Representations of knowledge – how the brain brings to mind

Human experience is defined by our marked ability to learn about the world and to make meaningful interactions with the things around us. As we grow and develop, we learn that a dog is a friendly animal that is similar yet distinct from a cat. That a tiger – although a cat – is dangerous and to be feared; that it is expected to appear amongst trees in the jungle. We somehow master the art of using a knife, whose function can change multiple times during as simple a task of making a sandwich as we cut bread, slice meat and scoop and spread butter. All of these are examples of knowledge acquired through life, and this must be housed in some way within the neurocognitive processes of our brain. Questions on the nature of knowledge go back as far as history can track, and the conversation spans multiple research fields from philosophy and psychology to cognitive science, neuroscience and computer intelligence. In this essay, I will present an integrative account of the research into human knowledge acquisition, discussing various ideas and models from a range of disciplines. What insights can we get from philosophical theories of knowledge, and on the other end, what evidence do we have for neural mechanisms of knowledge? The question of knowledge itself is a vast topic that is impossible to cover completely in one coherent review, but the aim of this work is to provide an overview and introduction to the key ideas spanning across different fields' exploration of this puzzling topic.

What is knowledge? What do we mean by a representation?

Firstly, we must address what is meant by knowledge. If we take insights from epistemology, we observe that the ongoing debate regarding the definition of knowledge highlights how difficult a task it is to fully capture its meaning. Although classically defined by Plato as being something that is justified, true and believed, many struggle to fully accept this definition. Much later, Ludwig Wittgenstein suggested that knowledge may be a case of family resemblance; it is a means of clustering concepts together so that we can be pointed towards relevant features. But when we say that we know something, exactly what type of knowledge are we talking about? One can know what something is and how to use it, or, one can know someone else. Knowledge can change given a context. Although often discussed in terms of categorisation behaviour, especially in the computer intelligence literature, knowledge isn't merely just attaching a label to a word. Humans are capable of expansive abstract knowledge that generalises across domains and examples. We have a grasp of concepts such as the fact that the phrase "bread and butter" just *feels* more correct than "butter and bread" (Morgan & Levy, 2016). Clearly, even defining knowledge is tricky, thus, for the purposes of this argument I point instead towards typical definitions of knowledge as it pertains to memory of useful information.

Let us consider here knowledge as something that can be accessed from our long term memory and employed for a specific function. This leads naturally to the branch of memory called explicit semantic memory. It concerns knowledge of general facts and information about the world, resulting in a generalisable understanding of what things mean. This lacks specificity in time or place (Tulving, 1972) in contrast to episodic memory of experiences and autobiographical events, an equally interesting type of knowledge that will not be discussed here. Indeed, semantic cognition refers to our ability to use, manipulate and generalise knowledge as we experience our life and environment, eloquently described by Lambon Ralph and colleagues as the thing that "transforms the sensory cacophony into a symphony of meaning" (Lambon Ralph et al., 2016). This serves to support and enable our behaviour, both verbal and non-verbal - although much research is in the lexical domain – and enriches our human experience so much so that the neurodegenerative loss of such knowledge in certain types of dementia can severely impact upon one's quality of life. By narrowing the discussion of knowledge and its representation to semantic knowledge and cognition, we enable a thorough transdisciplinary discussion of this age-old topic.

Knowledge as concepts, representations and relations – from philosophy to psychology

In order to know something, the mind must be able to house the content. The information must be transduced from physical, sensory or experiential input into abstract or long-term knowledge. Thus emerges the idea of a mental representation which, although commonly used in cognitive science, is not universally accepted (see Rowlands (2017), for a review of this argument). Representational theory of mind, which dates back to Aristotle, describes how cognitive states and processes are constituted by the occurrence, transformation and storage in the mind and brain of information-bearing symbols or structures called representations (Pitt, 2020). In this way, representational states can be activated and used when we encounter certain stimuli or environmental cues. For example, the perception of a dog involves the observation of many types of input from the animal's motion and the textures of its fur to the sound of its bark. All of this sensory input must be interpreted and transduced by the brain, and represented in a cognitive or neural mechanism. Indeed, this conceptual representation houses myriad semantic information such as what a dog is, how it relates to other animals (for example, that it is similar to a cat in that it is expected to be a pet) and perhaps even emotionally valent memories of one's own dog. Thus, a mental representation is not as simple as just the fact that the perceived object is attached to the verbal label "dog". The representation houses a rich semantic knowledge for the concept that extends beyond the basic functional output of attaching the label to the referent.

What exactly are concepts, and how are they handled by the brain? This attempt at defining a concept and discussing how we represent them requires the introduction of some theories for knowledge representation, stemming from the philosophical debate and psychological evidence. The classical theory of concepts relies on the idea of a definitional structure whereby concept C is composed of simpler concepts that express C's necessary and sufficient conditions (Margolis & Laurence, 2019). For example, the concept triangle is represented by its definition of having three edges and three vertices. Despite all successive theories stemming from the definition-based classical theory, it falls short in many ways. Mainly, most attempts to find successful definitions for a concept fail and it does not account for numerous empirical findings from psychological experiments.

One such finding is the highly-reproduced typicality effect. This reveals that certain members of a category are seen to be more representative of that category than others. For example, reaction times are faster for responding 'True' to the sentence 'A robin is a bird' than to the sentence 'A chicken is a bird', suggesting that a robin is seen as more typical for the concept of bird (McCloskey & Glucksberg, 1978). Although the classical theory is not exactly inconsistent with such findings, it simply does nothing for explaining them. Instead, Eleanor Rosch's prototype theory adds some explanatory power to our idea of a concept by taking a probabilistic rather than definitional approach (Eleanor Rosch & Mervis, 1975).

Rosch and Mervis propose that members of a category share family resemblance – harking back to Wittgenstein’s philosophical discourse of the same (Wittgenstein, 1958). In their psychological experiment, participants were asked to list all attributes they could for 20 examples of 6 categories across a range of typical examples. For example, the category ‘Furniture’ with the exemplar ‘Chair’ being the most typical and ‘Telephone’ the least. This experiment showed that there were actually very few properties that were shared by all instances of the category, which would have been expected by the definitional approach of classical theory. Instead, more typical members shared many attributes with more members of the group and this degree of feature similarity could be quantified in what the authors called a family resemblance score. In other words, when family resemblance was high, so too was typicality rating. In this way, a prototype for a category can be found which is the item that has the highest overall family resemblance to other members of the category, or is simply the statistically average member of the category. Thus, Rosch’s prototype theory argues that a mental prototype is formed that represents the average picture of a concept, even if such an average has never been experienced. This facilitates generalisation to other members of a category as the mind represents the concept in a probabilistic feature-based manner. Note that the prototype model inherently fails to capture the spread of concepts’ exemplars, which can be highly-variable, because of its reliance on finding an average prototype. An alternate theory to this is the exemplar approach, the power of which is illustrated by the experiments of Storms *et al.* (2000). It was found that the similarity between instances of a category was a better predictor of categorisation performance than Rosch and Mervis’ (1975) family resemblance. This provides conflicting empirical evidence to the prototype theory, instead suggesting that concepts are represented by stored examples that are all linked to the category name.

The feature-based approach of Rosch’s prototype model is nonetheless attractive, and is further formalised in Moss, Tyler and Taylor’s conceptual structure account (Taylor et al., 2007). The focus here is on the internal structure of a concept, which comprises its features that have been shown to have the most prominent effects. Of particular interest are the relationships between features and the degree of correlation between them. Using co-occurrence of semantic properties as a key relation of concepts is interesting, as it links to evidence from infant statistical learning (Saffran & Kirkham, 2018) and computational distributional semantics (Bruni et al., 2014) whereby the statistical regularities of the environment are learned by an agent so as to glean organising structure, relationships and therefore meaning from the world. Relational co-occurrence structure of this sort has also been shown in the visual domain, suggesting that it is not a route to meaning in language alone (Sadeghi et al., 2015). This idea of representing a concept by a type of similarity comparison accounts much better for observed typicality effects, but falls short when we extend the idea of a concept to more abstract judgements that require reflection such as goal-directed ad-hoc concepts e.g. "things to bring to the beach". These often lack clear prototypes, and instances of the concept can exhibit widely different characteristics which makes feature comparison very difficult.

Murphy & Medin (1985) illustrate this idea for the concept of 'drunken actions', describing how, even in the absence of underlying feature similarity, people can coherently implicate their knowledge of intoxication to enable classification of a drunk individual based on their behaviour. They argue that concepts must fit an underlying theory about the world. This 'theory-theory' of concepts describes classification as being similar to scientific theorising in which causal relations are particularly important for making judgements about category membership (Margolis & Laurence, 2019). It is well-suited for explaining the more abstract or reflective types of categorisation that the prototype theory fails to explain, but it is still flawed in that it has difficulty explaining how different people come to represent concepts in such similar ways, despite hugely different experiential input over time. Furthermore, the theory-theory does little for describing the influences of sensory information on our concepts, an important consideration that is lacking from any of the theories presented above.

Grounding concepts in our senses – embodied knowledge

Traditional approaches to information processing actually assume that the representation of knowledge is in the form of an amodal, internal symbol system that lies independent from the brain's modal sensory regions. There have always been arguments in opposition to this amodal view (Markie, 2017). While the rationalist views of Plato and more modern philosophers such as Descartes, Leibnitz and Kant argue against concepts that are grounded in sensory experience, a complete lack of consideration for sensory components seems unfounded. The idea that modal inputs are important for the mind's representations of knowledge goes back once again to ancient philosophers such as Aristotle. More recent empiricists including Locke and Hume argue that all concepts should be derived from sensory experience. The importance of perception is stated in David Hume's principle of association, as he believed that all knowledge is derived from experience and must be analysable in terms of perceptual content (Hume, 2003). His views went so far as to refute the existence of any innate ideas or theories. More recently, Lawrence Barsalou presents influential scientific arguments for knowledge being grounded in experience. His grounded cognition framework wholly rejects the need for amodal symbols and instead argues for an embodied view of concepts in which all knowledge comes from some form of sensory input (Barsalou, 2008).

Grounded cognition states that it is very unlikely that the brain contains the types of amodal symbols that standard theories of cognition often assume. It places an importance on the idea of simulation, the re-enactment of the perceptual, motor and introspective states that are acquired during experience and subsequently re-activated when we need to draw upon our knowledge. In this way, all knowledge is thought to have a basis in some sensory component. There is plenty of evidence that sensory and motor systems contribute to conceptual representation. For example, words such as “kick” or “lick” will result in motor cortex activation in the same somatotopic areas that would be activated when executing the action (Hauk et al., 2004). For the case of conceptual knowledge, Barsalou (2008) argues that there is plentiful empirical evidence corroborating the idea of simulation for conceptual processing. Referencing behavioural, lesion and neuroimaging studies he illustrates that the brain's perceptual areas in posterior regions are involved in conceptual processing and therefore simulation must be playing a role (Barsalou, 2008). However, it is arguable that the evidence provided doesn't necessarily preclude the involvement of what traditional approaches would call a symbol.

Indeed, there is undeniable evidence that perceptual brain regions are activated when recollecting and retrieving perceptual knowledge (e.g. Kellenbach et al., 2001) but that is not to say an entirely grounded mental state, as described by Barsalou, is being simulated upon retrieval. It is well-established that the early visual cortex receives top-down influence from frontal regions that facilitates object recognition (Bar et al., 2006) and top-down feedback provides a means of shaping attention towards task-relevant stimuli (Vetter et al., 2014). This is an inherently different type of state than would be activated during purely bottom-up perceptual experience, and begs the question: if top-down input from abstract attention-guiding regions is so important for modal regions' activation patterns during conceptual processing, then surely a completely grounded view is difficult to support? At some level, this signal must be considered amodal even if it is not the entirely symbolic representation favoured by rationalists. Furthermore, there is growing neural evidence for the phenomenon of replay within the hippocampus during which re-activation of neural sequences that were active in behaviour are thought to support consolidation, memory formation and retrieval (Pfeiffer, 2020). This replay is consistent with idea of simulation from grounded cognition but, if anything, it places the phenomenon in the abstract, amodal region of the hippocampus and not sensory cortex. Nonetheless, the involvement of modality-specific sensory experience for forming conceptual representations described by grounded cognition is important to consider. Although the details of this theory are often debated and questioned, I would argue that a completely symbolic view is equally unlikely; the first cortical port-of-call for any knowledge-forming experience is the sensory regions, and some element of grounded cognition must be present to build a representation.

Where is semantic knowledge represented in the brain? – Discussions from neuroscience

Any cognitive model or theory must have a neural basis. There are proven regions that are involved in the process of knowledge representation (Kiefer & Pulvermüller, 2012) as introduced in our discussion of grounded cognition above. Results from clinical lesion studies, neuroimaging and neuropsychology all give insight into the neural underpinnings of the theories for knowledge representation put forward by philosophers and behavioural psychologists. There are known category-selective regions in ventral temporal cortex (Downing et al., 2006; Kanwisher et al., 1997; Tanaka, 1996) and there is evidence that the ventral visual stream houses diffuse networks of semantic categories (Huth et al., 2012). However, insights from cases in which high-level conceptual processing goes wrong provide compelling evidence for an integrated site of semantic knowledge in the temporal poles.

Semantic dementia (SD) is a neurodegenerative disease in which patients exhibit a behavioural deficit in the ability to access conceptual knowledge despite otherwise preserved cognitive function. For example, an SD patient may fail to retrieve the word 'dog' when shown an image of a dog, and they may not even have an understanding of what it is they are being shown. This cognitive decline is accompanied by extremely specific atrophy of the anterior temporal lobe (ATL), which worsens progressively with age and disease progress (Czarnecki et al., 2008). Imaging of SD patients versus age-matched healthy controls shows a reliable bilateral atrophy of the temporal pole, with more pronounced degeneration in the left hemisphere (Mummery et al., 2000). Moreover, the degree of semantic memory impairment in patient groups correlates with the extent of atrophy in the left ATL, proving a clear link between pathological and behavioural disease profile. This finding is highly reliable and the pathology is very predictable, with patients showing deficits in accessing meaning across all conceptual domains rather than specific categorical deficits (Hodges & Patterson, 2007). Lesion studies implicating the ATL in conceptual processing are corroborated by transcranial magnetic stimulation (TMS) evidence in which inhibition of the ATL leads to an SD-like behavioural profile, with categorisation deficits for both abstract and concrete concepts (Pobric et al., 2009). As such, the temporal pole is widely thought to be an important site of conceptual knowledge processing.

It is interesting to note that the functional loss of the ATL described above affects both abstract and concrete concepts. Although sometimes thought to be the sole site of abstract knowledge representation, emerging evidence instead suggests that the ATL acts as a convergence zone integrating modality-specific inputs to retrieve and recollect concepts (Damasio, 1989). In this way, it can be thought of in terms of Lambon Ralph and colleagues' 'hub-and-spoke' theory of semantic representation. According to this model, the ATL acts as a modality-invariant hub to which sensory, modality-specific 'spokes' communicate bidirectionally through white matter connections (Lambon Ralph et al., 2016). The theory accounts for how coherent and generalisable concepts are built in the mind from sensory experience and how the learned features and concepts are then mapped to broader semantic knowledge making it a more a unifying theory that aligns with Barsalou's grounded cognition as well as symbolic ideas. If we refer pack to Rosch's prototype theory, the features of a concept would be learned and represented in the sensory spokes of the cortex and the ATL then acts as the calculator of typicality or family resemblance. The grounded sensory representations would occupy the spokes, which are then processed by a high-level amodal hub. Functional imaging reveals that, alongside ATL atrophy, there is a reduction in hub-spoke functional connectivity in SD patients, providing further neurophysiological bases for this unifying model of semantic cognition (Guo et al., 2013).

Although the importance of sensory experience seems obvious for typically developing semantic knowledge, recent evidence suggests there is an alternate means of acquiring knowledge representations in the absence of sensory input. By comparing congenitally blind individuals to healthy controls using resting-state fMRI and behavioural data, it was shown that the two groups possess a significantly similar representational space for object-colour knowledge e.g., that a cherry is red, and this is more similar to apples which are also red than to oranges (Wang et al., 2020). Despite this, there were subtle differences in the neural underpinnings of the two groups' semantic representations. While both exhibited activation of the typical semantic processing areas including ATL, the congenitally blind individuals lacked ventral occipitotemporal activation that was present in the sighted group. The functional connectivity of visual nodes was more tightly bound to the ATL in the sighted versus patient group. The authors conclude that there are two distinct types of knowledge representation present in the human brain: the first emerges from sensory-derived codes and the second from unembodied language and cognitively derived codes. They propose that, instead of ATL being the pre-defined location for representing abstract concepts, it is instead the abstractedness of how knowledge was acquired that determines whether it is coded in the ATL. To elaborate, the knowledge of object-colour associations was learned by both groups using abstract, lexical experience and thus resulted in ATL activation during a discrimination task. This learned representation was remarkably similar across participants, regardless of whether they had been blind from birth. However, in the sighted group, this abstract representation was accompanied by a locus of activation in the visual cortex – an alternate means of knowledge representation that results from visual experience. So, if knowledge was learned in an abstract way then it is housed in an abstract region (ATL), and if it was learned with a sensory component it would also be represented in sensory cortex. This is an interesting finding that raises many questions about how we view typical routes to knowledge representation, and the results are especially concerning for theories of grounded cognition. Indeed, the ventral occipitotemporal activation reflects what Barsalou would argue is simulation, but it is difficult to reconcile the strikingly similar behavioural and representational profiles of the congenitally blind individuals who had never received grounded visual input.

Zooming in – knowledge at the systems and computational level

Human lesion and imaging studies provide invaluable evidence for functional, modular and anatomical accounts of cognition in the brain. The results presented above point clearly to posterior sensory regions and temporal association cortices (ATL) as the sites for semantic knowledge processing. However, the strengths of these macro-level methods are accompanied by weaknesses in their ability to give detailed mechanistic evidence for the implementation of cognitive processes. Thus, we turn to computational and cellular systems neuroscience to explore this more zoomed-in level of representation.

The representations of concepts in specialised regions must still have a micro-scale description at the neural level. Traditional approaches often considered the neuron itself as the basic unit of information, assuming that a concept is coded within one node of an entire neural network (Barlow, 1972). This idea was successfully modelled in computational semantic networks which represent meaningful structure between concepts in a hierarchical fashion (Collins & Quillian, 1969; Collins & Loftus, 1975). The connection of nodes within the semantic network explicitly models knowledge of a concept in a symbolic manner, with each node having an associated label that specifies the knowledge content. At different levels of the network, properties will be inherited in such a way that concepts are structured in a hierarchical tree. This localist approach to semantic modelling was influential, but has proven to be quite restrictive. Significant progress in semantic modelling was made with a move towards distributed theories where concepts are coded by multiple representational units, and arise from more widespread activation patterns.

A key distributional framework was proposed by McClelland and Rumelhart called Parallel Distributed Processing (PDP) (McClelland & Rumelhart, 1986). This spawned the field of connectionism, which went on to influence modern artificial neural networks and machine intelligence. Connectionism progressed computational models of knowledge from the seminal semantic networks and switched the focus from the neuron itself as the unit of information, instead considering the wider pattern of synaptic connection weights. As famously summarised by Hebbian theory, “neurons that fire together wire together” meaning that information can be stored from increased connection strength due to correlated activation (Hebb, 2005). In this way, cognitive activities emerge from interactions between large numbers of processing units in a distributed process across many brain regions. Knowledge is stored in the activity patterns of participating neurons within a section, or layer, of a network (Caramazza et al., 1990; Devlin et al., 1998; James L. McClelland & Rogers, 2003; Tyler & Moss, 2001). In PDP framework, learning of concepts necessarily arises through experience. Backpropagation of prediction error shapes stored representations by comparing the expected result to that which actually occurs, a method that is known to take place in the neural circuitry of the brain (Watabe-Uchida et al., 2017). These connectionist networks are sensitive to co-occurrence statistics of features, the same statistical information that is theorised to be important for forming internal structures of concepts by Taylor *et al.* (2007) in their conceptual structure account.

Computational theories use terminology borrowed from cellular neuroscience, and artificial neural networks describe their processing units as neurons. However, modelling alone does not provide definitive evidence for a distributed mechanism of concept representation or memory in the brain. With continued technological advancements in cellular systems level neuroscience such as the advent of opto and chemogenetics, it has been made clear that sparse ensembles of cells encode specific memories, and indeed that memory resides in the stable connectivity patterns between these distributed cells (Harel & Ryan, 2020). This makes exciting connections between computational insights and network neuroscience. It is well-established that neural populations can encode complex probabilistic information (Ma et al., 2006), and network-level coding units within the hippocampus termed neural cliques have been shown to have powerful abilities to abstract and generalise representations of external events, useful for cognitive functions (Lin et al., 2006). Powerful calcium imaging techniques in drosophila have revealed that olfactory associative memories emerge from a distributed synaptic memory code, implicating synaptic boutons as the important modifiable units for information storage (Bilz et al., 2020). This is an attractive result for thinking about cellular implementations of weight updates modelled by connectionist networks, suggesting that the site of weight changes is at a sub-synaptic scale. This reduces reliance on the idea that new connections must emerge from neurogenesis, instead the shapes of dendritic trees and arborisation patterns may house the most fundamental unit of the activation patterns that facilitate knowledge representation (Frank et al., 2018). A true implementation of backpropagation as it appears in connectionist networks is unlikely in the brain. Instead, there is some argument for the role of excitatory and inhibitory balance at the dendritic tree being an important regulator of synaptic weight connections (Iascone et al., 2020). The more obvious link of prediction error updates is in terms of dopamine prediction errors for learning (Holroyd & Coles, 2002; Keiflin & Janak, 2015). Although many questions remain open, the combinatory power of computational and cellular neuroscience for addressing this level of understanding mechanisms for knowledge representation is becoming increasingly exciting.

Engineering knowledge in artificial intelligence

Studies of human knowledge representation are inextricably linked to those in computer science. Computational models are of course widely used to study mechanisms of knowledge representation in the mind, but understanding and engineering knowledge into such systems is an endeavour in itself. In fact, a Google search for "knowledge representation" doesn't give results regarding the brain, but instead brings up entries on artificial intelligence (AI). Indeed, artificial neural networks are heralded in modern machine learning, and deep neural networks have been disruptive in the field of computer vision (Krizhevsky et al., 2012). Not only this, but these networks are increasingly used in neuroscience to study knowledge representations such as those for objects, or in reward-based reinforcement learning (Richards et al., 2019). Although biologically-inspired, these networks and their exact methods of learning do receive heavy criticism regarding biological plausibility (Lillicrap et al., 2020) and one should struggle to make an exact comparison between the artificial networks and the brain.

Furthermore, claims that these networks are achieving human-level behavioural performance should be called into question. In psychology, concepts are traditionally explored through categorisation (e.g. E. Rosch, 1975) and engineers in AI often sing the praises of neural networks because of superior performance in categorisation tasks (He et al., 2015). However, there seems to be an under-appreciated oversight in the field. Often those engaged in the AI scientific community simply accept that knowledge has been acquired because the artificial network is extremely successful at attaching labels to visual or textual stimuli. But this fails to account for the complex semantic concepts that are housed in the human brain. Yes, part of knowing that a tree is a tree is naming it and categorising it, but we cannot claim that an AI actually knows what this means in a wider semantic context. Furthermore, most models are trained in only one modality meaning they can lack grounding of the concept in another (Baroni, 2016). Some highly-cited papers do give further insight into this issue, but I would still argue that true conceptual meaning is missing from state-of-the-art AI. Looking at studies that explore the representational space of these networks, i.e. how similar and dissimilar does it represent objects, there are striking similarities to representations found in the human brain from neuroimaging (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014). Further arguments for the representation of complex concepts in artificial networks comes from linguistic models. Such networks have been shown to be able to perform word analogy problems, for example that ‘Queen’ minus ‘Woman’ plus ‘Man’ equals ‘King’ (Mikolov et al., 2013). It’s enticing to refer to ground-breaking findings such as these when making claims for robust artificial knowledge representations but once again it seems far-fetched to argue that there truly is meaning in these networks. Of course, relations and co-occurrence patterns are thought to be important for human conceptual processing as discussed through the prototype and exemplar approaches above, but human knowledge representation goes beyond this. Even the authors of the latest and greatest in AI, the language model GPT-3, warn against interpreting it as the beginnings of artificial general intelligence (Floridi & Chiriatti, 2020). The cutting-edge in machine intelligence is impressive, but the most hyped advances are hugely reliant on labelling, categorisation and shallow representations of meaning that cannot be compared to the vast abstract semantic capabilities of human cognition.

Conclusion

The question of knowledge representation is undeniably stimulating. A variety of thinkers from philosophy to computer science all strive to elucidate the mechanisms of how we come to know, but much remains to be answered. In this essay, I have attempted to give an integrated, transdisciplinary overview of the topic describing theories of concepts, macro-level evidence from neuropsychology and neuroscience and a network discussion of how representations might be implemented on a neural level. It seems that knowledge emerges necessarily from grounded experience with the world, while its storage in the brain could be in higher amodal centres. Knowledge is distributed and flexible, and no one theory can fully account for how it is acquired. In my own understanding of knowledge representation I find it useful to draw upon elements of each theory discussed here, with the unifying hub-and-spoke model being especially appealing. Finally, although computational accounts of knowledge representations have taught us a lot and are progressing rapidly, true artificial intelligence is still a long way off and building more meaningful semantics into these networks is a worthwhile endeavour. The accounts presented here go much deeper and each offers a wealth of insight. However, given the historical speculation regarding this topic I suspect that we will continue to remain ignorant to the complete workings of knowledge for many years to come.

Bibliography

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmidt, A. M., Dale, A. M., Hämäläinen, M. S., Marinkovic, K., Schacter, D. L., Rosen, B. R., & Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(2), 449–454. https://doi.org/10.1073/pnas.0507062103

Barlow, H. B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception*, *1*(4), 371–394. https://doi.org/10.1068/p010371

Baroni, M. (2016). Grounding Distributional Semantics in the Visual World. *Language and Linguistics Compass*, *10*(1), 3–13. https://doi.org/10.1111/lnc3.12170

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Bilz, F., Geurten, B. R. H., Hancock, C. E., Widmann, A., & Fiala, A. (2020). Visualization of a Distributed Synaptic Memory Code in the Drosophila Brain. *Neuron*, *106*(6), 963-976.e4. https://doi.org/10.1016/j.neuron.2020.03.010

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1–47. https://doi.org/10.1613/jair.4135

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual

Object Recognition. *PLoS Computational Biology*, *10*(12). https://doi.org/10.1371/journal.pcbi.1003963

Caramazza, A., Hillis, A. E., & Romani, C. (1990). The Multiple Semantics Hypothesis: Multiple Confusions? *Cognitive Neuropsychology*, *7*(3), 161–189. https://doi.org/10.1080/02643299008253441

Collins, A.M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240–248. http://www.academia.edu/download/31971250/Collins_Quillian69.pdf

Collins, Allan M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Czarnecki, K., Duffy, J. R., Nehl, C. R., Cross, S. A., Molano, J. R., Jack, C. R., Shiung, M. M., Josephs, K. A., & Boeve, B. F. (2008). Very early semantic dementia with progressive temporal lobe atrophy an 8-year longitudinal study. *Archives of Neurology*, *65*(12), 1659–1663. https://doi.org/10.1001/archneurol.2008.507

Damasio, A. R. (1989). The Brain Binds Entities and Events by Multiregional Activation from Convergence Zones. *Neural Computation*, *1*(1), 123–132. https://doi.org/10.1162/neco.1989.1.1.123

Devlin, J. T., Gonnerman, L. M., Andersen, E. S., & Seidenberg, M. S. (1998). Category-specific semantic deficits in focal and widespread brain damage: A computational account. *Journal of Cognitive Neuroscience*, *10*(1), 77–94. https://doi.org/10.1162/089892998563798

Downing, P. E., Chan, A. W.-Y., Peelen, M. V, Dodds, C. M., & Kanwisher, N. (2006). Domain specificity in visual cortex. *Cerebral Cortex (New York, N.Y. : 1991)*, *16*(10), 1453–1461. https://doi.org/10.1093/cercor/bhj086

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, *30*(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

Frank, A. C., Huang, S., Zhou, M., Gdalyahu, A., Kastellakis, G., Silva, T. K., Lu, E., Wen, X., Poirazi, P., Trachtenberg, J. T., & Silva, A. J. (2018). Hotspots of dendritic spine turnover facilitate clustered spine addition and learning and memory. *Nature Communications*, *9*(1), 1–11. https://doi.org/10.1038/s41467-017-02751-2

Guo, C. C., Gorno-Tempini, M. L., Gesierich, B., Henry, M., Trujillo, A., Shany-Ur, T., Jovicich, J., Robinson, S. D., Kramer, J. H., Rankin, K. P., Miller, B. L., & Seeley, W. W. (2013). Anterior temporal lobe degeneration produces widespread network-driven dysfunction. *Brain*, *136*(10), 2979–2991. https://doi.org/10.1093/brain/awt222

Harel, A., & Ryan, T. (2020). The memory toolbox: how genetic manipulations and cellular imaging are transforming our understanding of learned information. In *Current Opinion in Behavioral Sciences* (Vol. 32, pp. 136–147). Elsevier Ltd. https://doi.org/10.1016/j.cobeha.2020.02.016

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic Representation of Action Words in Human Motor and Premotor Cortex. *Neuron*, *41*(2), 301–307. https://doi.org/10.1016/S0896-6273(03)00838-9

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *Computer Vision Foundation, ICCV*. https://www.cv-foundation.org/openaccess/content_iccv_2015/papers/He_Delving_Deep_into_ICCV_2015_paper.pdf?spm=5176.100239.blogcont55892.28.pm8zm1&file=He_Delving_Deep_into_ICCV_2015_paper.pdf

Hebb, D. O. (2005). *The organization of behavior: A neuropsychological theory*. Psychology Press. https://books.google.com/books?hl=en&lr=&id=uyV5AgAAQBAJ&oi=fnd&pg=PP1&dq=donald+hebb&ots=mKmSArGVRs&sig=GsaXvmLIP8E9rruW02ItWL6WHaw

Hodges, J. R., & Patterson, K. (2007). Semantic dementia: a unique clinicopathological syndrome. In *Lancet Neurology* (Vol. 6, Issue 11, pp. 1004–1014). Elsevier. https://doi.org/10.1016/S1474-4422(07)70266-1

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709. https://doi.org/10.1037/0033-295X.109.4.679

Hume, D. (2003). *A Treatise of Human Nature*. Courier Corporation. https://books.google.com/books?hl=en&lr=&id=zHYO1Fh9_JMC&oi=fnd&pg=PR9&dq=a+treatise+of+human+nature&ots=QglDehm91d&sig=JZ3WuuYuYbcDu3RCt_bcgiZ7ztc

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–1224. https://doi.org/10.1016/j.neuron.2012.10.014

Iascone, D. M., Li, Y., Sümbül, U., Doron, M., Chen, H., Andreu, V., Goudy, F., Blockus, H., Abbott, L. F., Segev, I., Peng, H., & Polleux, F. (2020). Whole-Neuron Synaptic Mapping Reveals Spatially Precise Excitatory/Inhibitory Balance Limiting Dendritic and Somatic Spiking. *Neuron*, *106*(4), 566-578.e8. https://doi.org/10.1016/j.neuron.2020.02.015

Kanwisher, N., McDermott, J., Chun, M. M., Hanke, M., Halchenko, Y. O., Wu, Y.-C., Abdi, H., & Haxby, J. V. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *17*(11), 4302–4311. https://doi.org/10.1523/jneurosci.5547-11.2012

Keiflin, R., & Janak, P. H. (2015). Dopamine Prediction Errors in Reward Learning and Addiction: From

Theory to Neural Circuitry. *Neuron*, *88*(2), 247–263. https://doi.org/10.1016/j.neuron.2015.08.037

Kellenbach, M. L., Brett, M., & Patterson, K. (2001). Large, colorful, or noisy? Attribute- and modality-specific activations during retrieval of perceptual attribute knowledge. *Cognitive, Affective and Behavioral Neuroscience*, *1*(3), 207–221. https://doi.org/10.3758/CABN.1.3.207

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11). https://doi.org/10.1371/journal.pcbi.1003915

Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex*, *48*(7), 805–825. https://doi.org/10.1016/j.cortex.2011.04.006

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems*, *1*, 1097–1105. http://code.google.com/p/cuda-convnet/

Lambon Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2016). They neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*, 42–55. https://doi.org/https://doi.org/10.1038/nrn.2016.150

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, *21*(6), 335–346. https://doi.org/10.1038/s41583-020-0277-3

Lin, L., Osan, R., & Tsien, J. Z. (2006). Organizing principles of real-time memory encoding: Neural clique assemblies and universal neural codes. In *Trends in Neurosciences* (Vol. 29, Issue 1, pp. 48–57). Elsevier Current Trends. https://doi.org/10.1016/j.tins.2005.11.004

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438. https://doi.org/10.1038/nn1790

Margolis, E., & Laurence, S. (2019). Concepts. In *The Stanford Encyclopedia of Philosophy* (Summer 201). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/concepts/#RenIntEmpDis

Markie, P. (2017). Rationalism vs. Empiricism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2017). https://plato.stanford.edu/archives/fall2017/entries/rationalism-empiricism/

McClelland, J.L., Rumelhart, D. E., & Group, P. R. (1986). *Parallel distributed processing. Explorations in the Microstructure of Cognition* (2nd ed.).

McClelland, James L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(4), 310–322. https://doi.org/10.1038/nrn1076

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, *6*(4), 462–472. https://doi.org/10.3758/BF03197480

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119.

Morgan, E., & Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, *157*, 384–402. https://doi.org/10.1016/j.cognition.2016.09.011

Mummery, C. J., Patterson, K., Price, C. J., Ashburner, J., Frackowiak, R. S. J., & Hodges, J. R. (2000). A voxel-based morphometry study of semantic dementia: Relationship between temporal lobe atrophy and semantic memory. *Annals of Neurology*, *47*(1), 36–45. https://doi.org/10.1002/1531-8249(200001)47:1<36::AID-ANA8>3.0.CO;2-L

Murphy, G. L., & Medin, D. L. (1985). The Role of Theories in Conceptual Coherence. *Psychological Review*, *92*(3), 289–316. https://doi.org/10.1037/0033-295X.92.3.289

Pfeiffer, B. E. (2020). The content of hippocampal "replay." *Hippocampus*, *30*(1), 6–18. https://doi.org/10.1002/hipo.22824

Pitt, D. (2020). Mental Representation. In *The Stanford Encyclopedia of Philosophy* (Spring 202). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/mental-representation/

Pobric, G., Lambon Ralph, M. A., & Jefferies, E. (2009). The role of the anterior temporal lobes in the comprehension of concrete and abstract words: rTMS evidence. *Cortex*, *45*(9), 1104–1110. https://doi.org/10.1016/j.cortex.2009.02.006

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., … Kording, K. P. (2019). A deep learning framework for neuroscience. In *Nature Neuroscience* (Vol. 22, Issue 11, pp. 1761–1770). Nature Publishing Group. https://doi.org/10.1038/s41593-019-0520-2

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104(3)*(192). https://psycnet.apa.org/record/1976-00172-001

Rosch, Eleanor, & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Rowlands, M. (2017). Arguing about representation. *Synthese*, *194*(11), 4215–4232.

https://doi.org/10.1007/s11229-014-0646-4

Sadeghi, Z., McClelland, J. L., & Hoffman, P. (2015). You shall know an object by the company it keeps: An investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia*, *76*, 52–61. https://doi.org/10.1016/j.neuropsychologia.2014.08.031

Saffran, J. R., & Kirkham, N. Z. (2018). Infant Statistical Learning. *Annual Review of Psychology*, *69*, 181–203. https://doi.org/10.1146/annurev-psych-122216-011805

Storms, G., De Boeck, P., & Ruts, W. (2000). Prototype and Exemplar-Based Information in Natural Language Categories. *Journal of Memory and Language*, *42*(1), 51–73. https://doi.org/10.1006/jmla.1999.2669

Tanaka, K. (1996). Inferotemporal Cortex and Object Vision. *Annual Review of Neuroscience*, *19*(1), 109–139. https://doi.org/10.1146/annurev.ne.19.030196.000545

Taylor, K. I., Moss, H. E., & Tyler, L. K. (2007). The conceptual structure account: A cognitive model of semantic memory and its neural instantiation. In *Neural Basis of Semantic Memory* (pp. 265–301). Cambridge University Press. https://doi.org/10.1017/CBO9780511544965.012

Tulving, E. (1972). Episodic and semantic memory. *Organization of Memory*, 381–403. http://web.media.mit.edu/~jorkin/generals/papers/Tulving_memory.pdf

Tyler, L. K., & Moss, H. E. (2001). Towards a distributed account of conceptual knowledge. In *Trends in Cognitive Sciences* (Vol. 5, Issue 6, pp. 244–252). Elsevier Current Trends. https://doi.org/10.1016/S1364-6613(00)01651-X

Vetter, P., Smith, F. W., & Muckli, L. (2014). Decoding sound and imagery content in early visual cortex. *Current Biology*, *24*(11), 1256–1262. https://doi.org/10.1016/j.cub.2014.04.020

Wang, X., Men, W., Gao, J., Caramazza, A., & Bi, Y. (2020). Two Forms of Knowledge Representations in the Human Brain. *Neuron*, *107*(2), 383-393.e5. https://doi.org/10.1016/j.neuron.2020.04.010

Watabe-Uchida, M., Eshel, N., & Uchida, N. (2017). Neural Circuitry of Reward Prediction Error. *Annual Review of Neuroscience*, *40*(1), 373–394. https://doi.org/10.1146/annurev-neuro-072116-031109

Wittgenstein, L. (1958). *Philosophical investigations* (G. E. M. Anscombe (trans.) (ed.); 3rd Editio). Oxford: Blackwell.