BOOK REVIEWS

The Freedom of the Will. J. R. LUCAS. New York: Oxford, 1970. 181 p. $5.00.

Ten years ago, in "Minds, Machines and Gödel" [1] J. R. Lucas attempted to give careful expression to an argument that had enjoyed considerable uncritical acceptance among mathematicians and philosophers, to the effect that Gödel's incompleteness theorem showed that men were capable of feats no conceivable mechanical computer could duplicate and, hence, that men escaped the clutches of mechanistic determinism. Roughly a third of his new book is a refurbishment of his earlier provocative argument, and the rest sets the stage by dismissing all other likely candidates for solutions to the problem of free will. Lucas's final position is libertarian: only freedom from mechanistic causation can give men moral responsibility, but quantum mechanics makes plausible and Gödel's theorem proves that men enjoy just this sort of contracausal liberty. In order to make use of such a last-ditch defense of liberty, Lucas must find error in all more reconciliatory views, and his attempts to do this in the first two-thirds of the book bear out his modest claim that he has "not tried to do justice to them" (2). A less convincing array of persuasions would be hard to find.

The core of Lucas's argument is to be found in his three-page chapter 25, and I will concentrate on one gap in the argument presented there, because I believe it cannot be filled in any plausible way, and because, if Lucas's bibliography of his critics is exhaustive, it has not been examined in detail before. The strategy of Lucas's argument is to show that mechanistic determinism requires a characterization of a man as a certain sort of theorem-proving machine (on the shaky grounds that at least some men sometimes prove theorems, and, hence, if men are machines, they are theorem-proving machines). Then Gödel's theorem is rung in to establish a certain frailty in all such machines, but men are seen by empirical observation to overcome this handicap; ergo, men are not machines and mechanistic determinism is false of men.

Since Gödel's theorem is not about material objects or minds, but about abstract formal systems, if Lucas is to extract any antimechanistic consequences from Gödel, he must build a bridge between truths about formal systems and truths about the motions of physical objects. This he has failed to do. He opens chapter 25 with

[1] *Philosophy*, XXXVI, 137 (April/July 1961): 112–127.

the claim that "Gödel's theorem applies to physical determinist systems, which are sufficiently rich to contain an analogue of simple arithmetic," a statement in need of much more unpacking than he ever provides. These "physical determinist systems" are presumably theorem-proving machines, devices of the sort characterizable as Turing machines. A particular Turing machine is specified by a particular "machine table" of interrelated instructions to perform simple computations, and any entity that can be interpreted to "follow" those instructions is a *realization* of that Turing machine. Hence a particular Turing machine can be realized in a variety of ways: e.g., by a piece of mechanical or electronic hardware, or by a person following the instructions using pencil and paper (called "hand simulation" by programmers). Gödel's theorem shows that, for each Turing machine that produces a consistent output in a language capable of expressing the truths of arithmetic, there is a sentence of its language, viz., its Gödel sentence, that it can never prove in the course of producing its output, but which "we" can establish to be true. Of course it does not follow from this that if something realizes Turing machine k, it can never prove k's Gödel sentence; for suppose that Lucas hand-simulates (and thereby is a realization of) k, and while doing this is asked to prove k's Gödel sentence; this may well present no problems to him—he will merely cease to simulate k while he proves it. What he cannot do is prove the Gödel sentence *in the course of* simulating k, and angels, were they to try, could do no better. Taken this way, Gödel's theorem does not distinguish physical deterministic systems from any others; nonphysical, indeterministic, supernatural, even divine entities, if such there be, cannot get their hand simulations of Turing machines to come up with proofs of the associated Gödel sentences, even given eternity to work at it.

Nor is it the case that physically determined entities lack the protean capacity of persons (and indeterministic angels) to change Turing-machine guises. Any physical object can be interpreted to be a variety of different Turing machines, indeed many at once.[1] Depending on what events we wish to interpret as input-output symbol tokens (the other events will be "noise") and what physical states we wish to interpret as logical machine states, an object—animal,

[1] Gilbert Harman makes a similar point in "Three Levels of Meaning," this JOURNAL, LXV, 19 (Oct. 3, 1968): 590–602, p. 595/6. Choosing a preferred interpretation for an object must be either arbitrary or dependent on extrinsic considerations (e.g., what one wants to use the object for), a point I argued in "The Abilities of Men and Machines," read to the APA Eastern Division Meeting, Philadelphia, Dec. 29, 1970; see, *ibid.*, LXVII, 20 (Oct. 22, 1970): 835.

vegetable, or mineral—can be given different, simultaneously applicable interpretations as a Turing machine. Object *a* can realize Turing machines *k,l,m,* . . . all at once, and so from the premises that (i) *a* realizes *k*, and (ii) *s* is *k*'s Gödel sentence, it does not follow by Gödel that *a* cannot prove *s*, because in *a*'s guise as *l* (or *m*, or *n*, or . . .), *a* may well be able to prove *s*. Further, at any moment an event which, relative to one interpretation, is *noise* may bring it about that the conditional regularities on which that interpretation is based cease to hold, "breaking" the machine, and then the event sequence which under that interpretation would be the production of the unreachable theorem, its Gödel sentence, may occur. So by accident or design an object may begin to "follow" different "rules." Gödel's theorem is in this respect like proofs that certain positions are impossible in chess; these positions are unreachable so long as one does not break the rules, but a person or a chess-playing machine can break the rules easily enough—and nothing physical need rupture for this to occur.

But is there not going to be some overarching Turing-machine description of any physically deterministic object which includes all possible changes of program (somewhat like the machine table of the "universal" Turing machine that can be programmed to simulate any Turing machine) and from which the object cannot escape, as it were? There are two avenues to explore here. First, if there is a finest-grained physical description of the object (surely a large assumption) and if we have deterministic laws of nature governing the behavior of these finest grains, then it seems that we could in principle devise a finest-grained, most inclusive Turing machine interpretation for the object, where all events are tokens of symbols, all states are logical states, all laws of nature are interpreted as inference rules for the machine—in short where the object cannot cease to be the Turing machine in question, cannot "break" at all, since there is no event (no noise) that can break it. For such a machine it seems we might have an unconditional, you-can't-get-there-from-here proof from Gödel; we choose our interpretation of the symbols in such a way that the multitude of atomic perturbations of the thing are proofs of theorems in arithmetic, and can then define a sequence of events in which this object cannot participate, unless the laws of nature change. I cannot see that there are any logical obstacles to this interesting idea (there may be), but in any case it will not give Lucas what he needs (and he seems to suggest this interpretation at least once, on p. 165), for surely none of his observations about the cleverness of mathematicians shows that human beings would be

exempt from such a consequence. Taken in this direction, Gödel's theorem has implications about the theorem-proving capacities of, say, oak trees: though each oak tree, with its waving branches and falling acorns, can prove innumerable theorems (!) there is one it cannot prove: its Gödel sentence.

The other avenue is to claim that, if a man is a physically determined entity, we must be able to distinguish, by some finite, empirical test, precisely which of all the possible bodily events and states of that man are to be interpreted as symbol tokens of the man's output language or machine states of his proving mechanism. This would permit us to identify all other events once and for all as noise and to fix on one Turing-machine interpretation as the only correct one. If an upset stomach or a blow on the head or the sight of a pretty girl interfered with the normal operation of the theorem-proving system, we could say that the mechanism was in some strong sense broken or at least temporarily out of order. If it is tempting at all to adopt this view, it is because our models—actual hardware computers—have relatively obvious and definable purposes and, hence, relatively transparent functional structures; we do not suppose for a minute that the slight hum being emitted is an output symbol or that the dust on the top is an indicator of a computer's logical state. However, the purposes of a man (and this is a different and more complex sense of 'purpose') are not so readily circumscribed, and it is not at all clear what criteria could be invoked to separate symbol from noise. Yet this is the avenue Lucas thinks his determinist opponent must take. The determinist, if he is to have a satisfactory theory of human behavior, must "fix what descriptions of subsequent states or processes . . . are to be regarded as uttering statements or writing formulae (of any sort), or making calculations or drawing inferences (of any sort)" (131). The determinist must specify in finite fashion precisely what physical motions of the man are to be interpreted as his "producing as true" a proposition. But why? Nothing about the physicality of a man could force the determinist to fix his interpretation. He can as easily view a man as a continuously changing succession of "fragile" Turing machines, or refuse to play the Turing-machine game altogether.

Lucas's alternative is highly counterintuitive, for men do not sit around uttering theorems in a uniform vocabulary. They say things in earnest and in jest, make slips of the tongue, speak several languages, signal agreement by nodding or otherwise acting nonverbally, and—most troublesome for this account—utter all kinds of nonsense and contradictions, both deliberately and inadvertently.

Are we to suppose that we can use physical criteria to partition this multifarious "output" into the part that is the *intended,* error-corrected, theorem-proving part, and the part that is not? Whatever else "producing a proposition as true" is (Lucas seems to view the notion as unproblematical), it is an intentional action, and Lucas thus must maintain that the determinist is committed to the view that intentional-action types (at least those involved in proving theorems) can be defined in terms of a finite number of physical characteristics of situations and bodily motions (131). Since earlier in the book Lucas gives short shrift to the supposition that "rational" explanations (explanations of intentional actions, roughly) and the "regularity" explanations of the determinist can coexist, it is hard to see how he can require the determinist to define intentional-action types in a physical-determinist vocabulary. In any case, if Lucas could establish that all determinists are bound to this implausible view (and he gives no argument I can find), we would not need Gödel's theorem to discredit determinism, but, unless he can entice the determinist into accepting this view, he has no opponent against whom to play out his "dialectical" argument from Gödel.

The topics tackled in this book are notoriously slippery, and require precision in handling, but Lucas works impatiently, presenting objections sketchily and responding to them obliquely. He obtains some of his pet conclusions by patent *non sequiturs,* aided by a great deal of hand waving, and not a few elementary logical confusions and slips (e.g., pp. 39, 61, 72, 110). But he gives us a splendid display of his command of Latin, Greek, and Middle English, and, except for failing to give any reference for a crucial argument of Wolfgang Pauli's that he mentions (p. 112), his footnotes are erudite.

D. C. DENNETT

Tufts University

NEW BOOKS: ANTHOLOGIES

FRIEDMAN, JOYCE, *et al.: A Computer Model of Transformational Grammar.* New York: American Elsevier, 1971. 166 p. $15.95.

HUNTER, W. B., C. A. PATRIDES, and J. H. ADAMSON, eds.: *Bright Essence: Studies in Milton's Theology.* Salt Lake City: Univ. of Utah Press, 1971. ix, 181 p. $7.95.

KOCKELMANS, JOSEPH J., and THEODORE J. KISIEL, eds.: *Phenomenology and the Natural Sciences.* Evanston, Ill.: Northwestern, 1970. xxi, 520 p. $15.00.