millikan, Carey (2009) argues that allowing idiosyncratic knowledge to be a part of concepts and not distinguishing between conceptual and idiosyncratic properties is a devastating move. Her rationale for this view, and approach to distinguishing these two forms of content, is as follows.

First, Carey (2009) notes that we must psychologically distinguish belief revision from conceptual change—operations that function differently in development. Conceptual change is slow and dramatic, and leads to previously unavailable representational resources that support new forms of inference. For example, children undergo a conceptual shift when they transition from being able to count to a finite number (initially one, two, or three) to a generative, productive understanding of the successor function (that one can always add one and obtain a new number ad infinitum). These enable radically distinct kinds of thoughts. In contrast, learning that Szechuan peppers typically appear in Chinese restaurants might be only a revision of belief—quick to learn with minimal impact on thought processes. There is thus an inferential origin story that concepts have that other beliefs do not.

Second, she motivates the distinction by arguments from Fodor and Lepore (1992) that if we do not draw that line, it would follow that *every* belief is dependent on every other. If we do not have a principle by which we determine which beliefs are important to understanding the concept "apple," we are committed to saying that all of them are, including our beliefs about camels and nuclear reactors. The consequences of this are philosophically objectionable. It would lead to absurd conclusions, for instance, that two people cannot have the belief that table salt is a kind of seasoning if they disagree about other things, such as salt is good for one's health. To avoid this consequent absurdity, dubbed "holism," some table salt content must be privileged in determining the concept of table salt. However, Fodor and Lepore (1992) argue that privileged content amounts to an analytic/synthetic

distinction, which they take to be definitively debunked by Quine (1951). There are two general routes to countering Fodor and Lepore (1992). One can either deflate the supposedly absurd consequences of holism or defend a way of demarcating privileged, meaning-determining content. Carey (2009), following Block (1986), opts for the latter strategy.

Carey's (2009) proposal for how to draw this line is according to the causal processes by which we create new concepts: those concepts that were causally implicated in forming an initial new concept are relevant; others are not. Thus, there is not only a causal (referential) relation between a referent and a concept, but also a causal relation among concepts that are principled. Thus, we can draw the line between mental contents that are, and those which are not, relevant to a concept's meaning.

Could neuroscience take these concerns into consideration when mapping out memory systems? On theoretical grounds, Fodor and Lepore's (1992) argument does not seem to motivate drawing a distinction between *kinds* of knowledge: it simply argues for a need to distinguish which beliefs are relevant to which others, but not grounds for postulating distinct memory types, such as semantic and episodic. It is thus an argument about how to delimit the relations among beliefs but not that some beliefs X are always conceptual and other beliefs Y are always idiosyncratic. Another concern is that this distinction might not be empirically tractable. To trace the causal path between prior concepts and newly formed ones is currently beyond the capacities of empirical practice. Nonetheless, this challenge may be possible, and worthwhile, to meet in the future.

In the meantime, it could be possible to use participants' judgments about the status of different conceptual properties to make such distinctions. Prasada and Dillingham (2006, 2009) show that adults explicitly distinguish between properties that are equally prevalent for an object, but differ in their "principled connection" to that object's kind membership. For example, participants judge that the claim "dogs have four legs" is similar to the claim that "dogs, by virtue of being the kinds of things they are, are four-legged," but that being red, for a barn, amounts to only "barns, in general, are red"—but not that there is any meaningful relation between being a barn and being red. A range of related findings supports the idea that these distinctions are psychologically real and judged consistently. It is unclear whether this distinction maps onto the distinction between autobiographical and shared, but this could be measured also.

Finally, another important theoretical position on concepts (Machery, 2005, 2010) argues that psychological phenomena can reliably distinguish at least three kinds of representations typically called "concepts" by psychologists (Margolis & Laurence, 1999; Murphy, 2002). Sometimes, participants seem to use statistical summaries; other times, they retrieve individual exemplars and, yet other times, theory-like structures (such as causal models). If signatures of these phenomena are sufficiently stable and distinguishable, then a tripartite typology based on them is warranted. Furthermore, all three can be distinguished from "background," idiosyncratic knowledge by virtue of how automatically or prominently they are retrieved. While theorists such as Carey and Prasada would argue that more than ease of retrieval should characterize the distinction between concepts and idiosyncratic knowledge, the general idea of using dissociations among psychological tasks to determine a typology of long-term memory empirically is an appealing one.

In sum, it remains to be seen whether one can distinguish shared versus idiosyncratic aspects of knowledge objectively in ways that could be incorporated into neuroscientific operational criteria, and whether they would allow empirical traction on memory systems in the brain, but these paths also appear promising.

12.5.4 Biting the Meta-Cognitive Bullet

We propose a third solution, which simultaneously accepts that conceptual content is distinct from idiosyncratic personal experience but does not propose that this division is objective. Rather, this distinction is meaningful because concept users believe it is true and have ways of making it for themselves. The implication of this view is a very different way to approach the delineation of memory systems in cognitive neuroscience.

According to this perspective, the way memories are stored in the brain is influenced by the experiencer's own "meta-cognitive" assessment of whether the information they observed is personal and idiosyncratic versus intersubjective and shared. As inherently subjective observers, we can only make a best guess at to the extent to which some content is available to others. Nonetheless, many cues are available to help us do so, and these subjective judgments themselves could matter for how a memory is encoded.

Taking the example of Szechuan peppers, different circumstances of how I learned about their spiciness can support my own inferences about the intersubjectivity of this property: if a friend tells me that they are spicy, or

if I read about this in an encyclopedia, I can rest assured that at least some other people experience them that way and conclude that it is probably a true fact about them. On the other extreme, if I sample Szechuan peppers just once, I could doubt whether my experience is either the same as anyone else's or factual at all. I might store the latter as an episodic memory and not integrate it into my knowledge of the world. But if I make the same observation repeatedly across different times and places, the chance that it is generally true increases. As we suggested earlier, we can use both consistency of observation and evidence of intersubjectivity as ways to infer whether an observation exists independently of our fallible sensory experience. This inference does not have to be made explicitly to exert an influence over how information is stored.

Not coincidentally, then, such cues coincide with many of the operational characteristics of memory reviewed through this chapter. Observations that are aggregated (consistent across multiple experiences) are less likely to be due to chance, and those that are cross-modally accessible are not likely relics of an idiosyncratic perceptual channel, and more likely to be common across observers with diverse perspectives (or even common with ourselves at another time, perhaps in a dark room). The concern with definitional, not just typical, features and with observations that reflect an underlying cause, rather than surface properties of stimuli, also reflect the concern with understanding what something *really is*. These criteria may thus operate at a psychological and neural level to distinguish between accidents of sensory noise and possibly real, shared referents in the world that exist independently of us. In this framework, there is a coherent way in which these different properties of memory distinguish our own observed experiences. The upshot is that we are able to use the characteristics of our own experiences to infer which observations are shared by others, and that this inference can itself impact the memory systems in which that information is stored. The empirical prediction that follows is that a person's own judgment about their experiences should affect how those experiences are neurally stored. If an observer has reason to believe that a certain observation is of a stable aspect of reality available to others, that observation would be used to update semantic memory. If that observation is instead only about their personal experience and not reflective of generally available facts, it would be stored in episodic memory. These judgments can be affected by any of a set of cues such as the ones described throughout this chapter. The way that these judgments

are made by an observer would result in predictable and distinct neural and psychological outcomes for how that observation is stored in memory. The broader implication of such a finding would be that individuals' judgments about their own observations are a major determinant of how memories form. These remain open to empirical investigation.

12.6 General Conclusion

We have sought to identify how conceptual knowledge can be operationalized in cognitive neuroscience. We reviewed three characteristic features that have been successfully used to distinguish among types of long-term memory, including the extent to which a representation aggregates across experience, whether it is proprietary to a certain sensory input channel, and whether it can depart from physical stimulus similarity space. We found that these distinctions, while theoretically and empirically justified, did not map cleanly onto the notion of conceptual knowledge as it is used within cognitive neuroscience or in psychology: not all aggregated or cross-modal memories are considered concepts.

We argued that rather than just an error of terminology, this usage reflects an underlying intuition guiding the use of the term "concept" based on notions of sharedness. For example, many examples of memories that meet the operational criteria are ones that are personal and idiosyncratic—such as the typical contents of my grandmother's attic or my favorite color of bicycle. The guiding intuition that excludes such examples from the domain of semantic memory reflects historically influential ideas that characterize semantic memory as "shared" knowledge. Yet, this criterion has not itself become part of the formal operational definition for measuring semantic memory in cognitive neuroscience, and thus its empirical validity in predicting neural organization remains to be tested directly.

We sought here to determine whether sharedness should be considered part of the operational definition for semantic or conceptual memory by appealing to opposite theoretical positions on the topic. Millikan (2017) argues that sharedness is not a meaningful distinction for concepts, while Carey (2009) argues instead that it is essential. Taking Millikan's position entails largely discarding the semantic-episodic distinction in long-term memory and revising this ontology in cognitive neuroscience. Taking Carey's view requires justifying an objective basis for distinguishing between

shared versus autobiographical/idiosyncratic content, but proposals for how neuroscientists could test such divisions remain to be developed.

We offered a third view, which argues that individual observers are capable of making the distinction between shared and autobiographical aspects of their own observations themselves subjectively. We predict that these subjective judgments themselves influence how observations are allocated among memory systems. For observers, these judgments are motivated by their ultimate concern with establishing whether their observations reflect intersubjective reality, and they can rationally make use of a variety of cues to guide their judgments. By systematically manipulating such cues experimentally, we expect to be able to affect whether observers classify an observation as personal or shared and, consequently, how that observation is allocated to their own memory systems. Overall, we believe that the theoretical alternatives we have outlined offer robust empirical predictions, and testing them has the potential to redefine our theories of memory systems.

Notes

1. "Concepts" in cognitive neuroscience are thought to be part of long-term (semantic) memory (e.g., Schacter & Tulving, 1994). It is possible that there are other cognitive domains in which concepts are a distinguishable part, but we focus on memory here.

2. We use "experience" broadly here: experiences need not be sensory, but also based in thought, language, and imagination.

3. These authors challenge that view; we turn to their arguments in section 12.4.2.

4. Mark Wilson's (2008) work against classical conceptions of concepts shows that in actual science, they do not even behave in the way many philosophers have thought they needed to in order for our paradigm of knowledge to be successful. Nonetheless, Millikan's theory was met with a number of objections and counterintuitive consequences. Rather than address them all here, we instead highlight the power of her motivating basics and the subsequent opportunity to develop alternatives within a "Millikan tradition." Approaching cognition as being confronted with some basic problems to solve and considering signs and communication systems as evolved solutions to such problems is an underexplored and potentially powerful way of addressing traditional philosophical issues about content—or, perhaps in some cases, learning to recognize false puzzles and dilemmas they have been taken to raise.

5. Millikan does, however, draw a distinction between those observations for which we do not form unicepts for referents which we do not care to same-track as such. Her examples are perceptual constancy (seeing the color of the wall as constant despite

differences in illumination), phonemes (categories of speech sounds), and oriented edges such as Gabor patches. These do not involve unicepts because we do not care to track them as such—to gather information *about* edges, *about* phonemes, or *about* colors. This may or may not be psychologically true. This distinction between things we track and things we do not could map onto the functional divisions we earlier described between semantic memory and the structural description system, but it would not draw the same line between semantic and episodic memory.

References

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences of the United States of America, 106*(18), 7345–7350.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*(2), 157–162.

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition, 13(3)*, 263–308.

Beauvois, M.-F. (1982). Optic aphasia: A process of interaction between vision and language. *Philosophical Transactions of the Royal Society B: Biological Sciences, 298*(1089), 35–47.

Bedny, M., Caramazza, A., Grossman, E., Pascual-Leone, A., & Saxe, R. R. (2008). Concepts are more than percepts: The case of action verbs. *Journal of Neuroscience, 28*(44), 11347–11353.

Bedny, M., Caramazza, A., Pascual-Leone, A., & Saxe, R. R. (2011). Typical neural representations of action verbs develop without vision. *Cerebral Cortex, 22*(2), 286–293.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences, 15*(11), 527–536.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex, 19*(12), 2767–2796.

Block, N. (1986). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy, 10*(1986), 615–678.

Bracci, S., & Op de Beeck, H. (2016). Dissociations and associations between shape and category representations in the two visual pathways. *Journal of Neuroscience, 36*(2), 432–444.

Capitani, E., Laiacona, M., Pagani, R., Capasso, R., Zampetti, P., & Miceli, G. (2009). Posterior cerebral artery infarcts and semantic category dissociations: A study of 28 patients. *Brain, 132*, 965–981.

Caramazza, A., Berndt, R. S., & Brownell, H. H. (1982). The semantic deficit hypothesis: Perceptual parsing and object classification by aphasic patients. *Brain and Language, 15*(1), 161–189.

Caramazza, A., Hillis, A. E., Rapp, B. C., & Romani, C. (1990). The multiple semantics hypothesis: Multiple confusions? *Cognitive Neuropsychology, 7*(3), 161–189.

Carey, S. (2009). *Origin of concepts*. Oxford: Oxford University Press.

Casasanto, D., & Lupyan, G. (2015). All concepts are ad hoc concepts. In E. Margolis & S. Laurence (Eds.), *The conceptual mind: New directions in the study of concepts* (pp. 543–566). Cambridge, MA: MIT Press.

Chun, M. M., & Turk-Browne, N. B. (2008). Associative learning mechanisms in vision. In S. J. Luck & A. Hollingworth (Eds.), *Visual memory* (pp. 209–245). Oxford: Oxford University Press.

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional organization of social perception and cognition in the superior temporal sulcus. *Cerebral Cortex, 25*(11), 4596–4609.

Devereux, B. J., Clarke, A., Marouchos, A., & Tyler, L. K. (2013). Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *Journal of Neuroscience, 33*(48), 18906–18916.

Dimsdale-Zucker, H. R., Ritchey, M., Ekstrom, A. D., Yonelinas, A. P., & Ranganath, C. (2018). CA1 and CA3 differentially support spontaneous retrieval of episodic contexts within human hippocampal subfields. *Nature Communications, 9*(1), 294.

Eichenbaum, H. B. (2004). Hippocampus: Cognitive processes and neural representations that underlie declarative memory. *Neuron, 44*(1), 109–120.

Eichenbaum, H. B., & Cohen, N. J. (2001). *From conditioning to conscious recollection: Memory systems of the brain*. Oxford: Oxford University Press.

Fairhall, S. L., & Caramazza, A. (2013). Brain regions that represent abstract conceptual knowledge. *Journal of Neuroscience, 33*(25), 10552–10558.

Fischer, M. H., & Zwaan, R. (2008). Embodied language: A review of the role of the motor system in language comprehension. *Quarterly Journal of Experimental Psychology, 61*(6), 825–850.

Fodor, J., & Lepore, E. (1992). *Holism: A shopper's guide*. Cambridge, MA: Blackwell.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science, 291*, 312–316.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

Haberman, J. M., & Whitney, D. (2009). Averaging facial expression over time. *Journal of Vision, 9*(2009), 1–13.

Haushofer, J., Livingstone, M. S., & Kanwisher, N. G. (2008). Multivariate patterns in object-selective cortex dissociate perceptual and physical shape similarity. *PLoS Biology, 6*(7), 1459–1647.

Hodges, J. R. (1992). Semantic dementia: Progressive fluent aphasia with temporal lobe atrophy. *Brain, 115*, 1783–1806.

Hodges, J. R., & McCarthy, R. A. (1995). Loss of remote memory: A cognitive neuropsychological perspective. *Current Opinion in Neurobiology, 5*(2), 178–183.

Hoffman, P., McClelland, J. L., & Lambon-Ralph, M. A. (2017). Concepts, control and context. *Psychological Review, 125*(3), 293–328.

Keil, F. C., Smith, W. C., Simons, D. J., & Levin, D. T. (1998). Two dogmas of conceptual empiricism: Implications for hybrid models of the structure of knowledge. *Cognition, 65*(2–3), 103–135.

Knowlton, B. J., Ramus, S. J., & Squire, L. R. (1992). Intact artificial grammar learning in amnesia: Dissociation of category-level knowledge and explicit memory for specific instances. *Psychological Science, 3*(3), 172–179.

Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning and Memory, 1*(2), 106–120.

Konkle, T. (2019). Emergence of multiple retinotopic maps without a feature hierarchy. *Journal of Vision, 19* (10), 90a–90a.

Kripke, S. (1972). *Naming and necessity*. Malden, MA: Blackwell.

Leshinskaya, A., & Caramazza, A. (2015). Abstract categories of functions in anterior parietal lobe. *Neuropsychologia, 76*, 27–40.

Leshinskaya, A., Contreras, J. M., Caramazza, A., & Mitchell, J. P. (2017). Neural representations of belief concepts: A representational similarity approach to social semantics. *Cerebral Cortex, 27*, 344–357.

Long, B., Yu, C.-P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences of the United States of America, 115*(38), E9015–E9024.

Machery, E. (2005). Concepts are not a natural kind. *Philosophy of Science, 72*(3), 444–467.

Machery, E. (2010). Précis of doing without concepts. *Behavioral and Brain Sciences, 33*, 195–244.

Mahon, B. Z. (2015). Missed connections: A connectivity constrained account of the representation and organization of object concepts. In E. Margolis & S. Laurence

(Eds.), *The conceptual mind: New directions in the study of concepts* (pp. 79–116). Cambridge, MA: MIT Press.

Mandler, J. M. (2004). *The foundations of mind: The origins of conceptual thought*. New York: Oxford University Press.

Manns, J. R., Hopkins, R. O., & Squire, L. R. (2003). Semantic memory and the human hippocampus. *Neuron, 38*(1), 127–133.

Margolis, E., & Laurence, S. (1999). Introduction. In S. Laurence & E. Margolis (Eds.), *Concepts: Core readings* (pp. 3–81). Cambridge, MA: MIT Press.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology, 58*, 25–45.

Martin, A., Wiggs, C. L., Ungerleider, L. G., & Haxby, J. V. (1996). Neural correlates of category specific knowledge. *Nature, 379*, 649–652.

Martin, C. B., & Barense, M. D. (2018). Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *ELife, 7*, e31873.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neo-cortex: Insights from the successes and failures of connectionists models of learning and memory. *Psychological Review, 102*(3), 419–457.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience, 4*(4), 310–322.

McRae, K., & Jones, M. (2013). Semantic memory. In D. Reisberg (Ed.), *Oxford handbook of cognitive psychology* (pp. 206–219). New York: Oxford University Press.

Miller, E. K., Nieder, A., Freedman, D. J., & Wallis, J. D. (2003). Neural correlates of categories and concepts. *Current Opinion in Neurobiology, 13*, 198–203.

Millikan, R. G. (2017). *Beyond concepts: Unicepts, language, and natural information*. Oxford: Oxford University Press.

Mishkin, M. (1997). Hierarchical organization of cognitive memory. *Philosophical Transactions of the Royal Society B: Biological Sciences, 352*(1360), 1461–1467.

Murphy, G. L. (2002). *Big book of concepts*. Cambridge, MA: MIT Press.

Myers, C. E., Shohamy, D., Gluck, M. A., Grossman, S., Kluger, A., Ferris, S., . . . Schwartz, R. (2003). Dissociating hippocampal versus basal ganglia contributions to learning and transfer. *Journal of Cognitive Neuroscience, 115*(2), 185–193.

Nadel, L., & Moscovitch, M. (1997). Memory consolidation and the hippocampal complex. *Cognitive Neuroscience, 7*, 217–227.

O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review, 108*(1), 83–95.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*(7), 739–744.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience, 8*, 976–989.

Peelen, M. V., Bracci, S., Lu, X., He, C., Caramazza, A., & Bi, Y. (2013). Tool selectivity in left occipitotemporal cortex develops without vision. *Journal of Cognitive Neuroscience, 25*(8), 1225–1234.

Peelen, M. V., Romagno, D., & Caramazza, A. (2012). Independent representations of verbs and actions in left lateral temporal cortex. *Journal of Cognitive Neuroscience, 24*(10), 2096–2107.

Prasada, S., & Dillingham, E. M. (2006). Principled and statistical connections in common sense conception. *Cognition, 99*(1), 73–112.

Prasada, S., & Dillingham, E. M. (2009). Representation of principled connections: A window onto the formal aspect of common sense conception. *Cognitive Science, 33*(3), 401–448.

Proklova, D., Kaiser, D., & Peelen, M. V. (2016). Disentangling representations of object shape and object category in human visual cortex: The animate–inanimate distinction. *Journal of Cognitive Neuroscience, 28*(5), 680–692.

Putnam, H. (1973). Meaning and reference. *The Journal of Philosophy, 70*(9), 699–711.

Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophical Review, 60*(1), 20–43.

Renoult, L., Davidson, P. S. R., Palombo, D. J., Moscovitch, M., & Levine, B. (2012). Personal semantics: At the crossroads of semantic and episodic memory. *Trends in Cognitive Sciences, 16*(11), 550–558.

Renoult, L., Irish, M., Moscovitch, M., & Rugg, M. D. (2019). From knowing to remembering: The semantic–episodic distinction. *Trends in Cognitive Sciences, 23*(12), 1041–1057.

Rey, G. (1983). Concepts and stereotypes. *Cognition, 15*, 237–262.

Richards, B. A., Xia, F., Santoro, A., Husse, J., Woodin, M. A., Josselyn, S. A., & Frankland, P. W. (2014). Patterns across multiple memories are identified over time. *Nature Neuroscience, 17*(7), 981–986.

Riddoch, M. J., & Humphreys, G. W. (1987). Visual object processing in optic aphasia: A case of semantic access agnosia. *Cognitive Neuropsychology, 4*(2), 131–185.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12*(1), 1–20.

Rogers, T. T., Lambon-Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review, 111*(1), 205–235.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 3*, 192–233.

Roy, J. E., Riesenhuber, M., Poggio, T., & Miller, E. K. (2010). Prefrontal cortex activity during flexible categorization. *Journal of Neuroscience, 30*(25), 8519–8528.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by eight-month-old infants. *Science, 274*(5294), 1926–1928.

Schacter, D. L., & Tulving, E. (1994). *Memory systems*. Cambridge, MA: MIT Press.

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus, 26*(1), 3–8.

Sekeres, M. J., Winocur, G., & Moscovitch, M. (2018). The hippocampus and related neocortical structures in memory transformation. *Neuroscience Letters, 680*, 39–53.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge: Cambridge University Press.

Shohamy, D., & Daw, N. D. (2015). Integrating memories to guide decisions. *Current Opinion in Behavioral Sciences, 5*, 85–90.

Simanova, I., Hagoort, P., Oostenveld, R., & van Gerven, M. A. J. (2012). Modality-independent decoding of semantic information from the human brain. *Cerebral Cortex, 24*(2), 426–434.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24*, 1193–1216.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Squire, L. R., Knowlton, B. J., & Musen, G. (1993). The structure and organization of memory. *Annual Review of Psychology, 44*(1), 453–495.

Striem-Amit, E., Wang, X., Bi, Y., & Caramazza, A. (2018). Neural representation of visual concepts in people born blind. *Nature Communications, 9*(1), 5250.

Tompary, A., & Davachi, L. (2017). Consolidation promotes the emergence of representational overlap in the hippocampus and medial prefrontal cortex. *Neuron, 96*(1), 228–241.e5.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–402). New York: Academic Press.

Tulving, E. (1984). Précis of elements of episodic memory. *Behavioral and Brain Sciences, 7*(2), 223–238.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology, 53*, 1–25.

Tulving, E., Hayman, C. A. G., & Macdonald, C. A. (1991). Long-lasting perceptual priming and semantic learning in amnesia: A case experiment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(4), 595–617.

Verfaellie, M., Koseff, P., & Alexander, M. P. (2000). Acquisition of novel semantic information in amnesia: Effects of lesion location. *Neuropsychologia, 38*(4), 484–492.

Wang, X., Peelen, M. V., Han, Z., He, C., Caramazza, A., & Bi, Y. (2015). How visual is the visual cortex? Comparing connectional and functional fingerprints between congenitally blind and sighted individuals. *Journal of Neuroscience, 35*(36), 12545–12559.

Warrington, E. K. (1975). The selective impairment of semantic memory. *The Quarterly Journal of Experimental Psychology, 27*(4), 635–657.

Warrington, E. K., & Taylor, A. M. (1978). Two categorical stages of object recognition. *Perception, 7*, 395–401.

Watson, C. E., Cardillo, E., Ianni, G., & Chatterjee, A. (2013). Action concepts in the brain: An activation-likelihood estimation meta-analysis. *Journal of Cognitive Neuroscience, 25*(8), 1191–1205.

Westmacott, R., Black, S. E., Freedman, M., & Moscovitch, M. (2004). The contribution of autobiographical significance to semantic memory: Evidence from Alzheimer's disease, semantic dementia, and amnesia. *Neuropsychologia, 42*(1), 25–48.

Wiggs, C. L., Weisberg, J., & Martin, A. (1999). Neural correlates of semantic and episodic memory retrieval. *Neuropsychologia, 37*(1), 103–118.

Wilson, M. (2008). *Wandering significance: An essay on conceptual behavior*. New York: Oxford University Press.

Winocur, G., Moscovitch, M., Rosenbaum, R. S., & Sekeres, M. J. (2010). An investigation of the effects of hippocampal lesions in rats on pre- and postoperatively acquired spatial memory in a complex environment. *Hippocampus, 20*(12), 1350–1365.

Yee, E., Chrysikou, E. G., Hoffman, E., & Thompson-Schill, S. L. (2013). Manual experience shapes object representations. *Psychological Science, 24*(6), 909–919.

Yee, E., & Thompson-Schill, S. L. (2016). Putting concepts into context. *Psychonomic Bulletin and Review, 23*, 1015–1027.

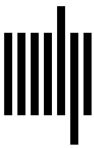
Neuroscience and Philosophy

Edited by: Felipe De Brigard, Walter Sinnott-Armstrong

Citation:

Neuroscience and Philosophy
Edited by: Felipe De Brigard, Walter Sinnott-Armstrong
DOI: 10.7551/mitpress/12611.001.0001
ISBN (electronic): 9780262367332
Publisher: The MIT Press
Published: 2022

The open access edition of this book was made possible by generous funding and support from MIT Press Direct to Open



The MIT Press

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif by Westchester Publishing Services. .

Library of Congress Cataloging-in-Publication Data

Names: Brigard, Felipe de, editor. | Sinnott-Armstrong, Walter, 1955– editor.
Title: Neuroscience and philosophy / edited by Felipe De Brigard and Walter Sinnott-Armstrong.
Description: Cambridge, Massachusetts : The MIT Press, [2022] | Includes bibliographical references and index.
Identifiers: LCCN 2021000758 | ISBN 9780262045438 (paperback)
Subjects: LCSH: Cognitive neuroscience—Philosophy.
Classification: LCC QP360.5 .N4973 2022 | DDC 612.8/233—dc23
LC record available at https://lccn.loc.gov/2021000758