What is Learning? A Definition for Cognitive Science

Jim Davies (jim@jimdavies.org)
2202B Dunton Tower, Institute of Cognitive Science, Carleton University
1125 Colonel By Dr., Ottawa, Ontario, K1S 5B6, Canada

Abstract

Many intuitive notions of "learning" do not support the diverse kinds of learning across different situations and learners. In this paper I offer a functional definition of learning from a cognitive science perspective, which attempts to account for the presence of learning in different physical substrates. The definition is that a particular event should be considered a good example of "learning" to the degree to which the following characteristics describe it: 1) a system undergoes change to its informational state or processing 2) the change is for the purpose of more effective future action, 3) the change is in response what the system experiences, and 4) the system executes the change, rather than some outside force. Episodes are better examples of learning according to how many of these characteristics they have. I discuss benefits and limitations of this characterization.

Keywords: learning; philosophy; conceptual analysis; cognitive science; functionalism; substrate neutrality

Introduction

According to Daniel Reisberg (Wilson & Keil, 2001), learning "can be understood as a change in an organism's capacities or behavior brought about by experience." The *Oxford Companion to Philosophy* defines it as "the acquisition of a form of knowledge or ability through the use of experience." These examples are reasonable and intuitive first passes, but are not defended.

Perhaps the simplest definition of learning would be "the creation of memory," but this merely pushes the definitional difficulty to the term "memory." Nevertheless, this discussion will assume that all memory-creating processes are examples of learning (as a sufficient condition), though I will not use the term in the definition.

In this article, I will present and defend a definition of "learning" for cognitive science. My goals are that this definition will cover all accounts of learning that we observe in natural and artificial systems, and reject cases of change that should not be considered learning.

My approach assumes a version of functionalism as applied to mental concepts: that many entities in our world should be defined not by their physical properties, but by how they interact in an information processing system.

The definition is that a particular event should be considered a good example of "learning" according to the degree to which the following characteristics describe it: 1) a system undergoes change to its informational state or processing, 2) the change is for the purpose of more effective future action, 2) the change is in response what the system experiences, and 3) the system executes the change, rather than some outside force. This is a "family resemblance" characterization, rather than one of necessary and sufficient conditions, though in this paper I will discuss the characteristics *as though* they were necessary for purposes of clarifying the benefits and drawbacks of including each one.

What Can Learn?

Learning is prototypically thought of as something animals do. But some plants have a limited form of memory, and the encoding of this memory can be considered a form of learning. The venus flytrap, for example, has hairs around its trap to detect the presence of food. These hairs have haptic sensors. But the trap will not close immediately upon triggering these sensors, which is good, because closing and opening the trap is expensive in terms of energy and, for opening it again, time. So the plant will only close when another sensor detects touch within 20 seconds of the first touch elsewhere—effectively detecting a bug walking across the plant. This prevents the trap from closing when hit with raindrops, twigs, or other non-food entities (Chamovitz, 2012). This is a very simple, very short-term memory. But even in humans we do not require that encodings be long-term to be considered memory, such as the phonological loop (Baddeley, 1992). Because we classify some explicitly short-term stores as memory in humans suggests that it is

sensible to refer to flytraps as systems that can, because they can create memories, learn.

Perhaps even more surprising are examples of learning recently discovered in single-celled organisms (Boisseau, Vogel, & Dussutour, 2016). Even slime molds can habituate to stimuli.

Our immune systems effectively remember past experiences to better deal with future infections. Immune system learning behaves a bit like classifier systems in artificial intelligence (Farmer, Packard, & Perelson, 1986), and can even overlearn, as seen in autoimmune disorders.

Beyond the vast variety of organisms that can learn, we also have an entire field of machine learning to consider: pieces of software created by human beings that learn. Neither immune systems, plants, AIs nor slime mold cells have nervous systems, but they are capable of limited kinds of learning, suggesting that there should not be a biological, let alone a neuronal, condition.

Human Learning at Different Levels of Analysis

When we examine human learning, we can see it happening at many levels of analysis. I will use a running example of learning to avoid eating food that makes one ill.

Neuroscientists now know a lot about how association and feedback can change neurons and how they communicate. This can happen, for example, through synaptic changes: neurons encode association through long-term potentiation and depression (associative learning; Cooke & Bliss, 2006), and can engage in supervised learning (Ishikawa, Matsumoto, Sakaguchi, Matsuki, & Ikegaya, 2014). An immediate nausea response could trigger an instance of supervised learning, "punishing" neurons that were involved with consuming the food, and their relationships to sensing that that food was present.

Synapatic changes in taste receptors allow us to habituate to bitter foods and drinks—children sometimes vomit when they first taste the foods that many adults enjoy. We evolved to dislike bitter foods, generally speaking, because they are more often lacking in nutrition (Sandell & Breslin, 2006). Eating bitter foods that don't sicken us gradually habituates the sensors in our tongue to the particular taste.

In addition to synaptic changes, the brain learns through creation, movement, destruction, and the changing of the shape of neurons.

Moving up to the information processing level, every major cognitive architecture has a theory of learning. The most popular control system used in cognitive architectures is the production system (e.g., ACT-R, Soar, EPIC, and OpenCog all use them). When something bad happens, recently-fired productions are "punished," making the system less likely to get itself in the same situation again. In connectionist architectures, learning changes connection weights in neural networks using learning algorithms such as backpropagation (Chauvin & Rumelhart, 1995).

At the behavioral level, we can describe a person's reluctance to eat a food that previously made them sick with the theory of conditioning.

I have shown how the same event of an *individual* agent learning not to eat a particular food can be effectively described as learning at different levels, but we might describe other examples of learning in distributed cognitive systems (Hutchins, 1995). A theater company might better learn how to market its performances, or a game development team might better learn how to use feedback from user testing to make better products.

The idea of distributed cognitive systems (and the related notion of extended minds) is controversial (Davies & Michaelian, 2016), but those who accept their existence would probably consider them capable of learning.

"Systems" Learn

To conclude this section, it is a mistake to define learning so that only humans and other animals are included. We can see learning in single-celled organisms, artificial intelligences, immune systems, and plants. Nor can we define learning as something only "agents" do. Distributed cognitive systems learn, but these are, perhaps, not best described as "agents" or "organisms."

As such, I suggest the term "system," meaning a complex of elements that engage in information processing in pursuit of goals or preferences, be they explicit (as in a person's desire to be not hungry) or implicit.

Information Processing

In the proposed definition, the changes to the system need to be changes to the *representation or processing of information*. For purposes of this definition, information is defined as anything that has a *representational* use in a system—be it symbolic or subsymbolic. That is, the information stands for something else, be it a physical referent in the world, a utility to the system, an internal category, or anything else.

To make this clearer, I will describe systems that are *not* information processors. A memory-foam mattress changes in response to your lying down on it. It does so for purposes of your comfort—so it even has a function. It was designed to adapt to the environment, just as machine learning programs were designed to adapt to theirs. The mattress even has the word "memory" in its label.

According to my definition, the mattress is a poor example, because the change is not informational. The change in topography of the mattress does not *mean* anything to anybody in its normal use (if you came home and found that nobody was on your mattress but there was a deformity, you might use that deformation to conclude someone had been on it recently. In this case, the mattress deformation becomes a representation (to you), and is arguably a part of some distributed processing information system including you and the mattress.)

Similarly, a knife is not learning when you sharpen it, and your muscles are not learning when they get stronger because of a workout.

But all of these cases are *merely* physical changes, and in learning, these physical changes are important only because they encode changes to information storage and processing. Changes in knife sharpness and muscle tone are functional changes, but not of information processing systems. Instances of biological plasticity that do not involve information processing (like the growth of a callus) are not considered learning.

The Purpose of More Effective Action

The intuitive notion of learning is that when the system learns something, it is somehow improved. It either knows something it didn't before, or is able to do something it couldn't before, or can do it better.

This characteristic poses some immediate problems, because not everything people learn is good for them. If people tell you something that isn't true, and you believe them, then you have learned something false. And even though some false beliefs might help us, we can assume that, in general, false beliefs lead to poorer behavior (mental or physical) in the future.

Some learned *behaviors* are bad for us. In the case of post-traumatic stress disorder (PTSD), we learn behaviors that are problematic in non-traumatic situations (such as diving beneath the table whenever a helicopter flies by, or having nightmares that plague one for years; see Levin, 2000). I'll refer to learning false things, and the learning of maladaptive behaviors as "bad learning."

For the definition to be able to include bad learning, it is insufficient to say that learning must *always* leads to better behavior. However, we can avoid the problem by saying that it learns with the *purpose* of better future behavior.

I will explain with an analogy to digestion. We might describe the purpose of digestion as altering large, insoluble food molecules into smaller molecules that can be used as nutrition. The fact that we can digest poisons and non-nutritional food does not mean that the *function* of digestion isn't to nourish the organism. A system can be used poorly without removing its function. For the same reason, just because we can engage in bad learning does not mean that the function of learning isn't to promote better future behavior, nor that those bad things aren't learned.

Similarly, we remember lots of true but trivial facts that we might not productively use (or, indeed, even retrieve) ever again. In these cases, too, these declarative memories are not being used for better future action. But they are encoded because they *might* be useful someday. The mind remembers things without the certainty of what, exactly, will and won't be useful in the future. Will it be important to remember that Jill was wearing a red sweater? Probably not, but if we need to describe her to someone else, that fact might turn out to be useful.

We can see how memory is biased in terms of what it *expects* will be useful, however. For example, people tend to better remember things likely to be relevant to future events. The existence of a push pin will be better recalled if it is on the floor, where it might be stepped on, than if it is safely in a box (Zwaan, Van den Broek, Truitt & Sundermeier, 1996). Words related to survival are better remembered than other words (Nairne, 2010).

When an organism gets hit in the head, and suffers some deficit as a result, we would not want to consider this learning. Although brain damage affects the information processing of a system, the purpose of being hit in the head (if there even is one) is not to promote better future behavior, so the definition excludes this.

Another challenging example is the deliberate, direct physical change to a brain. When a doctor performs neurosurgery, or prescribes psychoactive medication, the purpose is better future behavior. If, in the future, we are able to "download" skills directly into our heads, as is done in *Matrix* films, should this be considered learning? In this account it is a bad example of learning, because the system is not changing itself. However, if, somehow, somebody managed to brain surgery on oneself, then my account would have to accept that as learning, strange as it sounds.

We sometimes deliberately alleviate mental tiredness by taking a rest, drinking coffee, or eating something. These activities have the purpose (among others, perhaps) of better future behavior. And some of these examples are the agent changing itself. Although rest and consuming coffee and food might be best described at a biological level, rather than at an information processing level, it is likely that there is an information processing level of description of how these activities promote better behavior. My definition includes these activities as decent examples of learning. The only characteristic missing is "experience," because the psychological experience of doing brain surgery on oneself or drinking coffee is not what causes the change (beyond placebo effects).

This raises the question of what counts as "experience." A body can experience hair loss at a barber, arguably, but what we want to capture here should not include experiences irrelevant to a cognitive system. I suggest that we ignore consciousness and say that an experience is limited to what the sensory apparatus of the system can detect. For an immune cell, it has receptors for detecting pathogens. A committee has analogues to sensory apparatus in the sense organs of the people that make it up.

Should the system be required to change itself for it to be considered learning, or are outside forces acting on a system acceptable? I will deal with issues regarding this question next.

Cultural and Evolutionary Learning

Some might want to describe learning at the sociological level. For example, in Fiji there is a cultural taboo: pregnant and lactating women may not eat certain kinds of fish. It turns out that avoiding consumption of these fish reduces a woman's chances of being getting fish poisoning by 30% during pregnancy and 60% during breastfeeding (Henrich & Henrich, 2010).

It is common for cultural taboos to have practical value that the people in the culture are not aware of. Often these are framed in terms of religion (for an example, see Harris, 1978). These taboos are refined over the course of generations. No single individual need engage in learning for this to happen, though individuals encode the information state of the cultural system. If we look at culture as an evolving entity, and, in particular, the ideas in the culture as undergoing evolutionary selection, we can see how ideas that facilitate reproduction will have a better chance of enduring over the years than others (Richerson & Boyd, 2008).

What we observe, then, is that the society itself is doing the learning. The society, in this respect, is a cognitive system that is distributed over time, and we can observe the information changes it makes to act better in the future.

One might also look at a species as a system that learns through Darwinian evolution. Sweller and Sweller (2006) suggest that this happens, analogically mapping long-term memory with a genome; learning from other humans with biological reproduction; problem solving with

random mutation, and so on. Although I have not found an analogous argument for culture, it seems that one could easily be made.

But is evolutionary change "for the purpose of better future behavior?" We often can take a design stance to evolutionary processes to help us understand them, but biologists take great pains to make it clear that evolution is not goal-directed. Darwinian evolution is not purposeful (unless it is artificial selection, or is designed by a programmer in a simulation).

Specific behavioral phenotypes can be described as having purposes. As Daniel Dennett describes it, cuckoo chicks push other birds out of the nest. As scientists, we can ask why, and get a description at the level of neurons, but it is also profitable to look at the function of this behavior: to maximize resource acquisition from the cuckolded parent bird (1987). The function is a "free floating rationale." But application of this to the evolutionary process itself is more problematic. The products of evolution might be purposeful, even if evolution itself is not.

My point here is not take a strong stance on whether or not the changes to cultures and species that we see over time should count as learning, but to discuss how different definitions of learning would or would not include them. The definition I'm suggesting in this paper would render these poorer examples, because the changes are not for the purpose of better future behavior in genetic nor in cultural evolution, the changes are (arguably) not occurring through experience (can a culture or species experience something?) and finally because the system is not changing itself (this is clear for the genome, and possibly true for a culture). We still might metaphorically describe them as learning, and doing so might help us understand or teach these concepts.

Limitations of the Analysis

"Learning" happens to be a word in English, the *lingua franca* of science. However, we need to be careful not to assume that the existence of a word means that it necessarily refers to a natural kind. Other languages might break up the world in different ways, and ultimately whether learning exists in a way that happens to be captured by the English word for it is an empirical question.

This paper is in the tradition of a classical-styled conceptual analysis, looking for and suggesting conditions for what would count as an instance of "learning," and this is, admittedly, old-fashioned.

Is there a better way to do it? An earthquake can be described and explained using theories and equations from geology, but it turns out that these same theories apply to quakes that happen elsewhere as well—moons, starts, other planets, etc. Thus it makes sense to suggest that the idea of a "quake" extends beyond those that happen on Earth (United States Geological Survey, 2012).

This makes sense because we have a theory that is broadly, and successfully, applied. Admittedly, this is not happening with learning. Perhaps future descriptions of learning will be more theory-based. That is, we come up with a theory of learning (or a particular kind of learning), and then see to which phenomena in the world the theory can be productively applied. These future investigations might mean that "learning," as we conceive of it in English, isn't a sensible scientific category at all (Churchland, 1989, suggests that no sensible scientific categories should be based on folk psychology).

However, there is no general theory of learning yet, and if we think of cognitive science as the study of cognition independent of the substrate that supports it, it is helpful to have *some* idea of what we mean by learning. This paper is intended to be a start to the discussion, and more of a stepping-stone for future refinement rather than the final answer.

Conclusion

We've known for a long time that the search for necessary and sufficient conditions for concepts is often a fruitless task, so the definition should be seen as a list of family-resemblance features. My suggested definition is that an event is a better fit for the category "learning" depending on the degree to which the characteristics in the following list describe it:

1. The change happens to an information processing system

2. The change happens with the purpose of better future action
3. The change happens in response to the system's experience
4. The change is executed by the system itself, rather than some outside influence

This definition covers the intuitive and prototypical instances of learning, but renders as poor examples some processes that we might want to productively talk about as learning, such as evolutionary processes over species and cultures.

With hope, future research will ground the definition of learning in a theory of learning process, in contrast to my attempt to define it from a conceptual analysis.

Acknowledgments

Thank you for the helpful comments of the attendees of a brownbag talk I gave on this topic at Carleton University in 2016.

References

Baddeley, A. (1992). Working memory. *Science*, *255* (5044), 556.

Boisseau, R. P., Vogel, D., & Dussutour, A. (2016, April). Habituation in non-neural organisms: evidence from slime moulds. In *Proceedings of the Royal Society B*, 283(1829), 20160446).

Chamovitz, D. (2012). *What a plant knows: A field guide to the senses*. Scientific American: New York. Pages 50--63.

Chauvin, Y., & Rumelhart, D.E. (1995). *Backpropagation: theory, architectures, and applications*. Psychology Press.

Churchland, P. M. (1989). *A neurocomputational perspective: The nature of mind and the structure of science*. MIT press.

Cooke, S. F., & Bliss, T. V. P. (2006). Plasticity in the human central nervous system. *Brain*, *129*(7), 1659-1673.

Davies, J. & Michaelian, K. (2016). Identifying and individuating cognitive systems: A task-based distributed cognition alternative to agent-based extended cognition. *Cognitive Processing. 17(3)*, 307—319.

Dennett, D.C. (1987) *The Intentional Stance*, Cambridge, MA: MIT Press.

Farmer, J. D., Packard, N. H., & Perelson, A. S. (1986). The immune system, adaptation, and machine learning. *Physica D: Nonlinear Phenomena*, *22*(1), 187-204.

Harris, M. (1978). India's sacred cow. *Human Nature*, *1*(2), 28-36.

Henrich, J. & Henrich, N. (2010). The evolution of cultural adaptations: Fijian food taboos protect against dangerous marine toxins. *Proceedings of the Royal Society B: Biological Sciences, 277*(1701), 3715-3724.

Ishikawa, D., Matsumoto, N., Sakaguchi, T., Matsuki, N. & Ikegaya, Y. (2014). Operant conditioning of synaptic and spiking activity patterns in single hippocampal neurons. *The Journal of Neuroscience*. 34(14), 5044—5053.

Levin, R. (2000). Nightmares: Friend or foe? *Behavioral and brain sciences*, *23*(06), 965.

Nairne, J. S. (2010). Adaptive memory: Evolutionary constraints on remembering. *Psychology of Learning and Motivation*, *53*, 1-32.

Richerson, P. J. and Boyd, R. (2008). *Not by genes alone: How culture transformed human evolution.* Chicago, IL: University of Chicago Press.

Sandell, M. A., & Breslin, P. A. (2006). Variability in a taste-receptor gene determines whether we taste toxins in food. *Current Biology*, *16*(18), R792.

Sweller, J., & Sweller, S. (2006). Natural information processing systems. *Evolutionary Psychology*, *4*(1), 434—458.

United States Geological Survey, (2012). Earthquake hazards program. Retrieved 5 April from http://earthquake.usgs.gov/learn

Wilson, R.A., & Keil, F.C. (2001). *The MIT encyclopedia of the cognitive sciences*. MIT press.

Zwaan, R.A., Van den Broek, P., Truitt, T.P., & Sundermeier, B. (1996). Causal coherence and the accessibility of object locations in narrative comprehension. *Abstracts of the Psychonomic Society*, *1*, 50.