May 15th, 9:00 AM - May 17th, 5:00 PM

A Way to Describe and Evaluate Thought Experiments, or Trying to Get a Grip on Virtual Reality

Lawrence G. Souder
Temple University

A WAY TO DESCRIBE AND EVALUATE THOUGHT EXPERIMENTS, OR TRYING TO GET A GRIP ON VIRTUAL REALITY

Lawrence Souder
Center for Frontier Sciences
Temple University

Abstract:
The use of thought experiments seems to provoke much controversy, often in the form of charges of appeals to intuition. The notion of intuition, however, is vaguely defined in both the context of thought experiments and in philosophy in general. This vagueness suggests that the description of thought experiments is incomplete, and thus the prospect for their evaluation remains unfulfilled.

Previous analyses of thought experiments have come largely from philosophy where the focus has been on truth value and validity. But these approaches seem to view argument monologically; no accommodation of an audience response like intuition is possible. I try to show that van Eemeren and Grootendorst's pragma-dialectical model provides a framework for analyzing thought experiments and evaluating them because it treats thought experiments as part of a dialogue and as the result of a perspective.

***

Introduction

Thought experiments have been considered as a limiting case of experiment (Sorensen, 1992b). Moreover, in formal contexts they have been viewed as a way for readers to simulate certain conditions for testing some hypothesis (Nersessian, 1992b), and in less formal contexts they have provided an "opportunity to walk a mile in somebody else's shoes." (Steinem) This aspect of simulation suggests that the effectiveness of a thought experiment depends on how closely the one executed by the reader matches the one advanced by the writer. A thought experiment would be persuasive to the extent that it could be replicated by another, in the sense of Popper's "inter-subjectively testing." (Popper, p. 46)

In order to consider thought experiments as intersubjective phenomena, it is necessary to see them from the point of view of both the producer and the consumer. Such a requirement suggests a dialectical perspective where thought experiments are viewed as part of a dialogue. In what follows I try to show that thought experiments, whether scientific or philosophical, should be seen as a type of experiment and that their effectiveness depends on the intersubjective agreement about their conditions of execution. In order to describe and evaluate a given thought experiment, then, we need to see it as part of a dialogue. This requirement is also pragmatic because thought experiments are often found in dialectical pairs, where some original thought experiment has provoked an amended version that attempts to either reinforce or, more commonly, refute the original. This argument move is similar to what Perelman and Olbrechts-Tyteca call an amended analogy.

To this end I have tried to adapt and apply van Eemeren and Grootendorst's pragma-dialectical model of argumentation to the study of thought experiments. This model allows for seeing the dialectical tension between a given thought experiment and its variations. To illustrate the application of this model, I include a sample analysis of John Seale's Chinese Room thought experiment and several of its amended forms.

Thought experiments are experiments

In adapting the thought experiment to philosophical discussions, arguers often attempt to maintain a sense of experimental rigor. Some thought experimenters explicitly acknowledge the experimental nature of these examples. For example, Seymour in his criticism of Putnam's Twin Earth thought experiment says, "Simply by varying a certain object or property in the world, it is allegedly possible to vary the content of a subject's particular intentional state without a corresponding change in the subject." (p. 84) Vaihinger expresses a similar view about fictional examples:

> By varying the possibilities he imagines for this man left to his own devices, he is able to make a series of very subtle psychological observations. (p. 191)

Some philosophical thought experiments, in fact, are elaborations of actual experiments. Derek Parfit, for example, in his attempt to refute the necessity of the unity of consciousness, offers an invented example built around Sperry's experiments with commissurotomy.

> ...suppose that I have been equipped with some device that can block communication between my hemispheres. Since this device is connected to my eyebrows, it is under my control. By raising an eyebrow I can divide my mind. In each half of my divided mind I can then, by lowering an eyebrow, reunite my mind. This ability would have many uses. Consider [that] I am taking an exam, and have only fifteen minutes left in which to answer the last question. It occurs to me that there are two ways of tackling this question. I am unsure which is more likely to succeed. I therefore decide to divide my mind for ten minutes, to work in each half of my mind on one of the two calculations, and then to reunite my mind to write a fair copy of the best result. (p. 246)

Parfit says of such examples "[their] impossibility is merely technical." He believes that given enough time and resources such an experiment could be conducted.

Acknowledgments of the parallels between thought experiments and real experiments pervade the literature. Sorensen claims "the thought experiment is a limiting case of experiment." (p. 239) Nersessian focuses on the simulation of real events as the operative principle of thought experiments. For her a thought experiment is like a real one because it allows us to "draw conclusions about potential real-world situations we are not participating in at the time." (1992b, p. 291) Brown emphasizes, "The burden of any thought experiment rests on the establishment (in the imagination) of a phenomenon." (1986, p. 4) Even the conveyance of thought experiments seems modeled after the reports of real experiments. Shapin notes that Boyle's reports of his experiments were strongly narrative in order to allow readers to witness the experiment *in absentia*. Nersessian sees a parallel in thought experiments and calls the process "virtual witnessing." (1992b, p. 298)

The parallels between real experiments and thought experiments have been exploited, however, to privilege the former over the latter. Kuhn asks rhetorically on behalf of empiricists, "How, then, relying upon familiar data, can a thought experiment lead to new knowledge or to a new understanding of nature?" (1977, p. 241) The same parallels between real experiments and thought experiments have also been used to position scientific thought experiments over philosophical thought experiments. Wilkes claims, "Just as with real experiments, thought experiments presuppose that all relevant background conditions are included and specified." (p. 9) Wilkes believes that philosophical thought experiments should labor under the same constraints as do real experiments.

An experiment of any kind (real or imaginary, scientific or philosophical) needs to give the background conditions in order to clarify what factors are to be held constant and which vary. In her view philosophical thought experiments don't specify these background conditions to the extent that scientific thought experiments do, and so are problematic because their jumps from imaginary data to theory are too great.

But Wilkes' distinctions seem untenable. What constitutes relevant background conditions and whether they have been adequately specified are often uncertain even for scientific thought experiments. Wilkes herself acknowledges this uncertainty in the example she offers of the dispute between Bohr and Einstein over the latter's photon-in-the-suspended-box thought experiment. (p. 9) For this reason one must take a contextualist view of philosophical thought experiments. We cannot, as Wilkes does, reject them categorically as unreliable, for as Bunzl notes, "this notion of unreliability has to be understood relative to the goal of thought experiments as knowledge producing." (p. 227) But the goal of knowledge production cannot be assumed.

A Contextualist Approach to Analysis

Nevertheless, we can still make use of the criteria that Wilkes applies for rejecting philosophical thought experiments as a genre for the sake of evaluating them in context. Instead of asking whether a thought experiment is reliable, we can ask: to what extent do interlocutors agree about the context of the thought experiment; to what extent are the interlocutors' intuitions about the thought experiment's ordinary language terms congruent; and to what extent do the interlocutors agree about the establishment of a phenomenon. In general, we can ask to what extent are the interlocutors executing the same experiment in thought.

The emphases in a contextualist approach that are relevant to these questions are: (1) arguments are viewed in dialogue, (2) shifts in the dialogue are considered significant, (3) each side's account of its own perspective is integral to analysis. I propose that an adaptation of van Eemeren's, et al., pragma-dialectical model addresses these areas and is appropriate for the analysis of thought experiments.

While the pragma-dialectical model is intended for structured conversation, it seems applicable to an examination of amended thought experiments because of their dialogical character. Since I'm considering thought experiments in their original and amended versions, they can be considered pieces of a dialectical assembly and indeed constitutive of a dialogue. Van Eemeren, et al., themselves don't seem so rigid in their application of the model as to restrict themselves to conversations that are reproducible only via the notational system of conversation analysis. In fact, they assert explicitly at one point, "The model can provide a framework for interpreting and reconstructing the argumentative features of actual discourse, whether dialogic or monologic." (p. 34) I've accumulated over 100 variations on the Chinese Room thought experiment alone. Many of these (especially Dennett's and Churchland's) have left a trail of arguments and counterarguments to Searle's responses. This trail represents at least a virtual conversation, if not one in the ordinary sense.

The other tool to be added to the method proposed here is van Eemeren's, et al., technique of reconstructing reflexively reasoned accounts. In their analysis of an argument they attempt to describe each interlocutor's perspective of the issue, their understanding of their opponents perspective, and their understanding of the speech event they are participating in. Their analysis draws on their observations of the arguers' discourse and action, as well as interviews with the arguers. From this data Van Eemeren, et al., categorize the perspectives represented by the arguers and posit for each "a different dramatic structuring of the event." (p. 146) They then show how each arguer's perspective supplies a "subjective rationality" that brings coherence to each

interpretation of the speech event.

To illustrate this approach van Eemeren, et al., offer an analysis of the exchange between itinerant campus preachers and students. From their analysis of the speech event they derive two perspectives: the humanist and the fundamentalist. They describe the humanists's view of the exchange as "a confrontation over social policy," (p. 152) offering as evidence the humanists' use of the vocabulary of "the sociopolitical framework of secular morality." (p. 152) They report that the humanists describe themselves as "fair-minded and open to ideas different from their own." (p. 152)

Van Eemeren, et al., describe the fundamentalist's view of the exchange as one of the "categories of biblical drama." (p. 155) They cite the fundamentalists' use of moralistic terms like "lust," "fornicate," and "sinner." Van Eemeren, et al., note that the fundamentalists see themselves as "God's anointed representatives." (p. 156)

Van Eemeren, et al., bring the two perspectives together by asserting that not only do these perspectives supply their arguers with "a persuasive guide to the reality of the event," (p. 159) but they also maintain these perspectives by "placing the claims and conduct of others in ironic contrast to their own." (p. 160) So, though the perspectives may seem intractably opposed, they are interdependent because "they continually feed off of one another." (p. 165) The role of the foil, then, seems to balance that of the subject in so far as the crafting of arguments is concerned. As Globus notes, "we learn best who we are from our opponents, not our kindly collaborators who reinforce our own self-image, invincible or self-pitying as the case may be." (p. 1)

A Plan and Sample Analysis

Amended thought experiments, I propose, operate in a manner similar to the exchange between the fundamentalist preachers and the students, and are thus subject to the same kind of analysis. The sample analysis below attempts to examine several exchanges over John Searle's Chinese Room thought experiment by means of an approach consistent with the contextualist principles laid out above. I attempt: (1) to treat the exchange as a dialogue by considering these thought experiments as they evolve in their amended forms throughout the debate, (2) to identify significant shifts in the recounting of the experiment by noting which presentational aspects change and which remain constant, (3) to reconstruct reflexively reasoned accounts by citing stated positions from arguers' writings outside of the immediate context of the thought experiment.

Searle's argument has received a multitude of rejoinders, and as such they constitute an assembly of dialogical exchanges in the debate over artificial intelligence. But in many cases the responses have not been just to the issue of thinking machines but even to Searle's particular example of the Chinese room. The literature, in fact, contains a plethora of examples that attempt to clarify or modify Searle's original scenario in order both to refute and to affirm his argument.

Searle's example starts out as the first-person report of a man in a room with a pile of Chinese symbols and a translation manual:

> Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore (as is indeed the case) that I know no Chinese, either written or spoken, and that I'm not even confident that I could recognize Chinese writing as Chinese writing distinct from, say, Japanese writing or meaningless squiggles. To me, Chinese writing is just so many meaningless squiggles.

> Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that "formal" means here is that I can identify the symbols entirely by their shapes. Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch. Unknown to me, the people who are giving me all of these symbols call the first batch a "script," they call the second batch a "story," and they call the third batch "questions." Furthermore, they call the symbols I give them back in response to the third batch "answers to the questions," and the set of rules in English that they gave me, they call the "program." Now just to complicate the story a little, imagine that these people also give me stories in English, which I understand, and they then ask me questions in English about these stories, and I give them back answers in English. Suppose also that after a while I get so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view—that is, from the point of view of somebody outside the room in which I am locked—my answers to the questions are absolutely indistinguishable from those of native Chinese speakers. (1981, p. 355)

While it is true that Searle's prose style has a reputation for being casual, maintaining as he does here an informal, first-person stance, his choice of first person verbs in the Chinese room example is, I believe, more than just a matter of consistency of style. The emphasis on the first person singular seems central to Searle's argument. Elsewhere in a related discussion Searle makes a case for the subjective nature of consciousness. The upshot of this subjectivity is that the mental is "an irreducibly first person ontology." (1994 p. 95) As a result consciousness is not accessible to all observers in the same way. So by keeping the account of the Chinese room in the first person, Searle keeps the subjective aspect of the mental implicit and thereby encourages the reader to do likewise by running the simulation in his own mind. Searle's purpose here seems to be to segregate consciousness from other phenomena with respect to the way they can be studied, where the Cartesian distinction between *res cogitans* and *res extensa* is usually assumed. A first-person point of view serves to reinforce the subjective nature of the mental and so confers a certain epistemic privilege on itself.

Cole's amended version of the Chinese room does not cooperate with Searle's expectation that the readers will act as simulators of Searle's place in the scenario. Instead Cole takes a third person perspective and treats the scenario as one to be observed rather that simulated. Thus Cole says:

> Wang Hao is in a severe auto accident. Part of Hao's brain is surgically transplanted into John Searle's cranium with some neuron splicing. Sometimes [sic] thereafter, when Searle is addressed in Chinese, he finds himself, as he later describes the incidents in English, "making strange vocal noises" and "committing unnatural speech acts." Among Searle's auditors, English monoglots report that Searle insists he speaks no Chinese. But Chinese monoglots report that Searle speaks Chinese quite well, as well as some other language, but that he insists he speaks no English. (1984, p. 435)

Cole's conclusion is also in the third person: "...then it is not clear that he would not understand Chinese." (p. 435)

Elsewhere Cole offers yet another variation on the Chinese room, where the insistence on the third person is almost tedious:

> Suppose, for example, a person (Searle, in the original statement of the argument) who does not know Chinese sits in a room with instructions written in English (a "program") that tell one in detail how to manipulate Chinese symbols, producing strings in response to the strings given to one... Since the instructions tell one what to do entirely on the basis of formal or syntactic features of the strings, without ever mentioning (or revealing) meaning, one can generate Chinese sentences without any understanding... (1991, p. 400)

Cole begins his critique of Searle's argument by noting: "Searle's self-report of incomprehension of Chinese in his scenario conflicts with other evidence, notably the response in Chinese to Chinese questions." (p. 401)

As in Searle's original account, Cole's choice of person may seem a matter of stylistic consistency. But it seems too strategically connected with his philosophical commitments. Cole is an avowed functionalist; at the end of his argument he has optimistically declared, "Finally, it might even become possible to replace entire damaged neurons by functionally equivalent silicon-based electronic devices." (p. 411) Such sentiments are often accompanied in the AI debate by a rejection of introspection as a method of studying consciousness and an insistence instead on the importance of an analysis of environmental inputs and behavioral outputs. In view of these commitments Cole's insistence on the third-person perspective and his skeptical regard for "Searle's self-report" seem strategic.

Other commitments held by subjectivist and objectivist perspectives are evident in the recounting of Searle's Chinese room scenario. Here Churchland uses the language of behaviorism to summarize two features of the Chinese room:

> First, he [Searle] describes a SM [symbol manipulator] machine that realizes, we are to suppose, an input-output function adequate to sustain a successful Turing test conversation conducted entirely in Chinese. Second, the internal structure of the machine is such that, however it behaves, an observer remains certain that neither the machine nor any part of it understands Chinese. (1990, p. 34)

Searle's "answers to the Chinese questions" (1981, p. 355) in his original account become in Churchland's recounting "an input-output function." Searle described his answers as "equally good;" Churchland says they are "adequate to sustain a successful Turing test conversation." At the end of his original account Searle maintains his subjectivist perspective in concluding, "[I]t seems quite obvious to me in the example that I do not understand a word of the Chinese stories." Churchland recounts this part of the Chinese room example as, "however it behaves, an observer remains certain that neither the machine nor any part of it understands Chinese." So firm are Churchland's objectivist commitments, he seems unwilling to grant Searle a voice to his own self report.

Summary

I believe the examples above illustrate the contextualist principles of argument analysis I have brought together. The content of these examples imply a sense of dialogue; the amended versions of the Chinese Room experiment maintain many of the conditions set out by Searle's original example. Moreover, the accompanying commentaries

of the arguers themselves make a context of dialogue explicit. Cole, for example, prefaces his critique of Searle's account with, "we will do well to consider some additional thought experiments." (1984, p. 435) Within the exchange between Cole and Searle the shift from first-person to third-person narrative seemed to be more than just a matter of stylistic consistency, but rather was a result of prior philosophical commitments. These commitments to a subjectivist perspective in Searle's case and to an objectivist perspective in Cole's suggest that at least one aspect of what Tindale calls a mutual cognitive environment is missing in the context of their disagreement. Finally, I tried to illustrate how the accounts of the Chinese Room might be seen as reflexively reasoned by drawing on the other writings of Searle, Cole, and Churchland. Both Cole and Churchland seem to studiously avoid language that would acknowledge the possibility of speaking felicitously about consciousness in the subjectivist terms that Searle uses so glibly.

What I claim for philosophical thought experiments, Perelman has said more broadly of analogies:

> The whole history of philosophy could be rewritten, emphasizing not structure of systems, but the analogies that guide philosophical thoughts, the way these analogies reply to each other, change, and are adapted to each philosopher's view. (1979, p. 99)

Works Cited

Brown, James Robert, (1992). "Why Empiricism Won't Work," *PSA 1992*, vol. 2, 271-279.

Brown, James Robert, (1991). *The Laboratory of the Mind: Thought Experiments in the Natural Sciences*, London: Routledge.

Brown, James Robert, (1986). "The Structure of Thought Experiments," *International Studies in the Philosophy of Science: the Dubrovnik Papers,* i. pp. 1-15.

Bunzl, Martin, (1996). "The Logic of Thought Experiments," *Synthese*, 106, pp. 227-240.

Churchland, Paul M., and Patricia Smith Churchland, (1990). "Could a Machine Think?" *Scientific American*, 262 (1), 32-37.

Cole, David, (1991). "Artificial Intelligence and Personal Identity," *Synthese*, 88, 399-417.

Cole, David, (1984). "Thought and Thought Experiments," *Philosophical Studies*, 45, 431-444.

Dennett, Daniel C., (1991). *Consciousness Explained*, Boston: Little Brown.

Eemeren, Frans H. van, Rob Grootendorst, S. Jackson, and S. Jacobs (1993), *Reconstructing Argumentative Discourse*, Tuscaloosa, Alabama: University of Alabama Press.

Globus, Gordon G., (1995). *The Postmodern Brain*, Amsterdam: John Benjamins Publishing Company.

Kuhn, Thomas S., (1977). "A Function for Thought Experiments," in *The Essential Tension*, Chicago: The

University of Chicago Press.

Nersessian, Nancy J., (1992a). "How Do Scientists Think? Capturing the Dynamics of Conceptual Changes in Science," in ed. Ronald Giere, *Cognitive Models of Science*, Minneapolis: University of Minneapolis Press.

Nersessian, Nancy J., (1992b). "In the Theoretician's Laboratory: Thought Experimenting as Mental Modeling," *PSA*, vol. 2, pp. 291-301.

Parfit, Derek, (1984). *Reasons and Persons*, Oxford: Clarendon Press.

Perelman, Ch., (1979). *The New Rhetoric and the Humanities: Essays on Rhetoric and its Applications*, Dordrecht, Holland: Reidel

Perelman, C., and L. Olbrechts-Tyteca, (1969). *The New Rhetoric*, Notre Dame: University of Notre Dame Press.

Popper, Karl, (1968). *The Logic of Scientific Discovery*, New York: Harper & Row Publishers.

Puccetti, Roland, (1980). "The chess room: further demythologizing of strong AI," in *The Behavioral and Brain Sciences*, 3, 449-450.

Pylyshyn, Zenon W., (1980). "The 'causal power' of machines," in *The Behavioral and Brain Sciences*, 3, 442-444.

Searle, John R., (1980). "Minds, Brains, and Programs," *The Behavioral and Brain Sciences*, reprinted in Douglas Hofstadter and Daniel Dennett, edd., *The Mind's I*, New York: Bantam.

Searle, John R., (1994). *The Rediscovery of Mind*, Cambridge: MIT Press.

Searle, John R. (1981). "Minds, Brains, and Programs," in *The Mind's I*, eds., Douglas R. Hofstadter and Daniel C. Dennett, New York: Bantam Books

Shapin, S., (1984). "Pump and Circumstance: Robert Boyle's Literary Technology," *Social Studies of Science*, 14, pp. 481-520.

Sorensen, Roy A., (1992a). "Thought Experiments and the Epistemology of Laws," *Canadian Journal of Philosophy*, vol. 22, no. 1, 15-44.

Sorensen, Roy A., (1992b). *Thought Experiments*, Oxford: Oxford University Press.

Steinem, Gloria, (1994). *Moving Beyond Words*, New York: Simon & Schuster.

Tindale, Christopher W., (1992). "Audiences, Relevance, and Cognitive Environments," *Argumentation*, Vol. 6, No. 2, pp. 177-188.

Wilkes, Kathleen V., (1988). *Real People: Personal Identity without Thought Experiments*, Oxford: Clarendon Press.

View Commentary by D. Cohen

View Index of Papers and Commentaries

Return to Main Menu