In this paper I claim that thought experiments are examples of scientific modelling. Whilst I am not the first to suggest such an interpretation of thought experiments (Cooper 2005, Miščević 1992, and Nersessian 1992), the few philosophers of science who have taken on such a position have either limited themselves to the use of modelling in the service of explaining how thought experiments work, or have restricted unnecessarily the implication of a model-based account of thought experiments to the domain of cognitive science. This paper seeks to extend this early work on thought experiments by expanding its scope to include a discussion about: firstly, the nature of scientific modelling and theory generation; secondly, the role of analogic reasoning and metaphor in science and thought experimentation, and finally the place of thought experiments in science education. These steps towards both extending and unifying our theoretical understanding of thought experimentation is necessary if the extensive research in this established field of philosophy of science is to have significant impacts on thought experiment research in an educational context. It is worth making my agenda explicit: my aim is to develop a sound theoretical basis for the future implementation of thought-experimentation-based teaching and learning strategies into mainstream science education.

Introduction

Much of the momentum enjoyed by the thought experiment revival in the early 1990s was generated by the ongoing polemic between James Brown and John Norton (Brown 2004, Norton 2004a). While John Norton put forward the idea that thought experiments were arguments in disguise and nothing more (Norton 1991, 2004b), Brown suggested that thought experiments worked by steering the experimenter towards intuiting some platonic realm in which the laws of nature were manifestly self-evident (Brown 1991). As philosophers posited their own theories in opposition to the most striking limitations of these interpretations, less extreme positions on thought experiments began to emerge; ones that awarded greater significance to the agency of the thought experimenter in constructing their own meaning. As Gendler puts it: ‘…thought experiments rely on a certain sort of constructive participation on the part of the reader, and that the justificatory force of the thought experiment actually comes from the fact that it calls upon the reader to perform what I will call an experiment-in-thought’ (Gendler 1998: p 413). Gendler takes experiments-in-thought to mean ‘actual experiments’ performed by actual persons in real time.
In light of Gendler’s insight (see also Gendler 2004) it would be difficult to articulate thought experimentation simply in terms of a passive reception of the empirical claims that the originator of the thought experiment intended to convey. As in the interpretation of any text, a thought experiment requires action on the part of the thought experimenter. Put another way, we are called upon to avoid the often tacit assumption that there is a true experimenter-independent thought experiment out there; presumably an ideal form of the experiment that was developed and publicized by the original author in her efforts to support her scientific or philosophical claims. I believe that such an assumption operates in Norton and Brown’s respective interpretations of thought experiments. Taking their treatment of Galileo’s tethered spheres thought experiment as an example (Brown 2004, Norton 2004a), Norton’s reconstruction seems to suggest that Galileo’s claim that all bodies fall at the same rate (irrespective of their masses) is just sitting there ready to be unveiled by a formal chain of argumentation, whilst Brown’s platonic account presupposes a experimenter-independent reality built into every experiment.

Before leaving this introductory discussion on the polemic that surrounds thought experiments in the literature, a number of points need to be made. Undervaluing the active participation of the experimenter or assuming that the true content of thought experiments is somehow independent of them, is an error on a par with underestimating the social, cultural and historical forces that shape a thought experiment long before the instant when we are expected to conduct (or construct) the experiment using our own imaginations. These forces should be considered in all the varied manifestations of thought experiments. The first manifestation – namely that of thought experimentation as an active construction of meaning on the part of the experimenter – has already been alluded to in Gendler’s work and shall form the basis of my argument about the utility of thought experiments later in the paper. The others manifestations involve, respectively: an original, historical phase of thought experiment construction, which is often tentative and speculative; and a phase in which refined thought experiments are used to defend established theories.

Historically thought experiments begin with a real person making some scientific or philosophical claim. In the absence of suitable sense-extending apparatus, supporting her claims with the results of physical experimentation may be impossible. Yet she still has at her disposal the option of supporting her claims with results drawn from experiments-in-thought. Indeed, such a deficit of empirical data may be seen as synonymous with the absence of an appropriate language with which to describe and explain the underlying mechanisms of the phenomena at hand. The thought experiment in science is more than an experiment-in-thought; it is a practice in which an extraordinary language is fashioned from ordinary language with the express purpose of explaining or describing an extraordinary phenomenon in familiar terms. Thus, thought experiments usefully conscript metaphor in the service of meaning-making.

Like a pendulum, a thought experiment swings one way to push out the boundaries of what is known and on its return swing pushes against those who would dare question the scientific community’s current claims to knowledge. Thought experiments are as much rhetorical devices as they are forms of knowledge generation. This is evident when
thought experiments are presented to an audience with the aim of buttressing a consensus view – as is typical in undergraduate physics education for instance. In this context the thought experiment is presented not as an exploration of possible worlds but rather a didactic and folkloristic account of well-established scientific theories. We are not expected to get to the end of a presentation of one of Galileo’s thought experiments having missed the scientific moral of the story. Just as we tell our children the story of *Goldilocks and the Three Bears* to impress upon them an understanding of the property rights we accept as a social norms, so scientists use thought experiments in this manner to normalize scientific theory. The telling of thought experiments is a social mechanism by which theories are beatified.

Two consequences follow from this manifestation of thought experiments. Firstly, educators only utilize those thought experiments that support consensus scientific positions. As thought experiments are told and re-told, their message becomes more refined, literal, and increasingly aligned with current orthodoxy. It is not surprising that the only historical thought experiments to survive are those deemed successful. In effect, this robs ‘failed’ thought experiments the recognition they deserve in the advancement of scientific understanding. Secondly, it has the effect of demoting the contribution of the everyday thought experiments, the ‘what if’ scenarios that we use to plan future actions like catching a bus or planning a holiday. Thirdly, it lays a trap for students who would think that the content of thought experiments is fixed for all time and objective.

A model based account of thought experimentation allows us to unify the various manifestations of thought experiments into a single framework. However, since that framework itself is tied to the nature of science itself, a sketch of scientific model making and theory construction is necessary to the development of my argument.

**Science and Modelling**

It should be said from the outset that the long tradition of philosophy of science has to date failed to yield a consensus view on the nature of science: its ontology, its claims to knowledge (both theoretical and empirical), and of course its practices and methodologies. In the context of science education research it would be just as unwise to assume any greater certainty about our understanding about the nature of science in schools. Nevertheless, few would dispute the special place of modelling in the natural sciences. Consequently, I have taken as my starting point the realist account of science developed by Rom Harré, since he sees modelling as one of the primary tools by which scientists describe and explain phenomena in the world (Harré 2002, 2004).

According to Harré, science explores and expands the human *umwelt*; the environment to which humans have access. The human *umwelt*, in turn, dynamically constitutes the site from which the significance of scientific practices is drawn since it delimits what is afforded us and helps define what we are disposed to understanding. If we limit the human *umwelt* to include those entities only immediately accessible to naked sense perception (as the positivists would have us) then the scientific practices to which we are disposed – the world of possibilities available to us – would be significantly
impoverished. For what distinguishes the human *umwelt* from those of other species is our capacity to use and develop sense-extending instruments. Our *umwelt* is expanded with the development of every new tool. The moons of Jupiter, various interpretations the Copernican model of the heavens, the practice of telescopic Astronomy, all entered the human *umwelt* through Galileo’s peculiar use of the newly invented telescope.

The scientific extension of the human *umwelt* is achieved as much by the process of conception as it is by perception. Scientific practices presuppose the existence of entities that inhabit the imaginative space of human beings. For some of these entities, like the gravitational field, the prospect of there being instruments with which to see them directly is impossible. Others, like gravitational waves, may soon enter the realm of the observable with the imminent arrival of sufficient instrumental sensitivity. The fluidity of the internal and external boundaries of the human *umwelt* underscores the dynamic and ephemeral nature of scientists’ claims to knowledge and their assumption of what is taken to exist. It should be stressed that scientists cannot afford to give preference to thought over perception or *vice versa*. The nature of science – contrary to its conventional presentation – has the two alloyed together much like the tin and copper in bronze. Thought experiments, likewise, while characteristically about the manipulation of entities in the human imagination have the products of our senses woven into them.

Models are the primary tools by which scientists delimit the human *umwelt* and systematically describe and explore the relationships between the entities within it. Models are iconic representations of systems or processes in the real world. Although they may be material or imaginary things, they only count as models because of the particular way they relate to other entities in the world and because, as representations, they are constructed to serve some clearly defined purpose. So, a toy car serves as a model car not only because we relate some of its salient features to those of the real thing, but also because we do so with the purpose of studying say, the flow of air across its surface in the model world in order to ascertain (by analogy) how air might flow across the surface of a car in the real world.

The catalogue of models used in science bifurcate easily when we consider the distinction between what a model is *of*, the *subject* of the model, and what it is modelled *on*, the *source* of the model. The first branch comprises homeomorphic models in which the source and subject are identical. These models are directed towards an analysis or description of aspects of the perceptible world. By way of example, a 1:50 scale model *of* a house that is also modelled *on* a real house affords an analysis or description of the effects of light say, which on the building itself would prove too difficult. At one level the model house functions as an idealization of the real world house; it shares certain features with ordinary houses (such as a roof, walls, windows, flooring, etc.) but these features as they appear in the model house are deemed superior according to some standard. At another level the model house and the real house are related through abstraction. That is, for the sake of simplifying the analysis only certain properties or features of the real house are transferred to the model. So, there is no point in having 1:50 replicas of brass taps in a model house built to determine the lengths of shadows cast through the seasons. Finally, the model may be used to typify a class. For example, a
typical house modelled for town planning purposes may have 2.35 rooms, even though having 2.35 rooms makes no sense for a particular member of the type.

While *homeomorphic* models are used to simplify observed phenomena, *paramorphic* models are used to produce hypotheses about unobservable processes or generative mechanisms with the aim of explaining observed phenomena. These models rely on the source of the model differing from its subject, or more precisely, drawing upon a familiar or explicable set of phenomena as a source to make a representation of a subject whose underlying mechanism is not known. When Niels Bohr attempted to develop a model of the atom which would explain the distribution of spectral lines emitted by different elements, he took as his source a model that had existed for close to four centuries: the heliocentric model of the solar system. He chose the solar system because its salient features already had a large degree of resemblance to those of the atom (e.g. energy-distance relation), and it was hoped that the generative mechanism for the phenomena observed in the solar system could also serve as the unknown source of atomic phenomena. The efficacy of science stems from its judicious use of what is already known.

Of course the elements of what is known must be organized in a way that would take into account the relationships between them. In science this is achieved with a taxonomy or system of classification by which knowledge is arranged according to class, type and kind. In particular, a type-hierarchical taxonomy is an example which facilitates an explanation of modelling and thought experimentation. In a type-hierarchical taxonomy types are established for the entities extant in the human *umwelt* and arranged from the most general at the top level (supertype: *living thing*) to the most specific at the bottom level (token: *my cat ‘felix’*). A striking feature of a type-hierarchical taxonomy is that it ‘stores knowledge vertically, in the inheritance relation. To discover what is presupposed about a lower type one runs up the hierarchy through the nodes to the apex. Thus the species ‘cat’ is vertebrate, animal, living thing’ (Harré, 2002: p 41). Since there are an increasing number of types at lower levels of the hierarchy and multiple inheritance relations in the vertical direction, visual representations of a type-hierarchy resemble an inverted tree. With this analogy in mind we would consider scientific theories to be everything connected to a large branch within the structure along with its accompanying ‘branchlets’, ‘twigs’, and individual ‘leaves’: i.e. theories are chunks of the type-hierarchy.

Under this representation, certain chunks of the type-hierarchy serve as the model, others as the model’s subject and others as the model’s source. The process of modelling is a means by which likenesses and differences between the model, its subject and/or its source are brought to the foreground and is, therefore, an analogic mode of reasoning. More importantly ‘the use of analogy presupposes that model, source and subject are subtypes of the same supertype within a type hierarchy. They are related to one another via the inheritance relation’ (Harré, 2002: p 54). It is the inheritance relation that allows us to explain why some models are better than others in much the same way that it explains why we can excise the infinite number of trivial and negative analogies that arise when only comparing entities using structural isomorphism (Aronson, Harré & Way
For instance, consider the Bohr planetary model of the atom and its rival Thomson plum-pudding model. Only the former has a source (the solar system) that is connected to its subject (the atom) through a common supertype: namely, the type of systems with quantized energy levels.

The analogies or models that are available to the scientist are dependent on extant type-hierarchies that in turn must be determined empirically since the properties of any type within the hierarchy can change places between definitions and accidental attributes. Type-hierarchies are, therefore, dynamic representations of the human *umwelt*. In this picture, scientific theories are taken as segments of a type-hierarchy, with the inheritance relation doing all the work in ensuring the internal consistency characteristic of successful theories. This is the kind of consistency that Cooper (2005) identifies as an intrinsic feature of successful thought experiments. Theories, as sub-sets of type-hierarchies, can emerge or change as new empirical data makes it necessary to form new connections between types within the hierarchy.

**Thought Experiments as Models**

I shall use two examples of classroom thought experiments to illustrate how the proceeding framework for scientific modelling may be applied. The objective is to demonstrate firstly, how the most prominent theories put forward in the philosophy of science for thought experiments can be accounted for under a single unified interpretation, and secondly, how such a novel application of the proceeding theoretical framework results in a natural bridge between the philosophical and educational research contexts.

**Case Study One:**

A 24-year-old, female, pre-service, fourth-year Bachelor of Education (Primary) student was preparing a 1-hour teaching session for a group of five, grade five and six students following a recent total lunar eclipse. Her internet research on the subject lead her to the following quotation: “From the moon, during a [lunar] eclipse, you would see all the world’s sunsets at once”. Her initial impression was that this quotation could be used to explain to children why the moon was reddened during a total lunar eclipse. Why she thought the quotation carried such explanatory force she could not express, say that the colour of the eclipsed moon bore a striking resemblance to that of the sky at dusk. Moreover, it was the fact that objects such as buildings, people, mountains and clouds took on a reddish appearance during sunset that spurred her imagination.

I asked her to conduct the following thought experiment:

Imagine you are at a beach with your best friend. She is facing west towards the sun (which has just set) and you are facing her. Her face appears red. Let us assume for the moment that we know why her face appears red. Your friend takes a few steps backwards; her face still appears red. If she were to take a few more steps backwards her face would still remain reddened. Imagine now that instead of taking
small steps she made one giant leap into space out to a distance as far as the moon. So long as she travels in a direction such that the sun appears to her to have just set, then her face would remain illuminated – albeit faintly – by the reddish glow of the sky. Now imagine instead placing the moon at your friend’s location in space. It too would appear red.

“Ahh, I see! An eclipse is just a sunset. The sun goes behind the earth at sunset and that’s what happens during an eclipse” was her response to the thought experiment.

Case Study Two:

A small group of year-9 students were asked to discuss the possible answers to the following question: why is the sky blue. The following dialogue took place:

Amber: The sky is blue because it reflects the ocean, which is blue.

Stephanie: But what if the sky was really a mirror…what would you see?

Amber: Well, I would see my reflection…no, that’s too small, I would see the ground. But that would mean the sky would be brown on top!

Stephanie: Yeah, it can’t be that.

A model-based account of classroom thought experiments:

In the first case study the student is searching for a mechanism that explains why the moon is reddened during a total lunar eclipse. Her attempts to find such a casual mechanism begin, not with a theory, but rather with an analogy: the analogy between the reddening of the sky and other objects at dusk and the reddish appearance of the moon during an eclipse. This is a pre-verbal, aesthetic response in the affective domain of experience – yet it remains at the heart of scientific theorizing.

In order to develop an explanatory model for the astronomical phenomenon in question the student had to establish two things. Firstly, she had to select a source for her explanatory model that was (i) different from the subject of the model (in this case the reddening of the moon during an eclipse) and (ii) well understood. The analogy with sunset provided such a source. Secondly, she had to determine whether there was a supertype common to both the subject and the source. At first the only connection she could draw between the two phenomena was based on the coincidence of a secondary property; namely, colour. However, this explained nothing since the property “colour” served only to define inclusion of phenomena within a type, say the type of things that appear red. It was not until she had performed the thought experiment that a suitable candidate supertype emerged.

What the student had expressed so eloquently in her conclusion was that sunsets and eclipses fall under a common supertype; a supertype characterised by (i) the passage of
the earth in front of the sun, and (ii) the transmission of red light to objects along particular paths-of-alignment between the sun, the earth and the earth’s atmosphere. The thought experiment worked for her because it enabled her to see that a human being on a beach was also an object in space and that a sunset seen from the moon was the same as a sunset witnessed from a beach on earth. While she will undoubtedly go through life still recognising sunsets as sunsets and eclipses as eclipses (at the token level of her type-hierarchy), she now has at her disposal a new model that explains the generalised phenomena of “eclipse-sunsets” by connecting the theories (chunks of the type-hierarchy) associated with each respective phenomenon.

Like the first case study, case study two illustrates how thought experiments can be useful in examining explanatory models by taking analogies literally and experimenting with the consequences in the laboratory of the mind. This is what Stephanie demands of Amber when she poses the question “…what would you see?” In this case the thought experiment leads to the rejection of the “mirror” model used to explain the blueness of the sky. Amber is able to recognise the absurdity of the hypothesis because the thought experiment forces her to compare the respective theories associated with plane mirrors and the appearance of the atmosphere on mass. Students and teachers operating at the token level of the type-hierarchy will easily find individual points of resemblance or disparity between entities, but these do not constitute explanatory models. Explanation can only be derived from the unification of large chunks of the type-hierarchy from higher levels down.

If this is the case then Gendler’s ‘constructive participation’ can be taken to operate on a number of levels. The first participatory construction involves the development of a personal ontology or type-hierarchy: chunks of which explain such phenomena as sunsets or mirror images. Experimentation – particularly through personal experience and the use of novel or increasingly sensitive instruments – increases the number of ontological entities (expands the individual umwelt) and provides the means by which membership of types and kind is determined. At another level analytic modelling and in particular explanatory modelling – including thought experimentation – enables the construction of novel connections amongst the chunks of the type–hierarchy. The scientific enterprise amounts to nothing more than the collective construction of empirically justifiable connections within a commonly accepted/generated type–hierarchy. That a student’s ontology and the connections within it do not conform to that of the scientific community is a question of scale not of methodology. In both cases the construction of the ontological framework is a dynamic one as new scientific- or folk-theories necessitate the emergence of new nodes and links in the type-hierarchy. The plurality and dynamism of type-hierarchies have important consequences for the educational use of thought experimentation. For instance, when historical thought experiments such as Galileo’s or Einstein’s are presented in the classroom, their success will be determined largely by the degree of resemblance between a student’s ontology and that of the scientific community endorsing the consensus model in question. Conceptual change is most likely when there is a shared set of types.
That the second case study appears on the surface to be a reductio ad absurdum argument seems also to add weight to John Norton’s argument-based account of thought experimentation. Like many other examples provided by Norton this thought experiment could be described as a reductio argument. However, describing a thought experiment as an argument does not imply that a thought experiment is reducible to an argument. The argument-like properties of thought experiments are derived from the inheritance relation that functions in all type-hierarchies. Since the properties of types on the lowest levels of the type-hierarchy presuppose those immediately above them, and these in turn, presuppose those above them, a vertical “thread” through the type-hierarchy has the form of a deductive argument in which the tokens act as the ‘conclusions’ that necessarily follow from the ‘premises’ established by the types and supertypes above them.

If we assume the rhetorical force of a thought experiment is derived by its logico-deductive structure then it becomes apparent why their use in education has frequently taken the form of an argument rather than a dialogue. Recall that a thought experiment (as a model) relies on a comparison of large chunks of a type-hierarchy rather than vertical strands within it. Thought experiments, unlike deductive arguments, rely much more on lateral ‘propositions’: they are comparison of networks of relations-amongst-things. If historical thought experiments are effective in converting students to a consensus model then it is likely to be in virtue of the underlying ontology of the model being somehow pre-established rather than generated ab initio. If we allow ourselves to over-interpret this ontology we might well convince ourselves, as Brown does, that what is pre-established can pass as what is a priori. There are no new entities generated in case study one, only new explanatory models.

Finally, there is another means by which thought experiments may lead to conceptual change without the necessity for new empirical data. Since theories are simply chunks of a type–hierarchy a new theory may be generated not only by partitioning new real estate in the ontological landscape but also by re-organizing the types within an existing theory. This amounts to substituting ‘literal’ inheritance relations with ‘metaphoric’ ones. Supertypes that ordinarily connected a number of subtypes are forced to make connections between new ones. This process is metaphorical in the sense that the re-connection leads to counterfactuals of the form: A is not B; if A were B then X would follow (eg the sky is not a mirror; if the sky were a mirror then we should see the image of the ground reflected in it). It is precisely in their ability to generate counterfactuals that we find the real value of thought experiments.

**Conclusion**

What I have demonstrated is that thought experiments are well described in terms of the structure and function of scientific modelling. The figurative language they employ and their reliance on analogy stems directly from the model-making and theory-making processes they exemplify. Thought experiments are useful to scientists and students of science because they fulfil the need to explore and expand that region of the human **umwelt** only accessible through the human imagination. Whilst such experiments-in-thought do not result in a prior knowledge, they do allow for a personal re-interpretation
of empirically derived knowledge that is arranged hierarchically according to type in an inheritance relation. Since the internal organization of these underlying type-hierarchies are subject to change in the light of new empirical data thought experiments are necessarily dynamic and not reducible to fixed arguments. As examples of scientific thinking they are a useful tool for teaching students about the nature of science.

References


