WHY DO THOUGHT EXPERIMENTS WORK?

Nancy J. Nersessian
Program in History of Science
Princeton University
Princeton, NJ 08540

Abstract

Thought experiments have played a central role in historical cases of major conceptual change in science. They are important in both constructing new representations of nature and in conveying those representations to others. It is proposed that research into the role of mental modelling in narrative comprehension can illuminate how and why thought experiments work. In constructing and "running" the thought experiment, we make use of inferencing mechanisms, existing representations, and general world knowledge to make realistic transformations from one possible physical state to the next and this process reveals impossibility of applying existing concepts to the world and pinpoints the locus of needed conceptual reform.

Introduction

Throughout the history of science there are numerous instances where scientists who have created major conceptual innovations employed thought experiments and used them as a means of conveying the new representations to others. Some, such as Einstein (1945), have claimed that a specific thought experiment was essential in their initial construction of the new representation. Using Einstein as an example, Wertheimer (1945) confirmed the centrality of that thought experiment in Einstein's "*Gestalt* switch", but did not analyze how and why it worked to create the conceptual change. Indeed, despite widespread recognition of the use and importance of thought experiments in conceptual change in science, there has been little analysis - by either psychologists or philosophers and historians of science - of how they function and why they are effective in bringing about change.

Within philosophy and history of science, earlier scholarship presented two poles of interpretation of their role in creating conceptual change. Duhem dismissed them as bogus precisely because they are "not only not realized but incapable of being realized" (Duhem 1914, p.202), i.e., they are not "experimental" in the customary sense. Koyré (1939, 1968), on the other hand, argued that their logical force is so compelling that they supplant real experimentation in the construction of new representations for phenomena. The "thought" part of the experiment predominates and shows the synthetic *a priori* nature of scientific knowledge.

Contemporary historians and philosophers of science by and large reject both extremes of empiricism and rationalism. They acknowledge that thought experiment, while not eliminating the need for real experiment, is an important heuristic for creating conceptual change in science. There have been a few recent attempts to analyze the function of thought experiments. While these are sketchy and limited, some do yield useful insights.

Two analyses focus on the empirical force of thought experiments. The notion that an experiment having real empirical consequences can take place in thought seems paradoxical. Yet, Kuhn claims:

"Thought experiment is one of the essential analytical tools which are deployed during crises and which then help to promote basic conceptual reform" (Kuhn, 1964, p.263).

He argues that thought experiments perform this function by showing that there is no consistent way, *in actual practice*, of using accepted existing concepts. That is, the thought experiment reveals that it is not possible to apply the conceptualizations we have of pehnomena consistently to real-world phenomena and that this practical impossibility translates into a logical requirement for conceptual reform. Gooding (1990) in his analysis of Faraday's experimental practices, picks up on Kuhn's analysis, rendering the empirical force of thought experiments in terms of their demonstration of what he calls the "impracticability of doing". Gooding is concerned to show how the real world experimenter's "procedural knowledge", i.e., tacit and explicit knowledge of practical skill, is utilized in constructing and manipulating thought experiments. Both of these analyses contain important insights about the relationship between the conceptual and the experiential dimensions of knowledge.

Two other analyses focus on the logical force of thought experiments. First, Brown (1987) claims that thought experiments are a species of *a priori*

reasoning, but they get at something that cannot be derived from logical argumentation. His positive suggestion is that they provide a special window through which the mind grasps universals. It seems that Brown is trying to capture the idealizing function of thought experiments. However, his approach through linguistic analysis alone does not afford the possibility of understanding their experimental nature and why they have empirical consequences. Second, Norton (1986) claims that thought experiments can, in essence, be reconstructed as, and replaced by, arguments. This analysis is the most sympathetic for philosophers who wish to restrict reasoning to logical argumentation - whether deductive or inductive. Certainly thought experiments contain an argument. However, as Norton himself acknowledges, the argument can only be constructed *after the fact*, i.e., the argument is not evident until after the thought experiment has been executed. Additionally, he supports his claim by noting the presentation of a thought experiment contains particulars irrelevant to the generality of the conclusion. While this is correct, his emphasis reveals that he has failed to see the *constructive function of the narrative form* in which thought experiments are customarily presented.

This paper proposes that construing thought experiments as a species of mental modelling offers the best possibility of explaining how it is that an experiment made in thought can have both the logical and empirical force to compel and communicate conceptual change that the historical record seems to indicate. While certainly not uncontroversial, the notion that mental simulation is a central aspect of cognition is widely accepted among cognitive scientists. However the potential role of "mental modelling" in conceptual change, both in science and in general, has not been investigated. My aim is to show that bringing thought experiments under the perview of mental models research will enable us to better understand their function and their relationship to other aspects of scientific thinking. The weakness of the analyses of Brown and Norton lie in their taking the traditional philosophical route of construing a thought experiment as a form of reasoning with propositions. What Kuhn and Gooding are trying to capture may, in fact, not be able to be represented propositionally. My hypothesis is that propositional representations cannot capture the experimental dimension of a thought experiment. Running a *mental simulation is required for a it to be both "thought" and "experimental"*. The analysis presented here conceives of the original thought experiment as the construction of a mental model by the scientist who imagines a sequence of events. She then uses a narrative form to describe the sequence in order to communicate the experiment to others, i.e., to get them to run the corresponding simulation. While thought experiments are used extensively in scientific thinking and teaching, we will focus on a few that have been instrumental in creating major conceptual change in science.

Case Studies: Galileo and Einstein

Although there are many great thought experimenters in the history of science, the ones who have attracted the most attention are Galileo and Einstein. Bowing to tradition and hoping to capitalize on what may already be familiar to the reader, I will give a brief presentation of a few of the thought experiments devised by Galileo and by Einstein to convey some sense of the variety such experiments display and to elicit some common features for further analysis.

Galileo's importance as a pivotal figure in the transition from the qualitative categories of aristotelian and medieval theories of motion to the quantitative representation of motion provided by Newton's mechanics is widely recognized. As shown in analyses by Koyré (1939) and Clavelin (1968), among others, Galileo drastically transformed the problem of how to go about constructing a mathematical representation of the phenomena of motion. That process, which Koyré called "mathematization", required constructing an idealized representation, quantifying this representation, and mapping the quantified representation back onto the real world. While it is now clear that Galileo must have performed many more real-world experiments than Koyré would have liked (See, e.g., Drake 1973; Naylor 1976; and Settle 1961), no one would deny the importance of his use of thought experiment in the mathematization process. Take, as example, his analysis of falling bodies (Galilei 1683, pp. 62-86).

THOUGHT EXPERIMENT G1: According to the aristotelian theory heavier bodies fall faster than lighter ones. This belief rests on a purely qualitative analysis of the concepts of 'heaviness' and 'lightness'. Galileo argued against this belief and constructed a new, quantifiable representation through a sustained analysis using several thought experiments and limiting case analyses. The outline of his use of these procedures is as follows. He calls on us to imagine we drop a heavy body and a light one, made of the same material, at the same time. We would customarily say that the heavy body falls faster and the light body more slowly. Now suppose we tie the two

bodies together with a very thin - almost immaterial - string. The combined body should both fall faster and more slowly. It should fall faster because a combined body should be heavier than two separate bodies and should fall more slowly because the slower body should retard the motion of the faster one. Clearly something has gone amiss in our understanding of 'heavier' and 'lighter'. Having pinpointed the problem area, Galileo then goes on to show that it is a mistake to extrapolate from what is true at rest to what happens when bodies are in motion. That is, he has us consider that when the two bodies are at rest, the lighter will press on the heavier, and therefore the combined body is heavier. But when we imagine the two bodies are in motion, we can see the lighter does not press on the heavier and thus does not increase its weight. What Galileo has done up to this point is use the thought experiment to reveal the inconsistencies in the medieval belief, the ambiguities in the concepts, and the need to separate the heaviness of a body from its effect on speed in order to analyze free fall. He then goes on, using the methods of thought experiment and limiting case analysis in tandem to show that the apparent difference in the speed of falling bodies is due to the effect of the medium and not to the difference in heaviness between bodies.

As the historian Clavelin has pointed out, it is crucial for quantifying the motion of falling bodies that 'heaviness' not be the cause of the difference in speed because then we could not be sure that motion would be the same for all bodies. Galileo went on to use a further thought experiment to demonstrate that the observed differences in speed should be understood as being caused by the unequal way media lift bodies.

THOUGHT EXPERIMENT G2: Galileo asks us to suppose, for example, that the density of air is 1, that of water 800, of wood 600 and of lead 10,000. In water the wood would be deprived of 800/600th of its weight, while lead would be deprived of 800/10,000th's. Thus, the wood would actually not fall (i.e., would float) and the lead would fall more slowly than it would in a less dense medium, such as air. If we extrapolate to a less dense medium, such as air, we see that the differential lifting effect is much less significant (e.g., 1/600 to 1/10,000 in air). The next move is to consider what would happen in the case of no medium, i.e., in extrapolating to the limiting case. With this move, Galileo says "I came to the opinion that if one were to remove entirely the resistance of the medium, all materials would descend with equal speed." (Galilei 1638, p. 75). Having performed the extrapolation in this way we can quantify this idealized representation of the motion of a falling body and know that it is relevant to actual physical situations; we need only add back in the effects of a medium.

Galileo repeatedly used thought experiments and limiting case analyses in tandem as shown by this example both in constructing a quantifiable representation of bodies in motion and in attempting to convey this new representation to others. Later I will propose the cognitive function of thought experiments and of limiting case analysis are much the same.

Einstein employed several thought experiments in developing the special and general theories of relativity. These thought experiments were central in his reconceptualization of 'space', 'time', and 'simultaneity'. He began his paper, "On the Electrodynamics of Moving Bodies" (1905), with the following thought experiment.

THOUGHT EXPERIMENT E1: Einstein asks us to consider the case of a magnet and a conductor in relative motion. There are two possibilities for explaining how a current is produced in the conductor. In the first case, the magnet is at absolute rest and the conductor moving. According to electromagnetic theory, the motion of the conductor through the magnetic field produces an electromotive force that creates a current in the conductor. In the second case, the conductor is at rest and the magnet moving. In this case, again according to electromagnetic theory, the motion of the magnet creates a changing magnetic field which induces an electric field that in turn induces a current in the conductor. However, with respect to the relative motions, it makes no difference whether it is the magnet or the conductor that is considered to be in motion. But according to the Maxwell-Lorentz electromagnetic theory, the absolute motions do create a difference in how we would explain the production of a current in the conductor. Since the explanatory asymmetry could not, in principle or in practice, be accounted for by the observable phenomena - the measurable current in the conductor - Einstein argued that this supported his conclusion that "the phenomena of electrodynamics as well as of mechanics possess no properties corresponding to the idea of absolute rest" (p.37).

Although we cannot discuss them here, two more thought experiments figure crucially in the complete analysis. The second thought experiment in the paper is the most famous one in which Einstein constructed an operational definition for the concept of simultaneity.

In a similar manner many thought experiments figured in Einstein's constructing and communicating the general theory of relativity, i.e., the

field representation of gravitational action. We will just consider one he presented in various formats but claims to have first conceived in 1907.

THOUGHT EXPERIMENT E2: Einstein (1917, pp.66-70) asks us to imagine that a large opaque chest, the interior of which resembles a room, is located in space far removed from all matter. Inside there is an observer with some apparatus. In this state, the observer would not experience the force of gravity and would have to tie himself with strings to keep from floating to the ceiling. Now imagine that a rope is connected to the outer lid of the chest and a "being" pulls upward with a constant force, producing uniform acceleration. The observer and any bodies inside the chest would now experience the very same effects, such as a pull towards the floor, as in a gravitational field. The experiment demonstrates that the behavior of a body in a homogeneous gravitational field and one in a uniformly accelerated frame of reference would be identical. Once we see that there is no way of distinguishing these two cases we can understand the importance of the Newtonian law that the gravitational mass of a body equals its inertial mass: these are just two manifestations of the same property of bodies. That is, we have a different interpretation for something we already knew.

Common Features

Before beginning to sketch a way of understanding thought experiments and their role in conceptual change in terms of the mental models framework, we need first to glean some common features of thought experiments from the narratives presented above. While there is great variety among thought experiments, in general, the ones presented here do exemplify important salient features.

FEATURE 1: By the time a thought experiment is public it is in the form of a narrative. The narrative has the character of a simulation. It calls upon the reader/listener to imagine a dynamic scene - one that unfolds in time. The invitation is to follow through a sequence of events or processes *as one would in the real world.* That is, even if the situation my seem bizarre or fantastic, such as being in a chest in outer space (E2), there is nothing bizarre in the unfolding: objects float as they would in the real world in the absence of gravity. The assumption is that if the experiment could be performed, the chain of events would unfold according to the way things usually take place in the real world.

FEATURE 2: A thought experiment embodies specific assumptions - either explicit or tacit - of the representation under investigation. It usually exposes inconsistencies or exhibits paradoxes that arise when we try to apply certain parts of that representation to a specific situation, such as 'heavy' and 'light' to falling rocks (G1). The paradox can take the form of a contradiction in the representation, e.g. it requires that an object be both heavy and light, or of something being not physically possible, e.g., observing the asymmetry required by electromagnetic theory (E1).

FEATURE 3: By the time a thought experiment is presented it always works and is often more compelling than most real-world experiments. We rarely, if ever, get a glimpse of failed thought experiments or avenues explored in the construction of the one presented to us. Some experiments, such as G2, could potentially be carried out - at least until the analysis extrapolates to the limit. Others, such as E1, underscore that doing a real-world experiment could not provide the data the theory requires. While others, such as E2, are impossible to carry out in practice, either in principle or because we do not yet have the requisite level of technological achievement. However, once understood, a thought experiment is usually so compelling in itself that even where it would be possible to carry it out, the reader feels no need to do so. The constructed situation is apprehended as pertinent to the real world either by revealing something in our experience that we did not see the import of before - e.g., the measurable current in the stationary and in the moving conductor is the same, so on what basis can we support the difference in theoretical explanation? - or by generating new data - e.g., in the case of no medium, lead and wood would fall at the same speed - or by making us see the empirical consequences of something in our existing representation - e.g., the attributes called 'gravitational mass' and 'inertial mass' are the same property of bodies.

FEATURE 4: The narrative presentation has already made some abstraction from the real-world phenomena. For example, certain features of objects that would be present in a real experiments are eliminated, such as the color of the rocks and the physical characteristics of the observers. That is, there has been a prior selection of the pertinent dimensions on which to focus, which evidently derives from our experience in the world. We know, e.g., that the color of a rock does not effect its rate of fall. This feature strengthens our understanding of the depiction as that of a prototypical situation of which there could be many specific instances. In more colorful narratives there may be more irrelevant features in the exposition, but these most often serve to reinforce crucial aspects of the experiment. For example, in one version of the chest - or "elevator" - experiment

(E2), Einstein depicts the physicist as being drugged and then waking up in a box. This colorful detail served to reinforce the point that the observer could not know beforehand if he were falling in outer space or sitting in a gravitational field. And, in the version discussed above, the opacity of the chest is to prevent the observer from seeing if there are gravitational sources around.

Thought Experimenting as Mental Modelling

Rendering thought experiments as a species of mental modelling supports the interpretation that when they are employed in conceptual change, they are "essential analytical tools" in the process. While there is an extensive "mental models" literature, only recent work by Qin & Simon (1990) attempts, explicitly, to capture the simulation process of a thought experiment, which they characterize as a "mental image". There are many things we do not yet understand about the processes involved in mental simulation and how these differ from propositional reasoning. Thus, I can only hope the sketch I present will persuade the reader that following this direction does offer good prospects for accounting for both the "thought" and the "experiment" aspects of thought experiments, and for explaining how they can be "essential tools" in the process of conceptual change.

We can only speculate about what goes on in the mind of the historical scientist in the original thought experiment. Scientists have rarely been asked to discuss the details of how they went about setting up and running such experiments. However, reports of thought experiments are always presented in the form of narratives that call upon the reader/listener to simulate a situation in his or her imagination. Thus, drawing on what we think we know both about the processes through which we imagine or "picture" in general, and through which we comprehend any narrative, may help us to answer that most perplexing question about thought experiments: *how can an "experiment" carried out in thought have such powerful empirical force?* The most pertinent aspect of mental models research for this analysis is that which investigates the hypothesis that understanding a narrative involves the construction of a mental model. In the case of thought experiments we need to understand how: (1) a narrative facilitates the construction of an experimental situation in thought and (2) thinking through the experimental situation has real-world consequences. Framed in mental models terms, the "thought" dimension would include constructing a mental model and "running" a mental simulation of the situation depicted by the model, while the "experimental" dimension comprises the latter and the connection between the simulation and the world.

Briefly, the mental models thesis about text comprehension is that understanding the meaning of a narrative involves relating linguistic expressions to models; i.e., the relationship between the words and the world is mediated by the construction of a structural analog to the situations, processes, objects, events, etc. depicted by a text (Franklin and Tversky 1990; Mani and Johnson-Laird 1982; Johnson-Laird 1983; McNamara and Sternberg 1983; Morrow *et al.* 1989; and Tversky 1990). What it means for a mental model to be a "structural analog" is that it embodies a representation of the spatial and temporal relationships between, and the causal structure connecting, the events and entities of the narrative. In constructing and updating a representation, the reader would call upon a combination of conceptual and real-world knowledge and would employ the tacit and recursive inferencing mechanisms of her cognitive apparatus.

That the situation is represented by a mental model rather than by an argument in terms of propositions is thought to facilitate inferencing. We can actually generate conclusions without having to carry out the extensive computations needed to process the same amount of background information propositionally. The conclusions drawn are limited to those that are directly relevant to the situation depicted. The ease with which one can make inferences in such simulative reasoning has suggested to some that mechanisms either used in - or similar to those used in - perception may be involved. If we do employ perception-like mechanisms here many inferences would be immediate.

To date, most empirical investigations of the hypothesis have focused on the representation of spatial information by means of mental models. The main disagreement has been over whether the representation involves a perception-like image or is "spatial", i.e., allows different perspectives and differential access to locations (See, Tversky 1990). It is not clear that this debate has any bearing on the function of thought experiments. Far more important is the question of how knowledge and inferencing mechanisms are employed in running and revising the simulation, and on this subject there is nothing beyond some explorations of how implicit knowledge of causal relations influences updating a model (Morrow *et al.* 1989).

Returning to thought experiments, the proposal offered here is that the cognitive function of the

narrative form of presentation of a thought experiment to others is to guide the construction of a structural analog of the prototypical situation depicted in it. Over the course of the narrative, we are led to run a simulation that unfolds the events and processes by constructing, isolating, and manipulating the pertinent dimensions of the analog phenomena. The constructed situation inherits empirical force by being abstracted from both our experiences and activities in, and our knowledge, conceptualizations, and assumptions of, the world. In running the experiment, we make use of inferencing mechanisms, existing representations, and general world knowledge to make realistic transformations from one possible physical state to the next. In this way, the data that derive from a thought experiment, while constructed in the mind, are empirical consequences that at the same time pinpoint the locus of the needed representational change.

Characterizing thought experiments in this way also provides new insight into limiting case analysis, which has figured prominently in conceptual change in science as well. This form of idealization can be construed as a species of thought experiment. In this species the simulation consists of abstracting specific physical dimensions to create an idealized representation, such as of a point particle falling in a vacuum. The isolation of the physical system in thought allows us to manipulate variables beyond what is physically possible. Just what dimensions produce the variation and how to extrapolate from these may be something we determine initially in real-world experimentation, but the last step can only be made in the imagination. In physics, it is the idealized representation that is quantifiable. However, the idealized representation is rooted in and relevant to the real world because it has been created by controlled extrapolation from it. We get from imagination to application to the real world by adding in some of the dimensions we have abstracted, again in a controlled process.

Conclusion

Thought experiments play a central role in conceptual change in science. In constructing and "running" the experiment, a scientist is able to demonstrate how and why specific concepts are not applicable to the world. This understanding forms the basis of problem-solving efforts to construct an empirically adequate conceptualization. However the initial thought experiment is constructed and run, it is conveyed to others in the form of a narrative. I have proposed that we understand thought experiments as a species of simulative reasoning and that we draw upon research into the construction and running of mental models during narrative comprehension to understand how they can function to create change. In linking the conceptual and the experiential dimensions of human cognitive processing, the mental models analysis offers the possibility of explaining how thought experiments demonstrate the undesirable real-world consequences of a conceptualization, thereby compelling change. To develop and test this hypothesis, it will be necessary to have a better understanding of the processes through which mental simulations are "run".

There are further implications of this proposal for cognitive science. Historical cases provide valuable data on the processes through which individual scientists change their representations of nature (Nersessian 1991a). If - as a number of cognitive psychologists are claiming - the processes of conceptual change in cognitive development and in learning are indeed like those of major scientific revolutions, the potential role of thought experiments in creating these kinds of conceptual change should be explored (See, Nersessian 1991b).

References

Brown, J.R. 1986. "Thought Experiments Since the Scientific Revolution." *International Studies in the Philosophy of Science* 1: 1-15.

Clavelin, M. 1968. *The Natural Philosophy of Galileo: Essay on the Origins and Formation of Classical Mechanics*. Cambridge, MA: MIT Press, 1974.

Drake, S. 1973. "Galileo's Experimental Confirmation of Horizontal Inertia: Unpublished Manuscripts." *Isis* 64: 291-305.

Einstein, A. 1905. "On the Electrodynamics of Moving Bodies." in *The Theory of Relativity*, New York: Dover, 1952.

Einstein, A. 1917. *Relativity: The Special and the General Theory*. London: Methuen, 1977.

Einstein, A. 1946. "Autobiographical Notes." in *Albert Einstein: Philosopher - Scientist*. (ed.) P.A. Schilpp. Evanston, Il.: Open Court, 1949.

Franklin, N. and Tversky, B. 1990. "Searching Imagined Environments." *Journal of Experimental Psychology* 119: 63-76.

Galilei, G. 1638. *Two New Sciences*. (trans.) S. Drake. Madison: University of Wisconsin Press,1974.

Gooding, D. 1990. *Experiment and the Making of Meaning: Human Agency in Scientific Observation and Experiment*. Dordrecht: Kluwer Academic Publishers.

Johnson-Laird, P.N. 1983. *Mental Models*. Cambridge: Harvard University Press.

Koyré, A. 1939. *Galileo Studies*. Atlantic Highlands, N.J.: Humanities Press, 1979.

--------. 1968. *Metaphysics and Measurement*. Cambridge.MA: Harvard University Press.

Kuhn, T.S. 1964. "A Function for Thought Experiments." in *The Essential Tension: Selected Studies in Scientific Tradition and Change*. Chicago: University of Chicago Press, 1977.

Mani, K. and Johnson-Laird, P.N. 1982. "The Mental Representation of Spatial Descriptions." *Memory and Cognition* 10: 181-187.

McNamara, T.P. and Sternberg, R.J. 1983. "Mental Models of Word Meaning." *Journal of Verbal Learning and Verbal Behavior* 22: 449-474.

Morrow, D.G., Bower, G.H., and Greenspan, S.L. 1989. "Updating Situation Models during Narrative Comprehension." *Journal of Memory and Language* 28: 292-312.

Naylor, R. 1976. "Galileo: Real Experiment and Didactic Demonstration." *Isis* 67: 398-419.

Nersessian, N. J. 1991a "How Do Scientists Think? Capturing the Dynamics of Conceptual Change in Science." in *Cognitive Models of Science*, (ed.) R. Giere. *Minnesota Studies in the Philosophy of Science 15*. Minneapolis: U. of Minnesota Press.

Nersessian, N.J. 1991b. "Constructing and Instructing: The Role of 'Abstraction Techniques' in Developing and Teaching Scientific Theories." in *Philosophy of Science, Cognitive Science, and Educational Theory and Practice*, (eds.) R. Duschl and R. Hamilton. Albany: SUNY Press.

Norton, J. 1986. "Thought Experiments in Einstein's Work." Unpublished manuscript.

Qin, P. and Simon, H.A. 1990. "Laboratory Replication of Scientific Discovery Processes." *Cognitive Science* 14: 281-308.

Settle, T. 1961. "An Experiment in the History of Science." *Science* 133: 19-23.

Tversky, B. 1990. "Induced Pictorial Representations." Unpublished manuscript.

Wertheimer, M. 1945. *Productive Thinking*. New York: Harper.