Thought Experiments can be Harmful

R. I. Damper,

Image, Speech and Intelligent Systems Research Group,
Department of Electronics and Computer Science,
University of Southampton,
Southampton SO17 1BJ, UK.

Abstract for Conference on Model-Based Reasoning (MBR'01), May 17–19, Pavia, Italy

"One real experiment is worth far more that a half century of debate about the meaning of a thought experiment." (Gribbin 1984)

"Experience teaches us, however, that there is no such thing as a thought experiment so clearly presented that no philosopher can misinterpret it." (Dennett 1995)

This paper presents an argument in three parts:

1. Thought experiments are a necessary and inescapable part of the process of theory formation and scientific discovery.
2. Thought experiments can be helpful in pinpointing and illuminating key theoretical issues in a discipline. At best, when a critical level of development is reached, they may even trigger revolutionary advances.
3. In other disciplines, where the theoretical underpinnings are less developed or absent, the value of thought experiments is uncertain. It is possible that they could be misleading—even harmful.

So how can we tell if a given thought experiment is in category 2 or 3? I will give examples of various (mostly rather famous) thought experiments, and attempt to identify the features which mark them out as being either helpful or harmful.

1. Thought experiments are necessary. The number of questions which can be posed, from a scientific perspective, about the way that nature works is obviously infinite, so we cannot hope to answer more than a fraction of them by practical experiment. Also, certain disciplines such as evolutionary biology and economics do not lend themselves to experimentation. Although computer simulation can play a part (Casti 1997), it remains the case that we cannot avoid frequent recourse to 'thinking our way through' a problem, i.e., to thought experiment.

Much of the material taught to us as fact during our early scientific education has effectively to be taken on trust: Neither the time nor the resources are there to verify everything for ourselves by practical experiment. We accept and learn (mostly) what we are taught because we believe implicitly in (to borrow the terms of Kuhn 1962) the 'paradigm' which defines 'normal science', because we are attracted to the apparent rigour, rationality and predictive power of the paradigm, and because this acceptance is the price of entry into the profession and community of science. Once our knowledge and understanding of our discipline reaches a sufficient level, however, we no longer need such heavy reliance on trust. Thanks to the predictive power of scientific theories, it becomes possible to answer novel questions by thought experiment.

But this way lies danger! Before Galileo, the result of the following thought experiment would have been 'obvious' (but wrong) to any thinking person: Drop separate light and heavy masses together from a height (in a vacuum) and observe which accelerates fastest. Clearly, the majority answer will be relative to the current state of knowledge, but the simple fact of being in a majority is no guarantee of being right. Yet a thought experiment relies on obtaining a consensus answer (albeit from a select group of specialists) if it is to offer something more than paradox.

The discussion of Kuhn (1962, pp. 99–101) on the relation of Einsteinian to Newtonian mechanics bears on this point. The latter is an excellent theory within its range of validity—that of low relative velocities—as its widespread, everyday use in modern engineering attests. Incorrect predictions of Newtonian mechanics arise from applying it outside of that range, which "… must be restricted to those phenomena and to that precision of

observation with which the experimental evidence in hand already deals" (p. 100). But acceptance of this maxim rules out thought experiments which, as we have seen, are necessary. As Kuhn (1962, p. 101) says: "Is it really any wonder that the price of significant scientific advance is a commitment that runs the risk of being wrong?" Thus, the same difficulties arise as with other modes of reasoning such as abduction and inductive inference.

2. Thought experiments can be helpful. In a number of justly famous and brilliant thought experiments *circa* 1905, Albert Einstein revolutionised the whole basis of modern physics (see Jammer 1966, 1974 for comprehensive details). Partly, his extensive use of Gedankenexperiments was a 'forced move', reflected the (then) technical difficulty of performing practical experiments with elementary particles or bodies moving at relativistic velocities. But an additional factor was that this *modus operandi* perfectly matched Einstein's incisive ability to identify and illuminate the nub of the matter. However, the technical difficulties of performing practical experiments have slowly receded. One of the more exciting stories in science is surely Sir Arthur Eddington's expedition to the 1919 solar eclipse in West Africa during which the bending of starlight by the sun's gravitational field was observed, so confirming Einstein's predictions. But his predictions were not always so triumphally right! In 1935, Einstein, Podolsky, and Rosen described the famous EPR thought experiment, in which two particles interact and then fly apart, retaining some 'imprint' of the interaction. At a later time, a measurement is made of the state of one particle which has implications for the state of the other. Space does not allow us to develop the important theoretical point at issue here (see Penrose 1989, pp. 361–369 for an accessible description) except to say that it bears on the influence one particle can have on the other once they are no longer in the same locality.

Einstein and Bohr disagreed fundamentally on the outcome with Einstein arguing that the intuitively correct ('common sense') answer was at odds with quantum mechanics. The latter seemed to require that one particle must somehow 'know' about the state of the other which Einstein maintained was contradictory. Accordingly, quantum mechanical description could not be considered complete. A brilliant experiment by Aspect and his colleagues (Aspect, Grangier, and Roger 1982; Aspect, Dalibard, and Roger 1982) has largely resolved the issue in favour of Bohr. Although it would be a mistake to see this as the last word on the matter, there was indeed an 'entanglement' of states even though the particles were far from each other's locality.

So what do we learn? First, Gedankenexperiments can be blindingly insightful as in the majority of Einstein's predictions. Second, a thought experiment which gives birth to a paradox (cf. the EPR paradox) can be nonetheless useful, laying down an agenda for subsequent work aimed at resolving it.

Before continuing, we should note that Turing's seminal (1936) paper is properly a thought experiment. He effectively invented a (virtual) digital computer— the Turing machine—so allowing him to solve a problem in mathematical logic. In this sense, his thought experiment was a trigger to subsequent revolutionary advances in computer technology. Helpful indeed!

3. Thought experiments can be harmful. Thought experiments have been a popular investigative device in artificial intelligence, cognitive science and philosophy of mind, where the theoretical underpinnings are nowhere near as well developed as in physics. In AI, the classic thought experiment is Searle's (1980) Chinese Room argument (CRA): a computer program intended to 'understand Chinese' would not really do so because Searle himself could manually execute the same algorithmic steps while understanding nothing of Chinese. The argument has, of course, been thoroughly well debated (e.g., Harnad 1989; Penrose 1989; Copeland 1993; Boden 1994; Franklin 1995; Bishop and Preston; forthcoming, and the peer commentary appearing with the original article), yet it is surprising how few commentators remark on the practicality of doing what Searle proposes. An exception is Copeland (1993, p. 127) who writes of "the built-in absurdity of Searle's scenario". What Searle and others seem ready blithely to assume—the existence of a Chinese 'understanding' program able to pass the Turing test (Turing 1950)—is so far beyond the current capabilities of AI and computer technology as to amount to science fiction. What could we possibly learn from such a fanciful conception? There is no realistic way of resolving any paradoxes which arise, save appeals to common sense, and we know from the example of quantum mechanics how fallible this is.

One can conceive of two (at least) possible rejoinders. It could be said that Einstein's Gedankenexperiments were similarly fanciful: no one could chase after a light beam at the speed of light! Yet experimental tests of Einstein's predictions were on the verge of being practical—by observing binary stars, eclipses of the sun, etc. So there seems to be a matter of degree here. Another point of view might be that it is too early to pronounce on the CRA: in time, Searle's predictions might be proved (more or less) right or wrong by empirical means. My own feeling is that this will not happen: the proposed scenario is just too far from practical, experimental test. But perhaps some good can come out of the CRA if we substitute a task closer to the capabilities of current computer

programs than understanding Chinese. This direction was first explored by Puccetti (1980), who substituted the chess room for the Chinese room, although to my mind he did not press the point home.

Searle's CRA was chosen here for illustration, but there is no shortage of wildly implausible thought experiments in cognitive science and the philosophy of mind. One might mention the Twin Earth argument of Putnam (1975)—see Lloyd (1989) and Kim (1998) for discussion—which relies on confusing your earthly conception of some object with its apparently identical (but subtly different) counterpart in a twin world. Here, Dennett (1995, pp. 410–411) lays the argument bare by presenting "a more realistic example" which "could be" true, involving cats and Siamese cats. Next on my list is the thought experiment that actually convinced me that a paper such as this one was necessary. Dietrich (1989), in developing his argument that computational states involve content (semantics) as well as merely formal manipulation (syntax), writes: "Imagine that I had an exact duplicate made of me yesterday" (p. 123). Well, yes, imagine.

Finally, to negate the impression that thought experiments could never be of any great value in this area, I offer Braitenberg (1984) as a clear counter example. We have built 'vehicles' similar to those proposed in Braitenberg's series of thought experiments in synthetic psychology, with interesting results (Damper, French, and Scutt 2000). Here, of course, the value of Braitenberg's contribution lies in not departing too far (if at all) from what is practical. *[1600 words]*

References

Aspect, A., J. Dalibard, and G. Roger (1982). Experimental test of Bell inequalities using time-varying analyzers. *Physical Review Letters 49*(25), 1804–1807.

Aspect, A., P. Grangier, and G. Roger (1982). Experimental realization of Einstein-Podolsky-Rosen-Bohm Gedanken-experiment – A new violation of Bell inequalities. *Physical Review Letters 49*(2), 91–94.

Bishop, M. and J. Preston (Eds.) (forthcoming). *Essays on Searle's Chinese Room Argument*. Oxford, UK: Oxford University Press.

Boden, M. A. (1994). New breakthroughs or dead-ends? *Philosophical Transactions of the Royal Society of London (Series A) 349*, 1–13.

Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology*. Cambridge, MA: MIT Press.

Casti, J. L. (1997). *Would-Be Worlds: How Simulation is Changing the Frontiers of Science*. New York, NY: John Wiley.

Copeland, B. J. (1993). *Artificial Intelligence: A Philosophical Introduction*. Oxford, UK: Blackwell.

Damper, R. I., R. L. B. French, and T. W. Scutt (2000). ARBIB: an autonomous robot based on inspirations from biology. *Robotics and Autonomous Systems 31*(4), 247–274.

Dennett, D. C. (1995). *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. London, UK: Penguin.

Dietrich, E. (1989). Semantics and the computational paradigm in cognitive psychology. *Synthese 79*(1), 119–141.

Einstein, A., B. Podolsky, and N. Rosen (1935). Can quantum mechanical description be considered complete? *Physics Review 47*, 777–780.

Franklin, S. (1995). *Artificial Minds*. Cambridge, MA: Bradford Books/MIT Press.

Gribbin, J. (1984). *In Search of Schrödinger's Cat: Quantum Physics and Reality*. London, UK: Wildwood House.

Harnad, S. (1989). Minds, machines and Searle. *Journal of Experimental and Theoretical Artificial Intelligence 1*, 5–25.

Jammer, M. (1966). *The Conceptual Development of Quantum Mechanics*. New York, NY: McGraw-Hill.

Jammer, M. (1974). *The Philosophy of Quantum Mechanics*. London, UK: Wiley.

Kim, J. (1998). *Philosophy of Mind*. Oxford, UK: Westview/Perseus.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press. (Pagination refers to 1996 third edition).

Lloyd, D. (1989). *Simple Minds*. Cambridge, MA: Bradford Books/MIT Press.

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds and the Laws of Physics*. New York, NY: Oxford University Press.

Puccetti, R. (1980). The chess room: Further demythologizing of strong AI. *Behavioral and Brain Sciences 3*(3), 441–442. (Peer commentary on Searle, 1980).

Putnam, H. (1975). The meaning of 'meaning'. In K. Gunderson (Ed.), *Language, Mind and Knowledge*, pp. 131–193. Minneapolis, MN: University of Minnesota Press. (Vol. 7 of Minnesota Studies in the Philosophy of Science).

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences 3*(3), 417–457. (Including peer commentary).

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society (Series 2) 42*, 230–265.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind 59*, 433–460.