Metascience
DOI 10.1007/s11016-013-9746-x

ESSAY REVIEW

The cognitive science of Feynmen

Paul Thagard: The cognitive science of science: Explanation, discovery, and conceptual change. Cambridge, MA: The MIT Press, 2012, 376pp, $40.00, £27.95 HB

Sanjay Chandrasekharan

There is a Feynman problem-solving algorithm, coined in jest by Murray Gellman. It goes like this:

1. Write down the problem
2. Think hard
3. Write down the answer

One popular approach in the cognitive studies of science seeks to understand step 2, by identifying significant features of the algorithm's output (the answer), and trying to figure out what kind of mental process could have generated an answer with these features. *The Cognitive Science of Science* by Paul Thagard reports many studies in this vein, though the volume covers a lot more novel ground, by examining the nature of explanation and conceptual change in many areas, and presenting models of the role of emotion in science. All these aspects are brought together in a single volume, and the domains of science discussed include resistance to climate change, epigenetic explanation of mental illness, and the connection between western and Chinese medicine. This breadth is commendable, and I learned quite a bit of science and history of science from this volume, but at times this wide scope gave the impression of reading a collection of disparate papers, rather than a single volume.

A large chunk of the book reports studies and computational models of scientific reasoning and explanation developed by the author and his colleagues. Most of these studies follow the structural/taxonomic approach mentioned above, where features of discoveries, explanations, and cases of conceptual change are examined, and salient features are extracted to develop theoretical accounts/explanatory schema and neural network models. A good illustrative example is scientific

S. Chandrasekharan (✉)
Homi Bhabha Centre for Science Education, Tata Institute of Fundamental Research, Mankhurd, Mumbai 400088, India
e-mail: sanjay@hbcse.tifr.res.in

Published online: 06 February 2013

discovery, which is considered one case of creativity. The author proposes a theory of creativity, which is captured by five theses:

1. Creativity results from novel combinations of representations.
2. In humans, mental representations are patterns of neural activity.
3. Neural representations are multimodal, encompassing information that can be visual, auditory, tactile, olfactory, gustatory, kinesthetic, and emotional, as well as verbal.
4. Neural representations are combined by convolution, a kind of twisting together of existing representations.
5. The causes of creative activity reside not just in psychological and neural mechanisms, but also in social and molecular mechanisms.

To support this schema, a proof-of-concept neural network model of creativity is presented, which combines visual patterns in what could be considered creative ways, using the mathematical mechanism of convolution. In a subsequent chapter, the author argues that this structural/taxonomic approach is non-trivial and productive, as there is opposition to it (therefore non-trivial) and the categories it proposes allow deeper examination in terms of the neural mechanisms underlying them (therefore productive). These claims are valid, and the intellectual work that has gone into developing these accounts provides significant clarity on the issues related to the complex problem of scientific cognition.

This is a good beginning for the nascent science of scientific thinking, but further progress requires addressing a set of issues, which I will motivate using two quotes by Alan Kay from the chapter on creativity in computer science in the book. Kay is discussing how his exposure to programs such as Sketchpad and languages such as Simula culminated in his development of Smalltalk and the novel object-oriented programming paradigm:

> The sequence moves from 'barely seeing' a pattern several times, then noting it but not perceiving the 'cosmic' significance, then using it operationally in several areas; then comes a 'grand rotation' in which the pattern becomes the center of a new way of thinking (165).
> The big flash was to see this as biological cells, I'm not sure where that flash came from but it didn't happen when I looked at Sketchpad. Simula didn't send messages either (165).

The authors of the chapter frame these comments in terms of analogical 'jumps', considering Sketchpad and Simula as local analogies, and biological cells as a distant analogy, that led to the development of Smalltalk. This structural approach, where the progression of Kay's thinking is captured using analogical structure, is one way to understand the process Kay is describing.

Another way to think about it is to consider the lengthy *process* and the series of *interactions* that Kay describes. In this view, there is a phase of interaction where the pattern is barely seen; a phase of interaction when the pattern is seen, but this seeing does not lead to understanding the pattern's significance; even using the pattern in specific instances does not automatically lead to the understanding of its significance; until one day, there is a 'grand rotation', which is based on a 'big flash'

that the pattern is similar to biological cells; the inventor does not know where the flash came from, but he is sure that it was not in his first encounters with Sketchpad and Simula, which suggests the origin of the flash is in the long series of (partially revealing) interactions described.

The account of discovery developed in the book ignores this series of interactions and the partial insights from them, which lead up to the big flash event. It focuses exclusively on the 'grand rotation' process. This phase gets most attention in popular accounts of discovery, so it is a good place to start developing a theory of discovery. However, an account that focuses exclusively on this phase does not address the following questions:

1. Why did the invention process take so long? Presumably, the features of Sketchpad and Simula, as well as the notion of biological cells, were in Kay's brain all along. Why couldn't the convolution process just take these elements and create the invention? Related to this, Sketchpad and Simula were used by many programmers in Kay's generation. What process allowed only Kay to execute the convolution(s) required to come up with Smalltalk?
2. Why is the process of analogy so difficult? In science education, a central problem faced by students is transfer (Bransford and Schwartz 1999), which involves seeing a given problem (say an oscillating circuit) as an instance of a learned pattern (simple harmonic motion). This is an instance of a local analogy, where the search space for patterns is very limited. However, even with these constraints, such analogical problems are very difficult for students. In the case of Kay, and also other cases discussed in the book, there is more than one analogy, and the way they come together is very complex. What process is used by inventors and scientists to find/create these complex multi-step analogies, given the difficulty in finding/creating even one?
3. What is the role played by the series of interactions and partial insights? Are they required, or are they unnecessary sidelights? If they are required, how do they contribute to the convolution process?
4. Why is the multi-modality of representations important? What role does this feature play in discovery and also convolution?
5. How would the convolution model explain discoveries made by collaborating teams, which is the norm in recent science?

Similar questions can be raised about the coherence account of explanation, which is also a structural account that focuses on the features of explanation, and not on the process by which an explanation is discovered/understood.

In my view, answering these questions systematically (rather than in an ad hoc fashion) requires taking seriously two fundamental biological principles emphasized by radical embodied cognition theory (which the author critiques in two places in the book) and its close cousins (situated cognition, ecological psychology, and dynamic systems theory), even when rejecting (as the author rightly does) their extreme claim that representations do not exist (Chandrasekharan and Stewart 2007; Chandrasekharan and Osbeck 2010). The two principles are:

1. The brain evolved to support action
2. The brain evolved for, and through, interaction with the environment

The first principle suggests that the central feature of neural representations is not their multi-modality, but their action-orientedness. This does not mean that all internal representations are motor representations. It only means that internal representations have a bias, such that they are manipulated, and changed, easier through actions and interaction with the environment.

The second principle suggests that the brain assumes the world, and therefore it can use the world as a resource. This does not mean all cognition requires interaction with the world, as some have claimed (Brooks 1991). It just means that the brain will access resources from the world, and offload processing (including imagination, Kirsh and Maglio 1994) to the world, if these are available options.

Essentially, from this viewpoint, while the author stresses that neural networks are processes, he ignores the fact that this is not an accident—the network process evolved to support two other critical processes, actions, and constant interaction with the environment. To incorporate these biological principles into an account of scientific cognition, it is not enough to add some environmental interaction to the neural network that currently implements features of the discovery. It requires re-conceptualizing the Feynman problem-solving algorithm, in the following non-Feynman way:

1. Write down the problem
2. Make structures in the world that correspond to the problem
3. Think hard using these structures
4. Revise 2; Revise 3 (many times); Revise 1 if needed
5. Write partial results
6. Collect partial results; Revise 1
7. Run 1–6 many times
8. Write down the answer

In this restructuring, which is closer to how many scientists make discoveries, scientific discovery is the result of an iterative and interactive process. Importantly, since step 1 is changed many times, a theoretical account of this process requires focusing on individual discoveries, rather than on general features of all discoveries. A generalization about the *process* is only possible after many such cases are studied in process terms.

The feature-based approach favored by the author has a deeper problem as well. Imagine that someone writes an original software program. Let us say it takes videos of a given person's movements as input and gradually turns her movements to that of another person. Now, to understand how the programmer created this program, we could look at the features of the final program code and extract patterns. One pattern could be that the program is a combination of many 'for' loops and 'while' loops. This feature is part of almost all programs (just as the combination of concepts is part of all discoveries). Given this generality of the two loops, we could try to see how they are represented in the brain, and what neural operation(s) would allow them to be combined. This approach is analogous to the

approach presented by the author, and it is both non-trivial and productive, as the author claims.

But would this approach tell us anything about the nature and process of invention and its cognitive bases, even in this particular case? Very unlikely, because the invention becomes an invention not through the features of the final program, but through its technological context (does this program already exist?), social context (do people want this?), and the programmer's interaction with the world—to understand people's gaits, to develop ways of capturing and transforming gaits algorithmically, and to test the invention using actual people. All this happens in the real world, and none of this context and process is encoded in the final program, particularly not in a fashion extractable just by studying the final program. The account based on general features of the final output is thus not investigating the actual process of discovery. It is providing an idealized account, of how features of a new concept could potentially emerge from pre-existing concepts in the brain. The discoveries provide only the starting point for this investigation, as they provide good instances of combining concepts. The investigation is thus about combining of concepts, not about discovery. Metaphorically, an investigation of the neural basis of the color red is not an investigation of apples or how apples come to be, even though red is a prominent feature of most apples.

The coherence model of scientific explanation is based on a similar structural approach, so these problems are applicable to this account as well. The account of conceptual change, based on explanatory schemas, also provides a structural, rather than process, account. It is unclear whether the structural model could be revised to fix these issues. It is possible that a separate process model needs to be developed, taking into account the role of interaction in discovery (as well as explanation and conceptual change), and the neural processes that support such interaction. I outline a preliminary account of discovery along these lines in Chandrasekharan (2009); (see also Chandrasekharan and Nersessian 2011; Nersessian 2008).

The problem of scientific cognition is very complex, and the cognitive science of science is in its early infancy. This book makes a significant contribution to clarifying the issues involved in studying scientific cognition, and outlining proof-of-concept models of possible neural processes involved in discovery, explanation, and conceptual change. However, these models are preliminary and do not account for the way extended interactions (with artifacts, people and the real world) and iterations change scientific cognition. Such process accounts usually come at a second stage in any science, and build on insights from preliminary structural models. I look forward to Cognitive Science of Science 2.0.

References

Brooks, R. 1991. Intelligence without representation. *Artificial Intelligence* 47: 139–159.

Bransford, J.D., and D.L. Schwartz. 1999. Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education* 24(3): 61–100.

Chandrasekharan, S., and T.C. Stewart. 2007. The origin of epistemic structures and proto-representations. *Adaptive Behavior* 15(3): 329–353.

Chandrasekharan, S. 2009. Building to discover: a common coding model. *Cognitive Science* 33(6): 1059–1086.

Chandrasekharan, S., and L. Osbeck. 2010. Rethinking situatedness: Environment structure in the time of the common code. *Theory & Psychology* 20(2): 171–207.

Chandrasekharan, S., and N.J. Nersessian. 2011. Building cognition: The construction of external representations for discovery. In *Proceedings of the Cognitive Science Society conference 2011*, ed. L. Carlson, C. Hölscher, and T. Shipley, 267–272. Austin, TX: Cognitive Science Society.

Kirsh, D., and P. Maglio. 1994. On distinguishing epistemic from pragmatic action. *Cognitive Science* 18: 513–549.

Nersessian, N. 2008. *Creating scientific concepts*. Cambridge, MA: MIT Press.