The Role of Agency in Distributed Cognitive Systems

Ronald N. Giere†

In previous publications I have argued that much scientific activity should be thought of as involving the operation of distributed cognitive systems. Since these contributions to the cognitive study of science appear in venues not necessarily frequented by philosophers of science, I begin with a brief introduction to the notion of a distributed cognitive system. I then describe what I take to be an exemplary case of a scientific distributed cognitive system, the Hubble Space Telescope (HST). I do not here reargue the case for conceiving of systems like the HST as distributed cognitive systems. Rather, I examine a question that arises once one has adopted the perspective of distributed cognitive systems, namely, the role of *agency* in a distributed cognitive system. Here I argue, contrary to several advocates of distributed cognitive systems, that we should regard the human components of distributed cognitive systems as the only sources of agency within such systems. In particular, we should *not* extend notions of agency to such systems *as a whole*.

1. Introduction. In previous publications (Giere 2002a, 2002b) I have argued that much scientific activity should be thought of as involving the operation of distributed cognitive systems. Since these contributions to the cognitive study of science appear in venues not necessarily frequented by philosophers of science, I will begin with a brief introduction to the notion of a distributed cognitive system. I will then describe what I take to be an exemplary case of a scientific distributed cognitive system, the Hubble Space Telescope (HST). I will not here reargue the case for conceiving of systems like the HST as distributed cognitive systems. Rather, I will be examining a question that arises once one has adopted the perspective of distributed cognitive systems, namely, the role of *agency* in a distributed cognitive system. Here I will be arguing, contrary to several advocates of distributed cognitive systems, that we should regard the human components of distributed cognitive systems as the only sources

†To contact the author, please write to: Center for Philosophy of Science, 831 Heller Hall, University of Minnesota, Minneapolis, MN 55455; e-mail: giere@umn.edu.

Philosophy of Science, 73 (December 2006) pp. 710–719. 0031-8248/2006/7305-0022$10.00
Copyright 2006 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/518772 Published online by Cambridge University Press

of agency within such systems. In particular, we should *not* extend notions of agency to such systems *as a whole*.

2. A Distributed Cognitive System. A canonical source for the concept of distributed cognition within the cognitive sciences is Ed Hutchins's study of navigation in his 1995 book, *Cognition in the Wild*. This is an ethnographic study of traditional "pilotage," that is, navigation near land, as when coming into port. Hutchins demonstrates that individual humans may be merely components in a complex cognitive system. No one human could physically do all the things that must be done to fulfill the cognitive task, in this case, repeatedly determining the relative location of a traditional navy ship as it nears port. There are, for example, sailors on each side of the ship who telescopically record angular locations of landmarks relative to the ship's gyrocompass. These readings are then passed on by the ship's telephone to the pilothouse where they are combined by the navigator on a specially designed chart to plot the location of the ship. In this system, no one person could possibly perform all these tasks in the required time interval. And only the navigator, and perhaps his assistant, knows the outcome of the task until it is communicated to others in the pilothouse. Those recording the locations of landmarks have no reason to know the result of the process.

One might wish to treat Hutchins's case merely as an example of *collective cognition*. The cognitive task—determining the location of the ship—is performed by a collective, an organized group, and moreover could not, in the given circumstances, physically be carried out by a single individual. In this sense, collective cognition is ubiquitous in modern societies. In many workplaces there are some tasks that are clearly cognitive and, in the circumstances, could not be carried out by a single individual acting alone. Completing the task requires coordinated action by several different people. So Hutchins is inviting us to think differently about many common situations. Rather than simply assuming that all cognition is restricted to individuals, we are invited to think of some actual cognition as being distributed among several individuals.

Hutchins's conception of distributed cognition, however, goes beyond collective cognition. He includes not only persons but also instruments and other artifacts as parts of the cognitive system. Thus, among the components of the cognitive system determining the ship's position are the "alidade," an adaptation of a surveyor's telescope, used to observe the bearings of landmarks, and the navigational chart on which bearings are drawn with a ruler-like device called a "hoey." The ship's position is determined by the intersection of two lines drawn using bearings from the two sightings on opposite sides of the ship. So parts of the cognitive process take place not in anyone's head, but in an instrument or on a

https://doi.org/10.1086/518772 Published online by Cambridge University Press

chart. The cognitive process is distributed among humans and material artifacts.

3. The Hubble Space Telescope. I suggest that Hutchins's navy ship be taken as a prototype not only for large scientific facilities such as the HST and CERN, but also for many smaller facilities such as laboratories for gene sequencing or for imaging human brains. Here I will focus on the HST.

The Hubble Space Telescope was launched on April 24, 1990, aboard the space shuttle *Discovery*. Since a dramatic mission to correct an embarrassing flaw in its mirror at the end of 1993, Hubble has produced genuinely revolutionary observations. In January of 2003, the Space Telescope Science Institute (STSI) released a remarkable image produced by the Advanced Camera for Surveys (ACS) aboard the Hubble. Scientists who have studied this image claim that it captures galaxies from which light was emitted roughly 13 billion years ago, when the universe was only 1 billion years old, which is to say, about one-fourteenth its present age.

The process that produced this image is far too complex for me even to attempt to describe in any detail here. In addition to the mirror, it involves electronic detectors sensitive to light in the infrared part of the electromagnetic spectrum. The output of the detectors is fed into an onboard computer and put into a form in which it can be transmitted to a tracking and data relay satellite, from which it is retransmitted to the White Sands complex near Las Cruces, New Mexico, from which it is again retransmitted by domestic satellite to the data operations control center at the Goddard Space Flight Center in Greenbelt, Maryland. From there it is routed to the data capture facility and finally on to the Space Telescope Science Institute in Baltimore, where it is studied by astronomers and other space scientists. Each step in this process in some way modifies the initial signal and contributes to the construction of the final image.

A remarkable feature of this particular image is that it involved gravitational lensing. During the exposure, the Hubble telescope was pointed directly at a massive cluster of galaxies, known as Abell 1689, estimated to be 2.2 billion light-years away. In accordance with the general theory of relativity, this mass acts like a lens by warping space around it and thus effectively bending light passing by.

Although it incorporates many more, and much more complex, instruments, as well as many more people, I think it is theoretically useful to think of the HST as a distributed cognitive system along the lines of Hutchins's navy ship. It is a system involving both humans and artifacts that produces cognitive outputs, such as authenticated claims about the

https://doi.org/10.1086/518772 Published online by Cambridge University Press

age of the universe. Such claims could not be produced by any one person or group of people without the complex instrumentation employed.

The main alternative to thinking of the HST as a full-fledged distributed cognitive system is to think of the instrumentation simply as producing evidence (inputs) for the humans who produce knowledge claims (outputs). Here all the cognition is localized in human brains (or minds). At most one has collective cognition, not full-fledged distributed cognition. My main reason for preferring to invoke distributed rather than individual cognition is that it provides a conception of cognition in science that unifies much thinking about science in the history, philosophy, and sociology of science within a general cognitive science framework. However, I will not here reargue the case for invoking distributed rather than individual or even collective cognition to understand scientific practices like that involving the HST. Rather, I wish to inquire how further to characterize the nature of the cognition within a distributed cognitive system.

4. Hybrid Systems. The still dominant perspective within the cognitive sciences is that cognition is computation. Within this perspective, then, as a distributed cognitive system, the HST is a computational system. There are, of course, weaker and stronger versions of computationalism. Hutchins himself invokes a weaker version to claim that the navigational system on his navy ship is a computational system even though the major instruments used are analog rather than digital. Nevertheless, the navigational system aboard the ship does utilize representational schemes, such as a nautical chart, and there are rules for manipulating these representations so as to achieve the cognitive goal, determining the location of the ship. This pretty well fits the strict definition of computation as the manipulation of representations according to explicit rules. So characterizing Hutchins's system from a computational perspective may be justified.

It is obvious that the HST incorporates a vast computational system. From the sensors that record light focused by the Hubble mirror all the way to the final image produced using advanced computer graphics, the system is computational in the strict sense of being fully digital. The representations are digitized and these representations are manipulated by an array of computers according to explicit rules.

The same cannot be said for the HST system *as a whole*. The "front end" of the system, that is, everything up to but not including the sensors, contains no representations at all, much less rules for the manipulation of representations. The most important part of this component of the system is the mirror. The mirror does not have a built-in representational structure. It is simply an optical instrument. Scientists, of course, have representations *of* the mirror which they used to design it and to under-

https://doi.org/10.1086/518772 Published online by Cambridge University Press

stand how it functions. But the mirror itself does not represent anything and is not computing anything. It is merely interacting with incoming light in a causal manner. Once the light emerges from the mirror system, it enters the sensors, which are designed to be representational. What they do with the light *is* governed by explicitly designed rules.

What about the "back end" of the system, so to speak, where scientists at the Space Telescope Science Institute study the images produced by the rest of the system and reach various conclusions about what has been detected? Here again, a computational perspective would have us think that, insofar as it produces cognitive outputs, this too is a computational system. Here I side with those who find this wholesale extension of the computational perspective not very illuminating. The cognitive output is produced by a complex interaction of numerous people and involves a variety of individual and collective judgments. No one has any plausible scheme for rendering this process as a rule-governed manipulation of representations. Far better, I think, to understand this process as the operation of a social and cultural system using concepts from various human sciences, including philosophy and the philosophy of science.

In fact, similar considerations apply even to the apparently computational "middle" part of the system. All of the operations here are programmed in standard ways. Nevertheless, considerable human judgment is involved in deciding what algorithms to apply at each of the many stages through which the signal travels from the sensors aboard the telescope itself all the way to production of the final images. Anyone familiar with computer graphics knows how much flexibility there is in producing images using these techniques.

I conclude that distributed cognitive systems like the HST are *hybrid* systems, part just physical-causal, part computational, and part social-cultural. I realize, of course, that to philosophers of science, who tend to be physicalists of some stripe, these distinctions can only be relative. At bottom, it is usually thought, these are all just complex physical systems. Even if one advocates physicalism in a very deflationary form, for example, as a methodological presumption rather than as a metaphysical doctrine, the presumption applies to distributed cognitive systems as a whole, not just to one part.

What makes the computers in the middle of the system computational in spite of being just physical systems is that they are designed to instantiate the syntactic forms of language and mathematics as well as rules for the manipulation of these forms. The mirror was not so designed. It was designed to focus light in a certain way. The temptation to think of the mirror as embodying the mathematical structure of theories of physical optics may be lessened by the realization that such structures are abstract models which at best provide only an approximate simulation of what in

https://doi.org/10.1086/518772 Published online by Cambridge University Press

fact goes on in the mirror. No one ever has, and presumably no one ever will, produce a literally exact simulation of any physical process. Physical theories are *our* representations of nature, not nature's own representations of itself.

At the other end, even presuming that humans are ultimately just complex physical systems, it is undeniable that humans are capable of self-consciously creating and operating representational systems. I have already stated my reservations about regarding humans as computational systems. This is no place even to attempt to justify these reservations. Instead, I would like to consider a related issue, the role or location of *agency* in a distributed cognitive system.[1]

5. Agency in Distributed Cognitive Systems. The fundamental idea in distributed cognition is that a system *as a whole* performs a cognitive task. The cognition is distributed throughout the whole system, including both humans and artifacts. Advocates of distributed cognition, such as Hutchins, a cognitive anthropologist, and Andy Clark (1997), a philosopher of cognitive science, have explicitly argued that other attributes besides cognition, including even having a mind, should likewise be applied to distributed cognitive systems as a whole. Prominent sociologists and anthropologists of science (Latour 1986; Knorr-Cetina 1999) have advocated similar views, though not from the perspective of cognitive science.[2] This turns out to be a major sticking point for philosophers, philosophers of science, and even philosophers of cognitive science.[3] Do distributed cognitive systems *as a whole* think, acquire knowledge, or have minds? These notions are all bound up with the notion of cognitive *agency*. Are distributed cognitive systems *as a whole* to be regarded as *cognitive agents*? The only plausible alternative view, I think, is to regard cognitive agency as restricted to the *human* components of a distributed cognitive system. There are difficulties with both of these alternatives, but, on balance, I will argue, the more restricted view is preferable.

The fundamental problem with extending cognitive agency to the whole system is that the concept of cognitive agency is tightly bound up with related concepts such as that of intention, responsibility, consciousness, and, in general, having a mind in what is (sometimes derisively) called the "folk psychological" sense of "mind." Consider, for example, a simple distributed cognitive system consisting of a single person working at a

1. See also Giere 2003 and Giere 2004.

2. I have commented on these two works in Giere 2002c and in Giere and Moffatt 2003, respectively.

3. For a detailed critique of distributed cognition and the idea of extended minds from the standpoint of analytical philosophy of mind, see Rupert 2004.

https://doi.org/10.1086/518772 Published online by Cambridge University Press

desktop computer. This system, as a whole, has the capacity rapidly to perform cognitive tasks, such as multiplying two 10-digit numbers, that could not be performed nearly so quickly, if at all, by a human acting alone. There seems little that is problematic about ascribing the successful performance of a cognitive task to the combined human-computer system. But there are already problems if we move to saying that the system as a whole "remembers" the answer as opposed to saying that the system as a whole has the capacity for retrieving the answer. Could the system as a whole "forget" the answer, as opposed to the answer being accidentally deleted by the person pushing the wrong key or lost due to a computer malfunction before being saved to a more permanent medium such as a hard drive? Could the system as a whole "regret" the loss of the answer? Could it be blamed for the mistake? Should it feel sorry that the answer was lost? Of course people "blame" their computers for all kinds of mishaps, but this can be seen as just a harmless anthropomorphizing of an artifact, or maybe an attempt to escape responsibility on the part of the human operator, but not a serious ascription of blame. More serious questions arise if we ask whether the system *as a whole* intended to multiply the two numbers, or, ultimately, whether the system as a whole has a mind of its own.

Such questions have been asked by philosophers and cognitive scientists for several generations, usually in the context of questions concerning whether computers can think, or exhibit consciousness. Part of the plausibility of positive answers, I think, is due to the simple fact that computers, especially personal computers, are relatively localized entities. Positive answers are much less plausible when we consider a much more extended artifact such as the HST. If we were to attribute standard features of human agents to this whole system, we would have to imagine a mind that extends from earth orbit to Greenbelt, Maryland. Worse yet, the particular system that produced the image of the ancient universe noted earlier incorporated a massive cluster of galaxies some 2.2 billion light-years away acting as a gravitational lens. So we would have to imagine a mind so extended that it takes 2.2 billion years for one part of the mind to communicate with another part, unless we are to suppose that mental states propagate much faster than the speed of light. Such ideas would be at best science fiction, not good science.

What, then, about the possibility of restricting cognitive agency to the humans in the system? The fundamental problem with this option, especially for philosophers and philosophers of science, such as Paul Churchland (1989, 1995), is that the bundle of concepts making up folk psychology looks like a primitive scientific theory, and as such seems clearly false, if not actually incoherent. In particular, the folk psychological theory seems to presume a notion of free agency that is incompatible with

https://doi.org/10.1086/518772 Published online by Cambridge University Press

even a liberal materialistic account of human organisms. Assuming that some such causal account of human action is true, folk psychology must be false. It is thus hardly a sound basis on which to build a viable account of scientific practice. This might be taken as an argument against adopting the notion of distributed cognition in the first place. Nevertheless, the benefits of having an account of scientific cognition that unifies cognitive, social, and historical aspects of scientific practice seem to me great enough to retain the notion of distributed cognition and thus to reject this argument.

Although, within philosophy, these issues are often discussed in the context of arguments for or against free will, I think philosophy of science itself has resources for bypassing much of this literature. The chief resource is the view of scientific theories comprised of families of models together with claims about the relative fit of models to the world.[4] On this view of theories, the question is not whether folk psychology, considered as a scientific theory, is true or false. Most if not all theoretical claims are, strictly speaking, false. The question is whether the various models generated by the principles of the theory fit actual situations of interest well enough for the intended purposes. On this measure, folk psychology does remarkably well. We are generally very good at assigning responsibility, and thus praise or blame, for peoples' actions, including their epistemic actions. Thus, in the operation of a complex distributed cognitive system such as the HST, we can often enough reliably determine who is or is not performing well at their assigned tasks, including the task of drawing theoretical conclusions from the final images, just as we can often enough reliably determine whether a piece of equipment is functioning correctly. That seems sufficient to distinguish the epistemic agency of humans from the causal agency of other parts of the system.

In terms of traditional discussions of free will, this treatment resembles what has been called an "error theory." We apply models that seem to include a notion of freely chosen action that is at odds with a scientific understanding of humans as also causal, though very complex, systems. So, ultimately, causal models would better fit the situation, although no comprehensive causal models currently exist. But to call this an error theory is misleading because that terminology presumes that a causal theory would be a true theory while the folk psychological theory is false. Understanding theories as incorporating sets of models, the situation is better described as one in which one set of well-known models can be applied very well for a wide range of human activities while a largely

4. This is the view of theories long advocated by van Fraassen (1980, 1989), Suppe (1989), Cartwright (1983, 1989, 1999), myself (Giere 1988, 1999, 2006), and, more recently, Paul Teller (2001, 2004).

https://doi.org/10.1086/518772 Published online by Cambridge University Press

unformulated set of models might ultimately provide a more detailed fit to human behavior for scientific purposes.

Of course, there is also a long tradition of "compatibilism," going back to Hume, according to which the ordinary notion of a free action is just an action that is not coerced. Compatibilism is even easier to understand in terms of models. Our ordinary notions of human action generate quite simple but easy to apply models, while the sciences of human behavior strive to generate much more elaborate causal models. This contrast between simple, easy to apply models and richer models that are more difficult to apply is commonplace in the sciences. A standard example from physics would be the contrast between the models of thermodynamics and those of statistical mechanics as applied to gasses.

Having rescued a viable notion of human agency, it is now possible to claim that it is only the humans in the system that come to know the results of observations and the conclusions drawn from them. They are justified in their claims in the standard sense that they know the capabilities of the whole system. They know just how reliable it is. But the system as a whole does not know anything. It is not an epistemic agent. Actually, as has long been assumed in the social study of science, although the scientists analyzing the data may individually and collectively claim to know the results, these results do not count as *scientific* knowledge until their claims pass through the process of peer review and publication. So the cognitive system that produced the *scientific* knowledge about the age of the universe is even more widely distributed than the HST system itself.

6. Conclusion. I conclude that thinking of systems like the HST as distributed cognitive systems does not commit one to ascribing to the whole system other attributes of ordinary human cognition such as knowing or having a mind. This implies that the term "cognitive" can be applied to the whole system without also applying such terms as "knows," "believes," etc. So I am suggesting that "cognitive" be used as a theoretical term of cognitive science, not a term of everyday life or of folk psychology. Theoretical sciences are free to choose their own technical terms as best suits their scientific needs.

In general, whether systems like the HST should be thought of as distributed cognitive systems and whether humans should be taken as the only source of agency in such systems are not straightforward empirical questions. Rather, an affirmative answer to these questions provides a theoretical framework for thinking about a range of scientific practices. The framework might then be used to generate more specific models of particular cognitive systems that provide a good empirical fit to the operations of those systems. I think this is the case for the HST. But judg-

https://doi.org/10.1086/518772 Published online by Cambridge University Press

ments of fit are relative to theoretical interests and goals. One of my goals is to have a theoretical framework that can simultaneously serve the interests of the history, philosophy, and sociology of science. The cognitive study of science, and particularly the concept of a distributed cognitive system, provides such a framework.

REFERENCES

Cartwright, Nancy D. (1983), *How the Laws of Physics Lie.* Oxford: Oxford University Press.

——— (1989), *Nature's Capacities and Their Measurement.* Oxford: Oxford University Press.

——— (1999), *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

Churchland, Paul M. (1989), *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. Cambridge, MA: MIT Press.

——— (1995), *The Engine of Reason, the Seat of the Soul*. Cambridge, MA: MIT Press.

Clark, Andy (1997), *Being There: Putting Brain, Body, and World Together Again*. Cambridge, MA: MIT Press.

Giere, Ronald N. (1988), *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

——— (1999), *Science Without Laws*. Chicago: University of Chicago Press.

——— (2002a), "Scientific Cognition as Distributed Cognition," in Peter Carruthers, Stephen Stitch, and Michael Siegal (eds.), *The Cognitive Basis of Science*. Cambridge: Cambridge University Press, 285–299.

——— (2002b), "Models as Parts of Distributed Cognitive Systems," in Lorenzo Magnani and Nancy Nersessian (eds.), *Model Based Reasoning: Science, Technology, Values*. New York: Kluwer/Plenum, 227–242.

——— (2002c), "Distributed Cognition in Epistemic Cultures," *Philosophy of Science* 69: 637–644.

——— (2003), "The Role of Computation in Scientific Cognition," *Journal of Experimental & Theoretical Artificial Intelligence* 15: 195–202.

——— (2004), "The Problem of Agency in Scientific Distributed Cognitive Systems," *Journal of Cognition and Culture* 4: 759–774.

——— (2006), *Scientific Perspectivism.* Chicago: University of Chicago Press.

Giere, Ronald N., and Barton Moffatt (2003), "Distributed Cognition: Where the Cognitive and the Social Merge," *Social Studies of Science* 33: 301–310.

Hutchins, Edwin (1995), *Cognition in the Wild*. Cambridge, MA: MIT Press.

Knorr-Cetina, Karin (1999), *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA: Harvard University Press.

Latour, Bruno (1986), "Visualization and Cognition: Thinking with Eyes and Hands," *Knowledge and Society: Studies in the Sociology of Culture, Past and Present* 6: 1–40.

Rupert, Robert D. (2004), "Challenges to the Hypothesis of Extended Cognition," *Journal of Philosophy* 101: 389–428.

Suppe, Fredrick (1989), *The Semantic Conception of Theories and Scientific Realism.* Urbana: University of Illinois Press.

Teller, Paul (2001), "Twilight of the Perfect Model Model," *Erkenntnis* 55: 393–415.

——— (2004), "The Law Idealization," *Philosophy of Science* 71: 730–741.

van Fraassen, Bas C. (1980), *The Scientific Image*. Oxford: Oxford University Press.

——— (1989), *Laws and Symmetry*. Oxford: Oxford University Press.

https://doi.org/10.1086/518772 Published online by Cambridge University Press