ON SOME PERNICIOUS THOUGHT-EXPERIMENTS

Richard M. Gale

THOUGHT-EXPERIMENTS play a prominent role in the practice of analytical philosophy, being one of the chief methods by which analyses are put to the test. Unfortunately, all too often they are misinvoked. They are advertised by their espousers as establishing a certain result when in reality they show something quite different. Because they are misdescribed, their positive upshot goes undetected. Once we grasp the real result of these pernicious thought-experiments, we can reconceive them and put them to a useful purpose. It is my aim to lay the groundwork for this reconceptualization through a botanization of different types of thought-experiments. It is hoped that this will make the pernicious thought-experiments stand out in clear relief from their legitimate brethren and give us a conceptual foundation for reconceiving them.

To understand the point of a philosophical thought-experiment, we must begin with the way in which philosophical analysis has traditionally been conceived—really misconceived. The philosophical analyst is supposed to articulate the sufficient and necessary conditions for the correct application of concept *C* in *every* case, actual as well as merely possible. Toward this end, the analyst first must locate some word *w* in her language that expresses or signifies *C*. She then articulates the rules that prescribe the conditions under which it is correct and incorrect to use *w*, which rules have a normative force in that the people who engage in this rule-governed practice are willing to offer and accept correction when a move deviates from these rules. She is describing, in other words, the rules of a particular kind of normatively rule-governed human practice—a language-game.

No sooner is this accomplished than Counter-Example Man does his stuff. Often this consists in the performance of a philosophical thought-experiment, that is, a description of a merely possible or counter-factual situation in which this analysis is supposed to break down. There are different ways in which an analysis can break down, and it will prove useful to botanize thought-experiments in terms of these ways.

One obvious way is that it faces a clear-cut counter-example in the

envisioned situation: the rules articulated in the analysis require us to use (or refrain from using) *w* in that situation but we do (or do not) not want to use *w*. Thought-experiments can be useful devices for presenting such counter-examples. In some cases they save a lot of time and effort. To refute the ancient analysis of man as a featherless biped it wasn't necessary to go to the trouble of shaving a rooster. It is sufficient to say, "Imagine that there is this shaven rooster...." In some cases it is not practically or even physically possible to bring about the imagined situation. I have no bone to pick with such thought-experiments. We could call them *clear-cut counter-example thought-experiments* on the basis of their upshot.

Another type of *bona fide* thought-experiment presents us with a border-line or undecidable case. The profferred rules for the use of *w* do not determine what is the correct thing to say in the imagined situation. That there are such possible cases for *w* should not surprise us. It is by now the greatest story ever told that no set of rules can be determinative for every possible situation. The results of such *undecidable-case thought-experiments* should have a sobering effect on the practice of analysis. The analyst can no longer boast that she is spelling out the rules for the correct use of *w* in *every* situation, actual as well as merely possible. The analysis must be restricted to clearly decidable cases.

Unfortunately, many undecidable-case thought-experiments are falsely billed as clear-cut counter-example ones. This often happens when *C* is multiple-criterial.[1] Our present concept of personal identity over time, for example, requires both spatio-temporal and psychological continuity. And a case in which only one of them is satisfied is a border-line or undecidable case. Locke thinks that his thought-experiment of the prince and the cobbler allegedly switching bodies because of the interchange of their psychological traits is a clear-cut counter-example to the requirement of physical continuity; but, according to my intuitions, it is only an undecidable case.

Most undecidable-case thought-experiments are crucially underdetermined in that they do not tell us whether the odd-ball case is exceptional or happens frequently, say the majority of the time. If it is rare it leaves the viability of our present language-game intact. For things do not have to go right always. It is sufficient that for the most part they do. But if the undecidable case is the rule rather than the exception the upshot of the thought-experiment is radically different. By seeing the difference in its result, the pernicious thought-experiment will be brought into clear relief.

The difference between a world in which an undecidable case is an exception and one in which it is the rule is the difference between a world in which we engage in our normatively rule-governed language-game with *w* and one in which we do not because it would be futile to do so. It is a serious misdescription of the latter as a world containing undecidable

cases; for in that world we do not play the language-game at all. Of course, we might play some new language-game that bears a family resemblance to our old game; but that is another matter. Whether the same concept *C* enters into both the old and the new language-game is not something that can be decided, since we lack criteria for the identity of a concept not only over time but also across possible worlds.

The reason why our present language-game involving w would not be played in the world in which undecidable cases happen most of the time is that it is futile, in general, to engage in a normatively rule-governed practice when most of the time it is unclear whether a given move is permissible or not. Imagine what it would be like to play chess when most of the time it is unclear whether a given move is permitted or not. It would be like playing craps with dice that have had their markings erased. The stronger might prefer this new "game," since it would present them with the opportunity of bullying their weaker opponents into accepting whatever they claim, on the basis of their memories, are the true markings. But this game would not be our present game of craps. Rather, this "game" is closer to what we call a "shakedown" or just plain highway robbery.

The distinction between the upshots of these two types of undecidable-case thought-experiments rests on an important but often overlooked distinction between the rules of a language-game and the presuppositions for playing this game. One presupposition for playing the game is that for the most part undecidable cases relative to the rules of the game do not occur. But there are more subtle presuppositions for playing a language-game that pertain to the form of life that is involved in the game and the empirical conditions under which it is viable. Such presuppositions involve the conditions under which the purposes or ends for which we engage in the game are realizable. Hopefully a couple of examples will help to clarify the notion of the point or value of a language-game.

Consider the language-game that involves the bestowing of proper names on people which are then used to identify and reidentify them. The rules for this "personal identity" language-game require that for a person P_1 to be numerically one and the same as some earlier person P_2 it must be the case that P_1 is both physically and psychological continuous with P_2. Our playing of this language-game involves certain ways in which people interact, along with the values we ascribe to such interactions. But these interactions have the values we impute to them only in certain types of empirical circumstances. To answer Wittgenstein's pregnant question as to why we give proper names to people but not the pieces in a set of silverware we must see the significant difference between the ways in which we relate to persons and forks (knives, etc.), which differences are quite separate from the fact that persons, unlike forks, answer when called by name. For all human intents and purposes there is nothing to choose

between one fork in the set and another. As the old saying goes, they all look alike (whatever esthetic value they have is shared in common) and fulfill the same function equally well and thus are completely interchangeable. We relate to them as Buberian "its."

How different it is with people. They are worthy of being singled out and traced over time, for they are our "thous"—our friends, lovers, compatriots, colleagues, foes, and the like. Mr. Rogers' seemingly tautological song "Only You are You" really is the exciting tautology that, as a matter of empirical fact, each person is unique and moreover unique in ways that are humanly important. Each person uniquely instantiates some set of properties. Even when two people share some generic or determinable property they usually have some different species or determinate of it; one has øness$_{459}$ and the other øness$_{358}$. Our language is not rich enough to specify these subtle but important differences; but we are able to intuit them: it gives each person a special aura. It is this humanly important uniqueness that makes people special. Only Mary Smith Mary-Smith-izes. If you want to experience that unique set of properties you must go to that bag of flesh and blood over there. It is this that explains our attachment to the flesh and the importance we place upon spatio-temporal continuity of the body in reidentification of persons. The properties that endear Mary Smith to us and make us want to relate to her as colleague, compatriot, lover, and so on, are uniquely expressed through that body.

Very strong emotions and attitudes of love, adoration, respect and affection enter into our playing of the personal identity language-game, for they enter into the forms of life that are realized by the playing of this game. These forms of life involve the special sort of relationships people have to each other and the sort of obligations and duties that they give rise to. Through these relationships we become committed and our lives become purposeful. So seriously do we take other people that we are willing to endow them with a special type of dignity by holding them morally responsible for their actions and thereby invite them to reciprocate by in turn holding us responsible. Our sense of worth as persons is based upon our being willing participants in the moral responsibility game.

It is an empirical presupposition for the forms of life that inform the personal identity language-game to be realizable that persons be unique in ways that have human importance. In the recent literature on personal identity numerous science-fiction type thought-experiments appear that envision counter-factual worlds in which we possess the technology to duplicate people at will, interchange their parts, replace them with synthetic ones, and so on. Sometimes the point of these thought-experiments is that we are too tied to the flesh, that we make too big a deal of spatio-temporal continuity, especially when we make it a necessary condition of personal identity over time. "Why should it matter," we are asked, "whether the woman who returns from a stay in

the hospital is spatio-temporally continuous with your former, beloved wife, rather than a clone that cannot be distinguished from her and fulfills all the same functions?"

What is perverse about these science-fiction thought-experiments is that they transport us, along with our present language-games and their forms of life, into the counter-factual world. And it is claimed that in such a world we face counter-examples to our analysis of personal identity or at least undecidable cases. What they fail to realize is that in this world we would not want to play our old personal identity language-game, since there would be no point or value in doing so: the empirical presuppositions for doing so are not realized. Thus, in this world we do not face counter-examples or undecidable cases, since there no longer is the normatively rule-governed practice of identifying and reidentifying persons. The question of whether the women who returns from the hospital really is my wife wouldn't arise. Furthermore, I would not have an emotional stake in how the question is answered. For the form of life involved in the playing of the personal identity language-game would gain no foothold in the imagined world.

Wittgenstein, whose views I am elaborating on, performs a thought-experiment in which he envisions a counter-factual situation in which there no longer would be any point in playing the personal identity language-game, because an empirical presupposition for doing so is absent.

> Imagine, for example, that all human bodies which exist looked alike, that on the other hand, different sets of characteristics seemed, as it were, to change their habitation among these bodies.... Under such circumstances, although it would be possible to give the bodies names, we should perhaps be as little inclined to do so as we are to give names to the chairs of our dining-room set. On the other hand, it might be useful to give names to the sets of characteristics, and the use of these names would now *roughly* correspond to the personal names in our present language.[2]

I have been trying to explain why we would not be "inclined" to play the game, why it would no longer be "useful" to do so. In his imagined case, as well as in the above medical science-fiction case, we would probably replace our old personal identity language-games with one in which we give names to sets of characteristics and functions. We would not say "Get me Napoleon" but instead "Get me a Napoleon" (while you're up). Wittgenstein's thought-experiment, unlike our science-fiction one, is not perverse because it is properly introduced as a description of a world in which we would not play the personal identity language-game, since there would no longer be any point in doing so, the form of life fostered by the old game no longer being realizable or even desirable in those empirical circumstances.

Another interesting case for my distinction between the rules of a language-game and the empirical presuppositions for playing the game is that of the criterion of identity for so-called *autographic* works of art, that is, works of art that permit of the possibility of a fake or forgery of an

individual work. (It is important to distinguish a forgery of a specific work of art from a fraudulently advertised instance or performance of a work of art: it was billed as a performance of Beethoven sonatas by Dame Myra Hess but they were played by Liberace in drag; he claimed it was Dickens' autograph copy of the manuscript of *A Tale of Two Cities* but it was a forgery.) The criterion of identity for an autographic work of art is spatio-temporal continuity with the unique object crafted by the artist. Part of the playing of the autographic art identity language-game involves our having very different beliefs, attitudes and emotions concerning the original than we have concerning a fake. The very rules of the language-game require us to make an invidious esthetic distinction between the original and the fake. The original, after all, is an original work and the fake is not—certainly an important esthetic difference. We prize, in general, the original in ways that we don't the fake, just as I prize someone who is spatio-temporally continuous with my former wife in ways that I don't someone who is not, for example, a clone.

No sooner is such an account given of our autographic art identity language-game than Counter-Example Man flies through the window—the one which was conveniently left open since the last Dracula film—and hits us with this sort of science-fiction thought-experiment:

> Imagine that we possess a super-duper duplicating machine. We can place the original *Mona Lisa* painting on it and reproduce it molecule by molecule, as many copies as you please. Certainly in such a situation it would be a case of misplaced sentimentality to go on insisting on spatio-temporal continuity as a necessary condition for a painting's being the (or an instance of) the *Mona Lisa*.

Again, we have a misinvoked thought-experiment whose upshot is quite different from the intended one. Instead of describing a possible world in which our present requirement of spatio-temporal continuity for the identity of a painting faces a counter-example or undecidable case it gives us a world in which we would not play the autographic art identity language-game since an empirical presupposition for doing so is missing. Our present requirement of spatio-temporal continuity for the identity of a painting is based on the contingent fact that paintings are unique in ways that are esthetically important and are non-duplicatable. If you want to see the special *Mona Lisa* set of properties and have the unique Mona-Lisa-type esthetic experience you must go that that canvas over there, just as you must go to that bag of flesh and bones over there if you want to have the unique Mary-Smith-type experience. Again, it must be reiterated that the above thought-experiment would not be perverse, but instead serve a useful function, were it introduced as showing a possible world in which we would no longer play our old language-game.

Given that, as a matter of empirical fact, we live in a world in which paintings and people are unique and not duplicatable, we are justified in

caring as much as we do about canvases and flesh and thus in playing the language-games we do.

NOTES

1. For a very insightful discussion of multiple-criterial concepts see Nicholas Rescher, *The Primacy of Practice* (Oxford, 1973), chapter VI. Rescher shows how the viability of such concepts depends upon the empirical presupposition that for the most part we are not confronted with undecidable cases.

2. *The Blue and Brown Books* (New York, 1958), pp. 61-62.