Understand ‘understanding’ with abstraction, counterfactuals, mental models, and the mindsponge

Ho Manh Tung

Ritsumeikan Asia Pacific University,

Beppu, Oita, Japan

December 20, 2020

What does it mean to understand? Everyone knows what it feels like to “finally get something” or have an “aha” moment. Yet, until now, there is no complete account of the mental mechanisms that explain human understanding (Mitchell, 2019). Melanie Mitchell (2019) devotes two chapters in her book *Artificial intelligence: A guide for thinking humans* to this problem and putting it in the context of A.I. engineering. Indeed, we live in a world where A.I. systems that lack human-like understanding have an increasing amount of input into our decisions. Thus, solving the problem of how human makes sense of the world is not purely theoretical anymore; it has real applications in the world now. In this essay, we will look at some compelling theories of human understanding. I capture the essential components of Mitchell’s (2019) account and add some ideas I thought might benefit this account along the way.

Mitchell’s account of understanding

Our understanding requires basic building blocks. Research into babies’ psychological development suggests human babies seem to have innate, bare-bone theories of the world. Those theories are called intuitive physics, intuitive biology, intuitive psychology (Pinker, 2003; Bloom, 2013). Some examples of those basic intuitions are when we drop something, it falls, or

living things are different from non-living things, people have beliefs, etc. Those core commonsense bodies of knowledge are the building blocks to generalize, predict, and act in the world.

Understanding is about predicting and simulating. Mitchell (2019) slightly differentiates between understanding as prediction and understand as simulation. However, it appears prediction is a sub-category of simulation. Thus, grouping these two aspects is more straightforward. The “*understanding as simulation*” hypothesis is proposed by psychologist Barsalou (1999, 2008). So when we understand a new concept or phenomenon, we are capable of simulating its as mental models. To understand, we must be capable of simulating a phenomenon’s different outcomes when different elements are added or subtracted from the original phenomenon. That is to say, in my opinion, the simulation aspect of understanding involves our ability to reason with the counterfactuals, which is extensively explored in the work of Judea Pearl (Pearl, 2019; Pearl and Mackenzie, 2018).

Understanding is about making metaphors and analogies from our commonsense physical knowledge. Mitchell (2019) uses the arguments and evidence in *Metaphors we live by*, written by philosopher Mark Johnson and linguist George Lakoff (2008), to make a case that our understanding of abstract concepts (such as time, love, sorrow, richness, etc.) is about making metaphors from our basic physical knowledge of the world. In recent years, the argument for the physical or spatial basis of our abstract knowledge has gained steam with empirical evidence from brain sciences. A 2014 Nobel Prize in Physiology or Medicine was awarded to a paper that explains how the “place” cells work, the cells in the hippocampus that code places in the world, for example. The mechanisms behind those place cells help explain how abstract concepts such as home entwine with emotions: a home is a place that evokes particular feelings (Costandi, 2013).

This thinking line can be called “the embodied cognition hypothesis,” that states our understanding is formed by the body acting in space. This way of thinking about thinking becomes more recognized in a Science paper titled *Navigating cognition: Spatial codes for human thinking* (Bellmund, Gärdenfors, Moser, & Doeller 2018). It is also more popularized with a recent book by Barbara Tversky called *Mind in Motion: How action shapes thoughts* (2019).

Understanding as an act of creative problem-solving

Putting everything together, Mitchell (2019)’s account essentially generates a picture of understanding as a creative problem-solving or a knowledge-creating process with the building blocks are our basic physical experiences of the world (movements, directions, temperature, pressure, depth, etc.).

The myth of perceptual-to-conceptual shift in understanding. In David Deustch’s account of knowledge creation, which is inspired by Popper’s conjecture and refutation, all observations are theory-laden (Deustch, 2011). What it means is that we learn about the world not through a shifting process from perception to concept, but our learning is conceptual from the start. This view resonates in a recent discussion of how infants learn about emotions (Hoeman, Xu & Barrett, 2019; Hoeman, Devlin & Barrett, 2020) and the view that the brain functions as an internal model of the body, generating prediction about future’s sensory input by drawing on from past experiences (Barrett, 2017).

Entering the mindsponge. To understand something, we first need to form a mental model of it and simulate it. The simulation outcomes are compared with outcomes of existing mental models to judge whether the new model is worth keeping. Here, I think it is worth introducing the concept of a mindsponge, devised by Vuong and Napier (2015), to explain the process of adopting or rejecting new cultural values. As each value is indeed a special kind of

abstract knowledge, the mindsponge model can be useful in explaining the process of understanding as well (Vuong, 2016).

In the original mindsponge model by Vuong and Napier (2015), the mind is conceived as a sponge that has a core (a set of fundamental values) and a buffer zone, where new ideas are attached to and subjected to a filtering process. To speak in the language of mental models and simulation, we can say the core of a mindsponge is where the layers of fundamental sets of mental models of the world reside. Those mental models have been tried and proved to be useful to the person. When a new value or an abstract knowledge is encountered by a mindsponge, the filtering process will happen as the mind builds new mental models of the world with that new knowledge and simulates them against the existing mental models in the core.

Understanding then happens when there is a match between the new and some of the existing mental models. When there is a mismatch, there is a feeling of confusion and doubt, which signals the need for further mental model building. The mindsponge analogy highlights the complexities in the process of adopting and rejecting a new understanding of the world. For example, there is a clear potential for the core of the mindsponge to be taken over by some mental models that self-justified and resisting to change. This is the reminder of the power of memes (Dawkins, 1989). As we all know, certain memes are very resistant to change and can hold back improvement in understanding.

And since the core can be expanded and revised with new knowledge, it means the building blocks of understanding might be our physical experiences, yet, I suspect understanding in adults rely on building new mental models using a higher level of abstraction, as it would be more cost-effective.

As the original mindsponge model deals with value, which is a special kind of knowledge with particular imperatives and emotions associated with it. And since emotion is likely abstract and conceptual categories (Barett, 2017), one can see how emotions, as abstract mental models

the brain generate to predict the sensory input from the body to interact with the world cost-effectively, can hijack the core of the mindsponge, and thus hijacking the understanding process. Cost-efficiency and inductive attitude are parts of the mindsponge model of the value-filtering process.

This essay has reflected on Mitchell's (2019) account of human understanding and linked it with other theories that complement it. However, I would like to end with this thought. Explaining the mental mechanisms of understanding is still an easier problem compared with explaining the phenomenology of understanding. The feeling of "what it is like to understand something" is perhaps the most natural aspect of the human experience. Yet, it appears to be a complete mystery when we try to explain that feeling in terms of some mechanistic mechanisms. Mitchell (2019) posits we form our understanding by analogizing from our physical, spatial experiences. But as our consciousness mediates our physical experiences, there needs to be further open, proactive research (Vuong, 2019, 2020) that addresses the phenomenology of understanding.

References

- Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1), 1-23. doi:10.1093/scan/nsw154
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4), 577-660. doi:10.1017/S0140525X99002149
- Barsalou, L. W. (2008). Grounded cognition. *Annual Reviews of Psychology*, 59, 617-645.
- Bellmund, J. L. S., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science*, 362(6415), eaat6766. doi:10.1126/science.aat67

Bloom, P. (2013). *Just babies: The origins of good and evil*. New York: Broadway Books.

Costandi, M. (2013). Emotional Renovations: How your brain twists together emotion and place. *Nautilus. Neuroscience*. Retrieved from <http://nautil.us/issue/8/home/emotional-renovations>

Dawkins, R. (1981). Selfish genes and selfish memes. In: *The Mind's I: Fantasies and Reflections on Self and Soul*, Hofstadter, D. R. & Dennet, D. C. (eds). New York, Basic Books, pp. 124-144.

Deutsch, D. (2011). *The Beginning of Infinity: Explanations that Transform the World*. London, UK: Penguin.

Hoemann, K., Devlin, M., & Barrett, L. F. (2020). Comment: Emotions Are Abstract, Conceptual Categories That Are Learned by a Predicting Brain. *Emotion Review*, 12(4), 253-255. doi:10.1177/1754073919897296

Hoemann, K., Xu, F., & Barrett, L. F. (2019). Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Developmental psychology*, 55(9), 1830.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. Chicago: University of Chicago Press.

Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans*. London, UK: Penguin.

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3), 54-60.

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Pinker, S. (2003). *How the Mind Works*. London: Penguin U.K.

- Rajagopalan, A. New evidence for the strange geometry of thought. *Nautilus. On biology*. Retrieved from <http://nautil.us/blog/new-evidence-for-the-geometry-of-thought>
- Tversky, B. (2019). *Mind in Motion: How Action Shapes Thought*. London: Hachette UK.
- Vuong, Q. H. (2016). Global mindset as the integration of emerging socio-cultural values through mindsponge processes: A transition economy perspective. In: J. Kuada (Ed.) *Global Mindsets* (pp. 123-140). London: Routledge.
- Vuong, Q.-H. (2019). Breaking barriers in publishing demands a proactive attitude. *Nature Human Behaviour*, 3(10), 1034-1034. doi:10.1038/s41562-019-0667-6
- Vuong, Q. H. (2020). Reform retractions to make them more transparent. *Nature*, 582, 149.
- Vuong, Q. H., & Napier, N. K. (2015). Acculturation and global mindsponge: an emerging market perspective. *International Journal of Intercultural Relations*, 49, 354-367.