THOUGHT EXPERIMENTS CONSIDERED HARMFUL

Paul Thagard
University of Waterloo
Draft 4, May, 2013

Thagard, P. (forthcoming). Thought experiments considered harmful. *Perspectives on Science*.

Abstract: Thought experiments can be useful in suggesting new hypotheses and in identifying flaws in established theories. Thought experiments become harmful when they are used as intuition pumps to provide evidence for the acceptance of hypotheses. Intuitions are neural processes that are poorly suited to provide evidence for beliefs, where evidence should be reliable, intersubjective, repeatable, robust, and causally correlated with the world. I describe seven ways in which philosophical thought experiments that purport to establish truths are epistemically harmful. In the philosophy of mind, thought experiments support views that run contrary to empirically supported alternatives.

Introduction

Thought experiments have been influential in philosophy at least since Plato, and they have contributed to science at least since Galileo. Some of this influence is appropriate, because thought experiments can have legitimate roles in generating and clarifying hypotheses, as well as in identifying problems in competing hypotheses. I will argue, however, that philosophers have often overestimated the significance of thought experiments by supposing that they can provide evidence that supports the acceptance of beliefs. Accepting hypotheses merely on the basis of thinking about them constitutes a kind of epistemic hubris with many negative consequences, including the acquisition of false beliefs and the blocking of more promising avenues of inquiry.

I will not attempt to review the extensive literature on thought experiments, which is summarized well by Brown and Fehige (2011). I begin by acknowledging the legitimate contributions that thought experiments can make to scientific and philosophical development. Thought experiments become harmful, however, when they are used as intuition pumps to provide evidence for the acceptance of hypotheses. Defending this claim requires a discussion of the nature and epistemic roles of intuitions and evidence. I argue that intuitions are neural processes that are poorly suited to provide evidence for beliefs, where evidence is supposed to be reliable, intersubjective, repeatable, robust, and causally correlated with the world. The harmfulness of thought experiments is illustrated by their effects on the philosophy of mind, where they have led to the widespread adoption of views that run contrary to empirically supported alternatives. As a philosophical substitute to the use of thought experiments, I propose a superior approach I dub "natural philosophy". Finally, I consider objections to my skeptical view of thought experiments based on Bayesian epistemology and the scientific role of computer simulations.

Thought Experiments Considered Valuable

Philosophers and historians have documented dozens of thought experiments used by leading scientists such as Galileo, Kepler, Newton, Maxwell, Einstein, Schrödinger, and Feynman (e.g. Brown 2011). It would take astonishing philosophical arrogance to argue that such a collection of leading thinkers were epistemologically incompetent. Rather, I want to identify the contributions made by their thought experiments to the development of scientific knowledge, which are primarily of two kinds. First, thought experiments can help to suggest new hypotheses, as when Einstein's imagined himself riding on a beam of light as a way of developing ideas that eventually became relativity theory. The generation of new hypotheses is obviously a crucial part of the growth of science, and no one can fault the use of thought experiments to generate novel concepts and explanatory principles, subject to subsequent empirical evaluation.

Second, thought experiments are often used critically to identify flaws in established or proposed theories. Examples of this use include Galileo's famous attack on Aristotelian physics based on imagining an object formed by tying together a light and a heavy object, and thought experiments by Einstein and Schrödinger aimed at showing conceptual problems in quantum theory. Such attacks can even provide some weak support for an alternative theory, based on the structure shown in figure 1. If theory 1 competes with theory 2, and theory 2 can be shown to generate a contradiction, then theory 1 receives some indirect support.

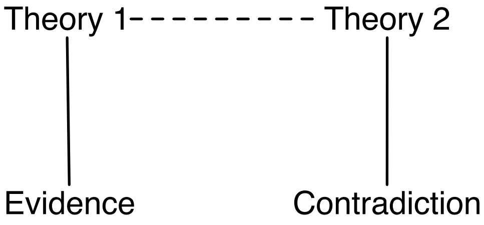

Figure 1. Indirect support for Theory 1 comes from identifying a contradiction in Theory 2. Here straight lines indicate coherence, and dotted lines indicate incoherence, in line with the theory of explanatory coherence described in Thagard (1992, 2000).

It is rare, however, in the history of science to find thought experiments to be used to the exclusion of empirical methods. Galileo conducted real experiments with inclined planes to show the superiority of his views over Aristotelian ones (Drake 1978), and he devoted considerable time to improving experimental instruments such as clocks telescopes, microscopes, and thermometers. Critics of quantum theory have tried (unsuccessfully, so far) to find experimental refutations of it. In contrast, philosophers from Plato's cave allegory to Searle's Chinese Room story have convinced themselves that their thought experiments are alone sufficient to establish important results. On the other hand, when philosophers use thought experiments merely to generate hypotheses or to show inconsistency in competing hypotheses, their reasoning has the same potential legitimacy found in the practice of scientists.

Accordingly, I do not want to say that the use of thought experiments by scientists is harmful, because I think they generally use them properly. Philosophers, in contrast, have often operated under the illusion that thought experiments alone can provide evidence for their theories. To understand this illusion, we need further discussion of the nature of intuitions and evidence.

Intuition in Thought Experiments

Some philosophers have contended that thought experiments are arguments (Norton 2004), but they do not seem to translate easily into standard forms for deduction or induction (Bishop 1999). More plausibly, thought experiments function as what Dennett (1991) called *intuition pumps,* providing a story that generates intuitive judgments about the kinds of situation that the story concerns. This function applies equally well to scientific stories such as Galileo's falling bodies and to philosophical stories such as Searle's Chinese room.

But what are intuitions? The ancient Platonic view, still popular among some mathematicians and philosophers, is that the mind has a special faculty for apprehending abstract objects such as concepts and mathematical structures. Such Platonism, however, is incompatible with a scientific world view that justifies the postulation of non-observed objects only on the basis that they provide causal explanations of what is observed. Justification in this way suffices for theoretically important entities such as electrons and viruses, but fails for abstract objects whose causal effects on equally non-corporeal minds are ineffable.

An alternative view for which evidence is progressively mounting is that intuitions are neuropsychological reactions generated by unconscious processes operating in parallel (e.g. Myers 2002). Compare object recognition: if you see an object consisting of wheels, frame, seat, and handlebars, your brain recognizes it as a bicycle with little conscious deliberation, using interactions among multiple brain areas. Similarly, when you hear a story, your brain processes the information in parallel to generate a judgment that can be emotional as well as cognitive, for example that Aristotle's explanation of falling is ridiculous or that understanding of language by digital computers is absurd.

I conjecture that intuitions result from the three primary neural processes that explain creative consciousness: encoding representation, neural binding, and interactive competition (Thagard forthcoming). Each of these processes is well understand both neurophysiologically and computationally (see for example, Eliasmith and Anderson 2003, Eliasmith 2013, Schröder and Thagard 2013, Thagard and Stewart 2011, Smith and Kosslyn 2007). Representation occurs in the brain when populations of neurons encode inputs by forming synaptic connections between neurons; these connections generate patterns of neural firing that correspond to the inputs. Then a concept such as *tree* is a pattern of firing in a population of interconnected neurons. This view of concepts diverges from the traditional philosophical view of concepts as abstract entities, but is justified because it can explain numerous empirical findings (Blouw, Solodkin, Eliasmith, and Thagard forthcoming).

Binding takes place between such representations through processes that produce new patterns of neural firing that combine previous ones, as in *green tree*. If *green* and *tree* are both neural processes, then their combination is also a neural process that can be generated by specifiable neural mechanisms that take neural firings as inputs and produce new firings that represent the combined concepts. Binding enables the brain to construct representations that go well beyond those provided by sensory inputs, such as *indivisible particle,* the original concept of an atom.

Finally, interactive competition is the process by which numerous bindings are evaluated in parallel by neural firing to determine what combination of representations makes sense of a current situation. Such interactions take place in simpler operations such as object recognition and sentence comprehension, but also in the more complex operation of story understanding (Kintsch 1998). Competition is most easily understand in simple neural networks in which concepts are represented by single artificial neurons. Suppose that you have a neuron for *tree* and another for *telephone pole,* each supported by other neurons that encode visual inputs. These input neurons such as one for *tall* can be connected by excitatory links to the neurons for *tree* and *pole,* with some neurons such as *leafy* supporting only *tree*. Competition occurs by virtue of inhibitory connections between *tree* and *pole* to ensure that firing in one of the neurons suppresses firing in the other. For example, if *tree* and *pole* both get excitation from *tall,* but only *tree* gets excitation from *leafy,* then the firing of *tree* will surpass and suppress the firing of *pole,* capturing the inference the viewed object is a tree rather than a firing pole. In more biologically realistic neural networks, concepts such as *tree* are distributed across many neurons, and competition is a much more complicated process (Grossberg 1987, Thagard and Stewart forthcoming).

Hence it is plausible to give the following account of how thought experiments work in people's minds. Through a combination of representation, binding, and interactive competition, people make sense of a story in a way that generates a reaction in the form of the intuitive sense that a claim is true or false, good or bad. This neural interpretation is compatible with the claim by various cognitive scientists that thought experiments require the construction of mental models, but fleshes it out with a more explicit account of how the brain makes mental models (see Thagard, 2012 ch. 4).

In contrast to Platonic apprehension, which is assumed to yield intuitions that are true, we have to ask about intuitions as neural processes whether they are reliable. The eminent cognitive psychologist Daniel Kahneman (2011) has argued that intuitions should only be trusted under two conditions: when there are identifiable regularities in the world, and when people have had ample opportunities to learn to recognize those regularities. For example, people such as doctors and firefighters sometimes have sufficiently frequent interactions with regular phenomena to be able to form reliable intuitions about likely diseases or fires. Experts have formed neural representations and bindings that correspond sufficiently well to what happens in the world that interactive competition among alternative hypotheses yields intuitions that approximate to reality. In contrast, there are many people with equal levels of confidence whose intuitions are not based on a history of interacting with aspects of a world that contain learnable regularities. For example, astrologers may assert with confidence the existence of connections between the arrangement of stars and planets at the time of people's births and their personalities and life events, but this intuitive confidence is not based on regularities about which they have had opportunity to learn.

Philosopher's intuitions about the nature of knowledge, reality, and morality are not much better than those of astrologers. The stories that philosophers concoct for their thought experiments rarely capture regularities because they have been crafted to support the views that the philosophers already hold. This is circular reasoning, not the employment of evidence to support a hypothesis. Philosophical discussions often consist of the swapping of contrary thought experiments, which led me to propose the maxim that for every thought experiment there is an equal and opposite thought experiment (Thagard 2010, p. 39). Philosophers sometimes claim that their own intuitive reactions toward stories they have constructed constitute evidence for their views, but close examination of the nature of evidence in the next section explodes this overconfidence.

My naturalistic account of intuitions differs markedly from the Platonistic account of thought experiments given by Brown (2011). He maintains that thought experiments serve constructively to obtain *a priori* laws of nature concerning relations between independently existing abstract entities. Then thought experiments perceive such entities in a way analogous to sensory experience of everyday objects. From the perspective of cognitive science, this analogy is seriously defective. The neural processes by which brains interpret sensory inputs are increasingly well understood, but the Platonistic interpretation of thought experiments requires a dualist metaphysics that has become implausible (Thagard 2010). Platonism requires abstract entities grasped by ineffable minds through unintelligible means. In contrast, naturalistic philosophy allies with cognitive science to consider thought experiments as neural processes that combine concepts in ways that may be suggestive but are never conclusive.

Sorenson (1992, p. 289) argues that thought experiments are simple but useful devices for determining the status of propositions with respect to truth, possibility, and necessity. He acknowledges that they have foibles and limited scope, and sometimes lead us badly astray. Nevertheless, he thinks that they can sometimes be useful, like compasses, even though "like compasses, there is mystery to how thought experiments work". On the contrary, the magnetic mechanisms that make compasses work are well known, and progress is ongoing to understand the neural mechanisms of representation, binding, and competition that make thought experiments useful for generating new hypotheses. Unfortunately for the philosophical use of thought experiments, these mechanisms provide no assurance that anything can be learned by thought experiments without empirical evaluation.

Evidence

The scientific revolution of the sixteenth and seventeenth centuries established powerful methods for investigating the world. Instead of relying on traditional texts and authorities, investigators collected their own evidence and generated new theories that explained the evidence better than traditional ideas. Rather than taking the pronouncements of Aristotle, Galen, and the bible as evidence, investigators sought evidence in systematic observations and experiments. On this new understanding, which survives in modern science, evidence has the following important characteristics.

1. Reliability

Kelly (2006) describes evidence as "something which serves as a reliable sign, symptom, or mark of that which it is evidence *of*". A source of evidence is reliable if it tends to yield truths rather than falsehoods (Goldman 1996). We now know that Aristotle, Galen, and the bible were often wrong, for example in Aristotle's belief that the brain's primary function is cooling the body. The alternative sources of evidence developed in the scientific revolution are not infallible, but many truths and few falsehoods have resulted from systematic observations using instruments such as telescopes and microscopes and from controlled experiments such as those practiced by early members of the Royal Society of London and many subsequent scientists. I will argue below that thought experiments have a poor record of reliability.

2. Intersubjectivity

The Royal Society took as its motto "Nullius in Verba", on nobody's word. Systematic observations and controlled experiments do not depend on what any one individual says. Rather they are intersubjective in that different people can easily make the same observations and experiments. The evidence provided by thought experiments in the form of intuitive reactions to stories is far from intersubjective: broad surveys show that there is much more variability in people's reactions to thought experiments than philosophers usually recognize (Knobe and Nichols 2008). Even among philosophers, there is often vehement disagreement about the import of famous thought experiments, and what agreement there is results as much from socialization as veridicality.

3. Repeatability

A major source of the intersubjectivity of systematic observations and controlled experiments is their repeatability: the same person or different persons can get similar results at different times. Replicability of results is not always easy, as some experiments are highly complex and some observable events such as supernova do not recur, but scientific evidence can usually be obtained more than once. Thought experiments can easily be repeated, but only with poor guarantees that similar results will occur at different times. Philosophers have sometimes repudiated their own earlier thought experiments.

4. Robustness

Another common characteristic of scientific evidence is robustness, which Wimsatt (2007 p. 196), characterizes as follows: “Things are robust if they are accessible (detectable, measurable, derivable, definable, producible, or the like) in a variety of independent ways." Robustness is more than repeatability – it means repeatability in different ways such as using different kinds of instruments. For example, evidence about neural functioning can be gained from many kinds of observation of the brain, such as damage caused by strokes, single cell recording, PET scans, fMRI scans, and transcranial magnetic stimulation. Thought experiments are supposed to be evidence of a special kind that suffices on its own, with no need to consult other methods, so they lack robustness.

5. Causal correlation with the world

Finally, a fifth feature of scientific evidence based on systematic observation or controlled experiments is that there is usually some basis for concluding that the evidence is causally connected with the world about which it is supposed to tell us. We know that telescopes and microscopes provide evidence because reflected light enters the eyes of observers and stimulates their retinas, providing a causal influence from what is observed to the human observation. In controlled experiments, the causality runs both ways, because the observer manipulates the world in order to observe the results, for example when Robert Hooke used an air pump to investigate pressure. The intuitions that philosophers generate about matters metaphysical, epistemological, and ethics are too far removed from interactions with the world to have any causal correlations. This distance is not a problem if thought experiments are merely used to generate hypotheses that can be evaluated with respect to evidence that does causally correlate with the world, but undermines claims that thought experiments by themselves generate anything that deserves to be called evidence. As I argued earlier, philosophers' intuitions are rarely based on expertise gained from interacting with the world.

Because of these five differences, we should judge the intuitions resulting from thought experiments as far inferior to scientific evidence based on systematic observations or controlled experiments. The problem, however, about thought experiments is not just inferiority, but actual harm that they can cause to the development of knowledge when used beyond their appropriate generative role.

The Seven Sins of Thought Experiments

There are many ways in which philosophical thought experiments that purport to establish truths are epistemically harmful. A bit luridly, I will call them the "seven sins of thought experiments".

1. Generating falsehoods

I claimed above that thought experiments are unreliable in often leading to false conclusions. The good reputation of scientific thought experiments comes from cases such as Galileo's falling bodies story where the investigator got the right answer, but there are also scientific cases where thought experiments have been used to support conclusions now viewed as false. For example, Newton's famous bucket thought experiment was used by him to support the idea of absolute space, which is incompatible with currently accepted relativity theory. It would be interesting to catalog other thought experiments in science that have led to superseded theories.

In philosophy, there are many thought experiments that have produced dubious conclusions. There is a battery of thought experiments designed to separate mind from body and reject the identification of mental processes with brain processes, including:

René Descartes' claim that he could imagine himself without a body, so he is essentially a thing that thinks.

Saul Kripke's claim that pain is not a brain process because he could imagine possible worlds in which is not.

Frank Jackson's claim that there is more to color than brain processes because he could imagine a neuroscientist who supposedly knew everything about color but learned something by gaining the experience of color.

David Chalmer's claim that the possibility of "zombies" (which are just like us physically but lack consciousness) shows that consciousness is not a brain process.

My basis for thinking that all of these thought experiments have yielded false conclusions is that in the past two decades there has been a huge accumulation of evidence in favor of the identification of mind and brain (see e.g. Anderson 2007; Smith and Kosslyn 2007;

Thagard 2010). The mind-brain identity hypothesis is part of the best explanation of a wide range of phenomena including perception, inference, and emotion whose neural mechanisms are becoming increasingly well understood.

Similar problems could be shown for many other famous thought experiments in the philosophy of mind, from Plato's bogus demonstration in the *Meno* that high-level mathematical knowledge is innate, to Putnam's Twin Earth thought experiment that mangles chemistry to show that "meanings just ain't in the head" (Thagard 2012, ch. 18). Searle's (1980) famous Chinese room thought experiment was taken by him and some others to prove that digital computers cannot represent the world, but there are currently robotic cars that navigate complex environments in ways that imply that they actually do have meaningful representations (Parisien and Thagard 2008). Hence philosophical thought experiments are unreliable in that they often generate false conclusions.

2. Underspecifying situations

Kathleen Wilkes (1993) argues forcefully that in philosophy thought experiments are usually both problematic and misleading. One of the problems she points out that makes thought experiments inferior to controlled experiments in science is uncertainty concerning the relevant background conditions. Philosophical thought experiments are underspecified about what they are taking for granted. For example, some personal identity thought experiments blithely assume the science fiction scenario of humans being transported from spaceships to planets, without considering any of the physical requirements for decomposing and recomposing trillions of cells. Brain-in-a-vat thought experiments omit calculations concerning the computational feasibility of pumping a full set of experiences into each brain. Underspecification allows philosophers to assume that whatever they can imagine is actually possible, but, as with Putnam's Twin Earth experiment, they are often engaging in ignorance-based reasoning.

3. Justifying ignoring relevant evidence

Wilkes (1993) shows that far more interesting conclusions about personal identity can be reached by reflecting on actual cases such as people who have various mental illnesses. One of the main flaws of thought experiments is that they appear to justify ignoring kinds of empirical evidence that are in fact highly relevant. The assumption is that philosophy should be aimed at discovering essences, properties that hold in all possible worlds, so there is no point in collecting evidence that only applies to the world we happen to inhabit. In contrast, the approach to philosophy that I sketch below claims that it is much more productive for philosophical reflections to build on what is known about this world, because there is scant hope of gaining knowledge about what is true of all possible worlds,

4. Circular reasoning

As my earlier discussion of intuition suggested, the philosophical method of thought experimentation is inherently circular, starting with a hypothesis and then making up pseudo-evidence to support the hypothesis. Scientific reasoning, in contrast, avoids circularity because the evidence it recruits from observation and experiment depends on successful interactions with the world, not just the pre-existing beliefs of the thought experimenter.

5. Blocking inquiry

Charles Peirce advocated the central methodological principle: Do not block the way of inquiry. Thought experiments used generatively and critically are consistent with this principle, but thought experiments used dogmatically are not, because they can be taken to imply that alternative views are necessarily false. It is fortunate that philosophical arguments against mind-brain identity and artificial intelligence have been ignored by scientific researchers in pursuit of their legitimate goals, the pursuit of which has yielded substantial progress. As discussed above, thought experiments in both science and philosophy can aid inquiry by generating new ideas worth exploring, but dogmatic thought experiments like the ones that abound in metaphysics and the philosophy of mind are impediments to further investigations.

6. Wasting time

Philosophers are mostly clever and industrious people who have much to contribute to understanding issues of great importance. It is sad to see them wasting time on generating and responding to thought experiments rather than dealing with ethical, epistemological, and metaphysical questions using what is known about the world and the people in it.

7. Casting philosophy into disrepute

The final harmful effect of thought experiments is that their overuse lends credence to the dismissal of philosophy by scientists and others who are serious about understanding the world. Here are some negative comments made about philosophy. Richard Feynman is supposed to have said that scientists are explorers, but philosophers are tourists, and that philosophy of science is about as useful to scientists as ornithology is to birds. It has also been claimed that philosophy is to cognitive science what tin cans tied to a car are to a wedding, that philosophy is to science as alcohol is to sex, and that philosophy is to science as pornography is to sex. These comments reflect the common view of philosophy as a navel-gazing, angel-counting enterprise divorced from matters worth thinking about, a view that is encouraged by use of dogmatic thought experiments to proclaim how things must be in isolation from empirical studies of how they are.

In contrast, I think philosophy has a great contribution to make to scientific and social issues, dealing with questions that are more general and more normative that usual scientific work (Thagard 2009). The kind of natural philosophy I advocate below should be better appreciated by non-philosophers.

Natural Philosophy

It is easy to imagine the embattled thought experimenter responding as follows. "But wait, if we don't get to use thought experiments to support our views about knowledge, reality, and morality, what would we do? Philosophers shouldn't just be the historians of past thinkers, nor should they devolve into science journalists picking up on ephemeral findings. Without thought experiments, philosophy has no way of answering eternal questions, so abandoning them is abandoning philosophy."

This lament ignores the long and distinguished history of philosophers who have addressed the most important issues in epistemology, metaphysics, and ethics in ways connected with scientific findings rather than made-up stories. A partial list includes the following: Aristotle, Locke, Hume, Mill, Peirce, Russell (after 1920), Dewey, Quine (after 1950), and Kuhn. Tying philosophy closely to science is sometimes called "naturalistic" , "naturalized", or "second" philosophy (Maddy 2007), but I would like to revive an old term and simply call it "natural philosophy". This term was commonly used for what we now call "science", before the term "science" took over in the 19th century, and investigators such as Newton called themselves natural philosophers. Here are some principles that constitute a start on a manifesto for a new version of natural philosophy.

1. Philosophy should aim to answer fundamental questions about the nature of knowledge, reality, and morality.
2. Answers to these questions, should come, not from highly subjective a priori reasoning, but from reflection on empirical and theoretical discoveries about the world.
3. Philosophy is concerned with truths about this world, not with unknowable conjectures about all possible worlds.
4. Philosophy differs from science in being more general, ranging across all of he sciences, and in being more normative, concerned with how things can be made better.
5. Thought experiments may be useful in generating new hypotheses, but they fail to provide evidence in support of hypotheses.

This naturalistic methodology provides ample room for philosophical investigations that should be much more productive than mere speculation.

Philosophers from Leibniz to Kripke and David Lewis have assumed that knowledge can be gained about what is true in all possible worlds, not just the world we actually inhabit. There are a number of problems with this assumption. First, it clearly goes beyond the methods known from scientific practice to produce reliable knowledge, including systematic observations, controlled experiments, and inferences to non-observable entities on the basis that they provide the best explanation of the evidence. Inferences about possible worlds lack a track record. Second, the method by which truths about all possible worlds is allegedly gained is through considerations of what is conceivable; but if the account of intuition given above is correct, then what people take to be conceivable is just a reflection of their current opinions, not any kind of transcendent reality. Hence knowledge of possible worlds besides the one we inhabit is unattainable, so philosophers should follow scientists in trying to figure out how things actually are, not how they have to be.

Objections: Computer Simulations and Bayesian Epistemology

I now want to consider two considerations that might mistakenly be taken to justify a stronger role for thought experiments in philosophy and science than I have been allowing. First, it has sometimes been suggested (Stuart, forthcoming) that computer simulations used in many areas of science are thought experiments, which might imply that they have a more central scientific (and potentially philosophical) role than I have allowed. This objection is potentially damaging to me, as I have been publishing computer simulations in cognitive science journals since the late 1980s.

From this experience, however, I can assure you that the function of computer simulations is very different from that of thought experiments aimed at directly justifying experiments. The major function of computer simulations is to connect theories understood as descriptions of mechanisms with empirical results that provide evidence for (and potentially against) those theories. The methodology is displayed in figure 2, which shows how scientists often move from theories about mechanisms to mathematical/computational models, then to running computer programs whose behavior can be compared with the results of behavioral or neural experiments. Computer simulations can also have a powerful generative role, because one way in which cognitive scientists develop theories about relevant mechanisms is to think about how to write a computer program that performs a desired task. But I know of no case in the cognitive or other sciences whether the mere act of producing a computer simulation has been taken as evidence for a theory.

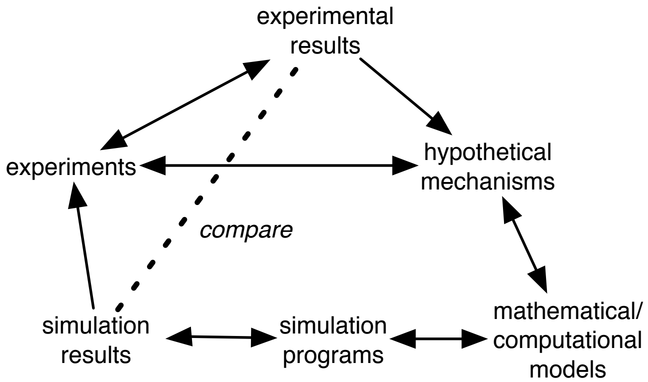

Figure 2. The role of computer models in developing and testing theories about mechanisms. Lines with arrows indicate causal influences in scientific thinking. The dashed line indicates the comparison between the results of experiments and the results of simulations. From Thagard 2012, p. 11.

The second defense of thought experiments is not one that I have seen in print, but it occurred to me as the result of reflection on Bayesian ideas that are currently highly influential in both philosophy of science and cognitive science (e.g. Glymour 2001). Bayes theorem is often interpreted as a probabilistic way of justifying hypotheses based on evidence, captured in the following formula: P(H/E) = P(H) X P(E/H) / P(E). In words, the probability of a hypothesis given evidence can be calculated by multiplying the probability of the hypothesis times the probability of the evidence given the hypothesis, all divided by the probability of the evidence. The relevance to thought experiments comes from the inclusion of P(H), the prior probability. Perhaps thought experiments in science and philosophy have no connection with evidence, but they might be taken to contribute to prior probabilities and hence to the acceptance of a hypothesis. Hence on Bayesian principles thought experiments might have a legitimate justificatory role! This role is much weaker than contributing a priori proofs of necessary truths, but nevertheless would counter my contention that thought experiments cannot contribute to justification.

We have to accept Bayes theorem as a straightforward deductive consequence of the axioms and definitions of probability, but interpreting and applying it to the justification of scientific and philosophical hypotheses is much more contentious. There are numerous problems with the Bayesian approach to hypothesis evaluation that I have discussed elsewhere (Thagard 2000, ch. 7; Thagard 2004; Thagard 2012, ch. 5). One is the problem of interpreting probabilities, which are variously understood as frequencies or degrees of belief. If probabilities are frequencies, they do not seem to apply to hypotheses or to relations between hypotheses and evidence, because there are no relevant populations. On the other hand, if probabilities are degrees of belief, they are purely subjective and have no connection with truths about the world. It may well be that thought experiments tend to increase people's subjective prior probabilities, but there is no reason to think that this result correlates with any sort of objective truth. Empirically, subjective degrees of belief do not map well onto human psychology (Kahneman and Tversky 2000), so the Bayesian approach to epistemology is more an exercise in *a priori* analysis than an account of how people can justify their beliefs.

A second major problem with Bayesian epistemology is that that it ignores many difficult problems of representing and computing complex inferential situations. For example, available algorithms rule out causation where the influences of variables run in a circle. However, such cycles are common in real world systems such as ecologies and climates. Without the idealization of acyclic graphs, Bayesian updating is computationally intractable. A third problem is that Bayesian calculations require large numbers of conditional probabilities about which information is rarely available, so they simply have to be made up. Hence the apparent elucidation of justification by Bayesian calculation is illusory.

Accordingly, I do not recommend Bayesian approaches to hypothesis evaluation, but they might be attractive to theorists who see them as offering a weak but legitimate role in contributing to justification via thought experiments.

Conclusion

Even if thought experiments do not succeed in establishing scientific or philosophical truths just by thinking, the study of them is still a fitting topic for philosophical investigation. I have argued that their appropriate roles are generative and critical, rather than justificatory. My methodology in reaching these conclusions has been naturalistic, looking at actual cases of thought experiments and taking into account what is known about the mental processes underlying them such as intuition and mental modeling. We can get a much better understanding of how thought experiments operate by considering how they rely on neuropsychological mechanisms such as representation, binding, and interactive competition. Seeing thought experimentation as a natural neural process undermines claims that it relies on a special, non-physical process such as divine communication or apprehension of Platonic forms.

Additional skepticism about the justificatory role of thought experiments results from contrasting the properties of the intuitions they generate with the features of good scientific evidence such as reliability, intersubjectivity, repeatability, robustness, and causal correlation with the world. Lacking these desirable properties of evidence, thought experiments used inappropriately can generate at least seven kinds of harm, including generating falsehoods, underspecifying situations, ignoring relevant evidence, circular reasoning, blocking inquiry, wasting time, and casting philosophy into disrepute. Neither considerations of computer simulations nor Bayesian epistemology suffices to revive the justificatory role of thought experiments. Hence I hope that philosophers will begin to follow scientists in restricting the use of thought experiments to the generation of hypotheses and the criticism of alternatives, thereby avoiding the epistemic harm that results from using them directly to justify the acceptance of hypotheses.

Acknowledgements. Thanks to Mike Stuart, Yiftach Fehige, and two anonymous referees for helpful comments. My research is supported by the Natural Sciences and Engineering Research Council of Canada.

References

Anderson, J. R. 2007. *How can the mind occur in the physical universe?* Oxford: Oxford University Press.

Bishop, M. 1999. "Why thought experiments are not arguments." *Philosophy of Science, 66*, 534-541.

Blouw, P., Solodkin, E., Eliasmith, C., & Thagard, P. forthcoming. "Concepts as semantic pointers: A theory and computational model." *Unpublished manuscript, University of Waterloo.*

Brown, J. R. 2011. *The laboratory of the mind* 2nd ed.. London: Routledge.

Brown, J. R., & Fehige, Y. 2011. "Thought experiments." *Stanford Encyclopedia of Philosophy*. Retrieved from http://plato.stanford.edu/entries/thought-experiment/

Dennett, D. 1991. *Consciousness explained*. Boston: Little, Brown.

Drake, S. 1978. *Galileo at work: His scientific biography*. Chicago: University of Chicago Press.

Eliasmith, C. 2013. *How to build a brain*. Oxford: Oxford University Press.

Eliasmith, C., & Anderson, C. H. 2003. *Neural engineering: Computation, representation and dynamics in neurobiological systems*. Cambridge, MA: MIT Press.

Glymour, C. 2001. *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT Press.

Goldman, A. 1986. *Epistemology and cognition*. Cambridge, MA: Harvard University Press.

Grossberg, S. 1987. "Competitive learning: From interactive activation to adaptive resonance." *Cognitive Science, 11*, 22-63.

Kahnemann, D. 2011. *Thinking fast and slow*. Toronto: Doubleday.

Kelly, T. 2006. "Evidence." Retrieved from http://plato.stanford.edu/entries/evidence/

Kintsch, W. 1998. *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

Knobe, J., & Nichols, S. 2008. *Experimental philosophy*. Oxford: Oxford University Press.

Maddy, P. 2007. *Second philosophy: A naturalistic method*. Oxford: Oxford University Press.

Myers, D. G. 2002. *Intuition: Its powers and perils*. New Haven: Yale University Press.

Norton, J. D. 2004. "Why thought experiments do not transcend empiricism." In C. Hitchcock Ed., *Contemporary debates in the philosophy of science* pp. 44-66. Oxford: Blackwell.

Parisien, C., & Thagard, P. 2008. "Robosemantics: How Stanley the Volkswagen represents the world." *Minds and Machines, 18*, 169-178.

Schröder, T., & Thagard, P. 2013. "The affective meanings of automatic social behaviors: Three mechanisms that explain priming." *Psychological Review, 120*, 255-280.

Searle, J. 1980. "Minds, brains, and programs." *Behavioral and Brain Sciences, 3*, 417-424.

Smith, E. E., & Kosslyn, S. M. 2007. *Cognitive psychology: Mind and brain*. Upper Saddle River, NJ: Pearson Prentice Hall.

Sorenson, R. A. 1992. *Thought experiments*. New York: Oxford University Press.

Stuart, M. forthcoming. Paul Thagard's skepticism: A refutation. *Perspectives on Science*, ADD DETAILS.

Thagard, P. 1992. *Conceptual revolutions*. Princeton: Princeton University Press.

Thagard, P. 2000. *Coherence in thought and action*. Cambridge, MA: MIT Press.

Thagard, P. 2004. "Causal inference in legal decision making: Explanatory coherence vs. Bayesian networks." *Applied Artificial Intelligence, 18*, 231-249.

Thagard, P. 2009. "Why cognitive science needs philosophy and vice versa." *Topics in Cognitive Science, 1*, 237-254.

Thagard, P. 2010. *The brain and the meaning of life*. Princeton, NJ: Princeton University Press.

Thagard, P. 2012. *The cognitive science of science: Explanation, discovery, and conceptual change*. Cambridge, MA: MIT Press.

Thagard, P. forthcoming. Creative intuition: How EUREKA results from three neural mechanisms. In L. M. Osbeck & B. S. Held Eds., *Rational intuition: Philosophical roots, scientific investigations* Cambridge: Cambridge University Press.

Thagard, P., & Stewart, T. C. 2011. The Aha! experience: Creativity through emergent binding in neural networks. *Cognitive Science, 35*, 1-33.

Thagard, P., & Stewart, T. C. forthcoming. Two theories of consciousness: Semantic pointer competition vs. information integration. *Unpublished manuscript, University of Waterloo*.

Wilkes, K. V. 1993. *Real people: Personal identity without thought experiments*. Oxford: Oxford University Press.

Wimsatt, W. C. 2007. *Re-engineering philosophy for limited beings*. Cambridge, MA: Harvard University Press.