Induction Week Seminar - Some Philosophical Thought-Experiments

How can you prove or disprove anything in philosophy? One striking way in which people have attempted to achieve this is by using *thought-experiments*. We have gathered together a few examples of philosophical thought-experiments. Please read them through as preparation for the Intro Week seminars.
Pick a couple and try to say:
Either: What it is that the thought-experiment establishes, and why. Or: Why the thought experiment does not show what it was intended to show.

The Madman and the Borrowed Axe

An early example, from Plato, *The Republic, Bk*.I:
Socrates uses a thought-experiment to attack the view that rightful conduct consists in paying one's debts and keeping promises: *Cephalus*: "... Now it is chiefly for this that I think wealth is valuable... For wealth contributes very greatly to one's ability to avoid both unintentional cheating or lying and the fear that one has left some sacrifice to God unmade or some debt to man unpaid before one dies. ..." *Socrates*: "... But are we really to say that doing right consists simply and solely in truthfulness and returning anything we have borrowed? Are those not actions that can be sometimes right and sometimes wrong? *For instance, if one borrowed a weapon from a friend who subsequently went out of his mind and then asked for it back, surely it would be generally agreed that one ought not to return it, and that it would not be right to do so*, or to consent to tell the strict truth to a madman?"

Smart and Smart against Negative Utilitarianism

Utilitarianism is a moral theory according to which the right action to perform is the one that maximizes utility; i.e., increases happiness and reduces suffering as much as possible. One variant on this is *Negative Utilitarianism*. The *Negative Utilitarian* principle, that our overriding moral duty is not to maximise happiness but to minimise suffering, has appealed to some (including Karl Popper; cf. *The Open Society and Its Enemies*, Routledge, London 1957, vol.I ch.5, note 6). It is supported by the thought that we do feel we are under an obligation to help alleviate suffering, whereas it is not at all clear why we should feel that we ought to make people more happy if they are already reasonably content.
However, Negative Utilitarianism appears to be refuted by the following thought experiment. Suppose there were something that you could do — pushing a button, releasing a gas, or whatever — that would painlessly bring about the extermination of all sentient life on Earth (and perhaps elsewhere in the universe as well, if there is more out there). This total euthanasia would be the most certain and secure way of minimising suffering: it would ensure that never again would there be any pain, any agony, any discomfort, any heartache, any grief, any sadness, any sorrow. So according to negative utilitarianism *you would be morally obliged to do it*, if such an act of total annihilation were available to you. But it would — wouldn't it? —be the most terrible act that any right minded person would avoid doing at all costs! So the negative utilitarian principle is wrong. (See R.N. Smart, 'Negative Utilitarianism', *Mind* vol 67, 1958, pp.542-3; J.J.C. Smart and B. Williams, *Utilitarianism: For and Against*, Cambridge University Press 1973, pp.28-9.)

Nozick's Experience Machine

"Suppose there were an experience machine that would give you any experience you desired. Superduper neuropsychologists could stimulate your brain so that you would think and feel you were writing a great novel, or making a friend, or reading an interesting book. All the time you would be floating in a tank, with electrodes attached to your brain. Should you plug into this machine for life, preprogramming your life's experiences? If you are worried about missing out on desirable experiences, we can suppose that business enterprises have researched thoroughly the lives of many others. You can pick and choose from their large library or smorgasbord of such experiences, selecting your life's experiences for, say, the next two years. After two years have passed, you will have ten minutes or ten hours out of the tank, to select the experiences of your *next* two years. Of course, while in the tank you won't know that you're there; you'll think it's all actually happening. Others can also plug in to have the experiences they want, so there's no need to stay unplugged to serve them. (Ignore problems such as who will service the machines if everyone plugs in.) Would you plug in? *What else can matter to us, other than how our lives feel from the inside?* Nor should you refrain because of the few moments of distress between the moment you've decided and the moment you're plugged. What's a few moments of distress compared to a lifetime of bliss (if that's what you choose), and why feel any distress at all if your decision *is* the best one?" Robert Nozick, *Anarchy, State, and Utopia*, Basil Blackwell, Oxford, 1974, pp.42-3.

Nozick himself concludes that "We learn that something matters to us in addition to experience by imagining an experience machine and then realizing that we would not use it." (p.44)

Q: Do you agree? Why would you refuse to use the Experience Machine?

Searle's Chinese Room

'One way to test any theory of the mind is to ask oneself what it would be like if my mind actually worked on the principles that the theory says all minds work on. Let us apply this test to the Schank program* with the following *Gedankenexperiment*. 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 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. Nobody just looking at my answers can tell that I don't speak a word of Chinese. Let us also suppose that my answers to the English questions are, as they no doubt would be, indistinguishable from those of other native English speakers, for the simple reason that I am a native English speaker. From the external point of view — from the point of view of someone reading my "answers" — the answers to the Chinese questions and the English questions are equally good. But in the Chinese case, unlike the English case, I produce the answers by manipulating uninterpreted formal symbols. As far as the Chinese is concerned, I simply behave like a computer; I perform computational operations on formally specified elements. For the purposes of the Chinese, I am simply an instantiation of the computer program.

'Now the claims made by strong AI are that the programmed computer understands the stories and that the program in some sense explains human understanding. ...' From J.R. Searle, 'Minds, Brains, and Programs,' *Behavioral and Brain Sciences* 3, 1980. *Roger Schank and other researchers in Artificial Intelligence at Yale were trying to develop a computer programme which would be capable of answering questions and maintaining a conversation. There has been considerable success in developing AI programmes of this type, provided the subject matter is not too free-wheeling.

Q: would a computer running such a programme be thinking and understanding questions posed?

The Ship of Theseus

In ancient Athens the ship in which Theseus had sailed to the palace of King Minos in Crete continued to be used for ceremonial purposes. But of course the timbers and metal fastenings from which it was made wore out with the passage of time. As each plank and nail became unfit for service it was replaced with a new piece of wood or metal. The original planks and nails were, however, carefully preserved and stored in a temple. Eventually all the original parts of the ship had been replaced by new material. Somebody had the idea that there was a more elegant way of storing them than in a pile inside the temple. Instead they were fitted together in just the same way in which they had previously been, recreating a whole ship.
After that there were two ships: the Ship in the Temple and the Ship at Sea.

Q: Which of these ships is the same ship as the Ship of Theseus? Both of them, neither of them, or just one but not the other?

Parfit on Sharing Out a Brain

'*My Division*. My body is fatally injured, as are the brains of my two brothers. My brain is divided, and each half is successfully transplanted into the body of one of my brothers. Each of the resulting people believes that he is me, seems to remember living my life, has my character, and is in every way psychologically continuous with me. And he has a body which is very like mine.' From Derek Parfit, *Reasons and Persons*, Clarendon Press, Oxford, 1987, pp. 254-5.

Q: Do I survive? As either or both?

Putnam's Twin Earth Thought-Experiment

One of the most discussed thought-experiments in modern philosophy was devised by Hilary Putnam in a paper in which he attempts to persuade us (and has succeeded in persuading many philosophers) that 'meanings ain't in the head'. He introduces it like this: "For the purpose of the following science-fiction examples, we shall suppose that somewhere in the galaxy there is a planet we shall call Twin Earth. Twin Earth is very much like Earth; in fact, people on Twin Earth even speak *English*. In fact, apart from the differences we shall specify in our science-fiction examples, the reader may suppose that Twin Earth is *exactly* like Earth. .. "... One of the peculiarities of Twin Earth is that the liquid called 'water' is not H_2O but a different liquid whose chemical formula is very long and complicated. I shall abbreviate this chemical formula simply as *XYZ*. I shall suppose that *XYZ* is indistinguishable from water at normal temperatures and pressures. In particular, it tastes like water and it quenches thirst like water. Also, I shall suppose that the oceans and lakes and seas of Twin Earth contain *XYZ* and not water, that it rains *XYZ* on Twin Earth and not water, etc." H. Putnam, "The Meaning of 'Meaning'", *Mind, Language and Reality*, Philosophical Papers Vol 2, Cambridge University Press 1975, pp.215-271, at p.223.

Q: Call the stuff on Twin Earth composed of *XYZ* 'twater'. Twater is just like water. So is twater water? Or should we say that because it isn't H_2O it cannot be water?

The Trolley Problem

"Suppose you are the driver of a trolley. The trolley rounds a bend, and there come into view ahead five track workmen, who have been repairing the track. The track goes through a bit of a valley at that point, and the sides are steep, so you must stop the trolley if you are to avoid running the five men down. You step on the brakes, but alas they don't work. Now you suddenly see a spur of track leading off to the right. You can turn the trolley onto it, and thus save the five men on the straight track ahead. Unfortunately... there is one track workman on that spur of track. He can no more get off the track in time than the five can, so you will kill him if you turn the trolley onto him. Is it morally permissible for you to turn the trolley?" From 'The Trolley Problem', Judith Jarvis Thomson, *Yale Law Journal* 94.6, 1985, pp.1395-1415.

Q: Is there a clear answer about what you should do in this case? Can you imagine ways of varying the example which would make the decision (even) more problematic?