«*GRUE*» and INNATE IDEAS

Robert M. Martin

The debate over innate ideas has a long philosophical history. Plato argued for the view that certain ideas, for example, the concepts of geometry, must be innate - in our minds at birth - because experience is inadequate to produce them. He did not, however, think that babies can, starting at the moment of birth, apply all their ideas. Experience is necessary to bring out these inborn ideas - to «remind» us of them. John Locke's work contains some of the best-known arguments for the contrary view - that all of our ideas are produced by experience. The view that there are innate ideas is a central feature of philosophical rationalism, and its denial is a central position of empiricism. Since the Seventeenth Century, concomitant with the rise of science, empiricism has, on the whole, been the dominant view; but the debate about whether some feature of human understanding is innate or learned has continued. Many scientists think, however, that this debate concerns questions which cannot be answered by philosophical methods, by thought alone; it is rather a debate which can be settled only by empirical research. And many psychologists think that most important human characteristics are a complex mixture of the learned and innate, so no simple answer can be given even by empirical methods. In what follows, I shall go against all of these modern tendencies : I shall argue, on the basis of philophical, not empirical considerations, that empiricism about ideas must be wrong.

My argument depends on a certain invented notion with which you may not be familliar : the notion of «*grue*.» This is not an ordinary English word, and you won't find it in your English/

Turkish dictionary. The word was invented by the American philosopher Nelson Goodman. (Goodman used the term to make a philosophical point quite different from the one I'll be talking about, one with which we need not be concerned.) «*Grue*» is the name of a certain colour. It can be defined in terms of ordinary colour words. First, let's stipulate a centain time, called time *T;* let this be the midnight that begins the year 2000 : 2400 hours, January 1, 2000. Now, we can define «*grue*» in terms of *T* and ordinary colour words, in these two clauses :

Before T, anything is *grue* if and only if it is green.

At *T* or after, anything is *grue* if and only if it is blue.

Thus right now, and for the next few years, grass and traffic lights that tell you to go are green, so they are *grue.* If grass and traffic lights remain green on January 1, 2000, however, they will no longer be *grue.* Right now, the sky on clear days, and my wife's eyes, are blue, so they are not *grue.* If, as expected, the clear sky and my wife's eyes both stay blue at *T* and thereafter, they will have turned *grue* at I.

(You can see why Goodman used the trem «*grue*» - he put together the first part of the word «green» and the second part of the word (blue.)

Goodman also invented the similar term «*bleen*» («green» + «blue»). As you might guess, this is the definition of «*bleen*» :

Before *T,* anything is *bleen* if and only if it is blue.

At *T* or after, anything is *bleen* if and only if it is green.

Thus, we expect, grass will turn from *grue* to *bleen* at *T,* and at that same time, my wife's eyes will turn from *bleen* to *grue.*

Now, when we use colour-words, we have an idea associated with them. The classical empiricists thought that this idea was a fainter mental copy of the mentel experience we have when actually seeing a coloured thing; but this position is controversial, and debating it need not concern us here. I shall use only this minimal and uncontroversial account of what an idea is : to have the idea associated with a word is to have the ability to identify those things to which the word applies, and those things

to which the word does not apply. Thus, most of us have the idea of green, because we are able correctly to sort out things into those which are green and those which are not. (Those of us who are colour-blind perhaps do not have this idea.) You now also have the idea of *grue*, because you are able to sort things into those which are *grue* and which aren't. Note that the sorting we make when distinguishing green from non-green thigs is exactly the same as the sorting we make when distinguishing *grue* things from non-*grue* things. A spinach-leaf goes into the green pile and into the *grue* pile. Note also that, when someone puts a spinach leaf into the *grue* pile, he is correct : that spinach leaf is in fact *grue* now. The fact that it is *grue* has nothing whatever to do with what colour it will be after time *T*. Perhaps it will still be green after *T*, in which case it will have turned *bleen* at *T*. Perhaps it will still be *grue* after *T*, in which case it will have turned blue. (Most likely, if that leaf still exists in January 2000, it will have turned brown or black by then, in which case it will then be neither green nor blue nor *grue* not *bleen.)*

Let me ask you to imagine that someone right now has the idea of *grue* when he uses the word «green». That means that whenever he is asked whether the word «green» applies to something, he looks at the thing, and decides whether or not it is *grue;* it he thinks it is, he answers «yes.» Note that there would be absolutely no difference between this person and the rest of us concerning the way colour-words are used. Despite the fact that he means «*grue*» when he says «green», he applies the word «green» exactly as the rest of us do. Nobody has and way for the next few years of finding out that the idea he has, associated with the word «green» is different from the one the rest of us have. In January, 2000, however, there will be a difference. The rest of us will say that grass is still green, but this person will claim that grass is no longer green : it has changed to a different colour, as have traffic lights that tell you to go, spinach leaves, and so on. He will be very puzzled to find out that everything he used to call «green» has changed colour overnight.

You might object to this story by claiming that nobody except a philosopher could have the idea of «*grue*» at all - it's a very pe-

culiar idea. In order to find out if something fits that idea, you must know what day it is. Around midnight on January 1, 2000, a person with that idea would have to look at a clock to make sure what time it is before he could identify the colour of a spinach leaf. But this objection is a mistake. The person does not have to look at a clock before telling what colour something is. Shortly before midnight, the spinach leaf looks *grue* to him - it fits his *grue* idea, and he says it's «green»; and shortly after midnight the spinach leaf no longer looks *grue* to him - it no longer fits his *grue* idea, and he says that it has changed colour, and it is no longer «green.»

A more precise way of putting this objection is to say that *grue* is not a legitimate colour - concept, since it includes a time in its definition, but ordinary colour - concepts to not. The reply to this objection is that we can define *grue* in terms of green plus a time; but someone who thought in terms of *grue* and *bleen* would have equal basis to criticize our ideas of green and blue because he would thing that our ideas need to mention times in their definitions. Here is how he would define green and blue in terms of his concepts :

Before T, anything is green if and only if it is *grue*.
At T or after, anything is green if and only if it is *bleen*.

and

Before T, anything is blue if and only if it is *bleen*.
At T or after, anything is blue if and only if it is *grue*.

The point is that there's nothing logically wrong with the idea of *grue*.

After T, then, we can imagine that arguments about colour would break out between people who associated the idea of green with the word «green» and those who associated the idea of *grue* with the word «green.» They would disagree about whether spinach leaves had changed colour - about whether they are still «green.»

Would there be a way of settling this argument? Suppose we put a spinach leaf into a machine which tells us the wave-length

of light reflected by the leaf before and after *T*, and the machine tells us that exactly the same wave-lengths of light are reflected before and after. Would this settle the argument and show that the person who thought that the leaf had changed colour was wrong? Perhaps not. That person might merely claim that this experiment shows that reflecting the same wave-lengths of light does not always mean being the same colour. Or perhaps this person would have a different idea associated with his words «the same wave-length», and would claim that it seems to him that the machine shows that the spinach leaf now emits a different wave-length of light.

In any case, the sudden outbreak of arguments at *T*, concerning what colour certain things were, would show that people had different ideas associated with their words. There is no way of knowing that before *T*.

Now consider some different ideas of *grue* and *bleen* - ideas such that the time in question, call it *T**, is a time already past. Suppose that *grue** and *bleen** are defined just as *grue* and *bleen* are, except that *T** is midnight, January 1, 1995. If anyone had these ideas associated with his words «green» and «blue» then several months ago we would have observed that he began saying and believing peculiar things He would have claimed that the sky, spinach leaves, etc., suddenly changed colours overnight. We would have had disagreements with him over what colours things were.

But this did not happen. In fact, both before and after *T**, there was general agreement among everyone about what colour the sky and spinach leaves were. This shows that a certain possible difference in our ideas was not in fact present. People might have had different ideas associated with the words «green» and «blue,» but they did not.

But this fortunate fact cannot be accounted for by the way people had learned to apply colour terms before *T**. People learn to use colour words by hearing others apply them. When we were small children, we heard how others used the word «green,» and tried out various ideas of how that word was used, until we were

able to use that word to apply to things in the same way everyone else did. But note that this process, by which I learned to use the word «green» before T^*, could not have taught me to associate the word with the idea of green in preference to the idea of *grue**. Associating the word with either concept would have worked just as well, because before T^*, everything that was green was *grue**, and everything that was not green was not *grue**. No learning experience before T could have corrected me if I happened to associate the word «green» with the idea of *grue**. Both associations would have worked equally well. Nobody could not have learned, before T^*, to prefer one idea to the other. But the fact is that we somehow managed to avoid this disagreement, and arguments over colours did not suddenly arise in January, 1995. How did we manage to do this? If we cannot have learned to prefer the idea of green to the idea of *grue**, then this preference must have been innate. The fact that we all chose to associate the same colour idea with colour words shows that there is an innate predisposition to certain ideas, and away from others. This is my argument for innate ideas.

This argument can be made more general. It applies not only to the ideas we associate with «green» and «blue», but to all other ideas as well. I shall conclude by a more abstract and general version of this argument.

The process by which we learn to apply category-words always involves a small and finite number of examples. We learn, for example, to distinguish apples from non-apples by means of a small and finite number of examples. Bu, logically speaking, there are an infinite number of different principles of categorization consistent with any finite sample used for learning. The fact that we all do continue to apply the words we learn in much the same way shows that we have the innate tendency to pick from among this infinite number of categorizations in much the same way, without having been taught to. Without this innate tendency, we would all be picking different categorizations, at random, consistent with the same finite sample used for learning, and there would be continual disagreements arising over new categorizations. But fortunately, we all are innately extremely limited, and

limited in the same way, about which ideas we can have. This is not to say, however, that all our ideas are fully formed at birth. Infants cannot, of course, sort out green and non-green things. But the learning process, if it is sucessful, cannot bring about just any idea; it is limited to those ideas for which we have an innate predisposition; the learning process, then, does not create ideas from nothing, but rather «reminds» us of those ideas we are born with a predisposition for. I agree, then, with Plato. John Locke argued that our minds were all «blank slates» at birth, and that we learned all our ideas experience. But if he were right, then there would be no continuing interpersonal agreement about categorization. He was wrong: our ideas must be, to a large extent, innate.

A closing note. The categorizations we use appear on the whole to correspond with constancies in the world outside us. That is to say: many things remain for long periods in the same category, and when we detect change, we can usually explain it. Note that if everyone had the concept *grue*, but that category did not correspond with the world, and things which were green before *T* remained green after T, then we would all be faced with a systematic widespread change our science could not account for. The fact that many of our ideas correspond with constancies in the world is easy to explain. Many of our ideas are simply learned from the world: its constancies cause us to have those ideas. Alternatively, if an idea is innate, then it's reasonable to think that there is an evolutionary explanation of this: an organism with ideas which correspond to the world's constancies would have better prospects for survival and reproduction than one which didn't; so ideas which match the world's regularities would tend to evolve by natural selection. But the peculiarities of *grue* make either of these explanations impossible. As I have argued, before *T* we could not learn that green was a better idea to have than *grue*. But if the aversion to *grue* is, as I have argued, innate, then this innateness cannot have an evolutionary explanation either; the reason for this is that this innate disposition must have been implanted in our minds before *T*, but before *T* the idea of green gave its holder no advantage over the holder of the idea of *grue*.

So the innate aversion to *grue* cannot have an evolutionary advantage. We all are, right now, much better off than we would have been if we had had the innate idea of *grue**, and we be thankful that we did not. But because the evolution of our minds took place before this advantage manifested itself at *T**, evolution cannot be thanked for this fortunate fact. We were just lucky. When one considers the large number of ideas we have which could not have been chosen over their logical competitors through learning or evolution, and when one considers the enormous success our ideas provide us - the facts that we can explain so many constancies in the world, and that we have such a large amount of interpersonal agreement - we can see that we have been, unexplainably, extremely lucky.