Papers in philosophical logic

DAVID LEWIS

Princeton University

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge CB2 1RP, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, United Kingdom
40 West 20th Street, New York, NY 10011-4211, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1998

Printed in the United States of America

Typeset in Bembo

Library of Congress Cataloging-in-Publication Data

Lewis, David K., 1941–

Papers in philosophical logic / David Lewis.

p. cm. – (Cambridge studies in philosophy)

Includes bibliographical references and index.

ISBN 0-521-58247-4 (hardback). – ISBN 0-521-58788-3 (pbk.)

1. Logic, Symbolic and mathematical. I. Title. II. Series.

BC135.L44 1998 97-6656

160–dc21 CIP

A catalog record for this book is available from the British Library.

ISBN 0 521 58247 4 hardback
ISBN 0 521 58788 3 paperback

13

Lucas against mechanism II

J. R. Lucas serves warning that he stands ready to refute any sufficiently specific accusation that he is a machine. Let any mechanist say, to his face, that he is some particular machine M; Lucas will respond by producing forthwith a suitable Gödel sentence ϕ_M. Having produced ϕ_M, he will then argue that – given certain credible premises about himself – he could not have done so if the accusation that he was M had been true. Let the mechanist try again; Lucas will counter him again in the same way. It is not possible to accuse Lucas truly of being a machine.[1]

I used to think that the accusing mechanist interlocutor was an expository frill, and that Lucas was really claiming to be able to do something that no machine could do.[2] But I was wrong; Lucas insists that the interlocutor does play an essential role. He writes that "the argument is a dialectical one. It is not a direct proof that the mind is something more than a machine; but a schema of disproof for any particular version of mechanism that may be put forward. *If* the mechanist maintains any specific thesis, I show that a contradiction ensues. But only if. It depends on the mechanist making the first move and putting forward his claim for inspection."[3] Very well. I

First published in *The Canadian Journal of Philosophy* 9 (1979), 373–376. Reprinted with kind permission from *The Canadian Journal of Philosophy*.

1 J. R. Lucas, "Minds, Machines and Gödel," *Philosophy* 36 (1961), pp. 112–27.

2 David Lewis, "Lucas Against Mechanism," *Philosophy* 44 (1969), pp. 231–33; reprinted as Chapter 12 of this volume.

3 J. R. Lucas, "Satan Stultified: A Rejoinder to Paul Benacerraf," *Monist* 52 (1968),

promise to take the dialectical character of Lucas's argument more seriously this time – and that shall be his downfall.

Let O_L be Lucas's potential arithmetical output (i.e., the set of sentences in the language of first order arithmetic that he is prepared to produce) when he is not accused of being any particular machine; and for any machine M, let O_L^M be Lucas's arithmetical output when accused of being M. Lucas himself has insisted (in the passage I quoted) that the mechanist's accusations make a difference to his output. Therefore we cannot speak simply of Lucas's arithmetical output, but must take care to distinguish O_L from the various O_L^M's.

Likewise for any machine M: let O_M be M's arithmetical output when not accused of being any particular machine, and let O_M^N be M's arithmetical output when accused of being some particular machine N. If the machine M, like Lucas, is capable of responding to accusations, then O_M and the various O_M^N's may differ.

We may grant Lucas three premises.

(1) (Every sentence of) O_L is true. For O_L is nothing else but everyman's arithmetical lore, and to doubt the truth thereof would be extravagant scepticism.
(2) O_L includes all the axioms of Elementary Peano Arithmetic. Lucas can easily convince us of this.
(3) For any machine M, O_L^M consists of O_L plus the further sentence ϕ_M, a Gödel sentence expressing the consistency of M's arithmetical output. It is Lucas's declared policy thus to respond to any mechanistic accusation by producing the appropriate Gödel sentence; and – ignoring, for the sake of the argument, any practical limits on Lucas's powers of computation – he is able to carry out this plan. (We may take it that a mechanistic accusation is not sufficiently specific to deserve refutation unless it provides Lucas with a full functional specification of the machine he is accused of being: a machine table or the like.)

pp. 145–46. See also J. R. Lucas, "Mechanism: A Rejoinder," *Philosophy* 45 (1970), pp. 149–51; and J. R. Lucas, *The Freedom of the Will* (Oxford, 1970), pp. 139–45.

Let the mechanist accuse Lucas of being a certain particular machine M. Suppose by way of *reductio* that the accusation is true. Then $O_L = O_M$ and $O_L^M = O_M^M$.

M is a machine. In the present context, to be a machine is not to be made of cogwheels or circuit chips, but rather to be something whose output, for any fixed input, is recursively enumerable. (More precisely, the set of Gödel numbers encoding items of output is recursively enumerable.) If the whole output of M, on input consisting of a certain mechanistic accusation, is recursively enumerable, then so is the part that consists of sentences of arithmetic: O_M^M, in the case under consideration.

Then there is an axiomatizable formal theory θ that has as theorems all and only the sentences of arithmetic that are deducible in first order logic from O_M^M. Further, θ is an extension of Elementary Peano Arithmetic: by premise (2) the axioms thereof belong to O_L, by premise (3) O_L is included in O_L^M, O_L^M – that is, O_M^M – is included in θ. Hence θ is the sort of theory that cannot contain a Gödel sentence expressing its own consistency unless it is inconsistent.

Is θ inconsistent? Apparently so. The Gödel sentence ϕ_M belongs to O_L^M, hence to O_M^M, and hence to θ.

Yet if ϕ_M is true, then O_L^M, which is O_L plus ϕ_M, is true by premise (1); hence O_M^M is true, hence θ is true and *a fortiori* consistent.

Lucas says that he can see that ϕ_M *is* true. Surely he means that he can see that *if* the accusation that he is M is true, *then* ϕ_M is true. If he meant more than that, the accusation – which he disbelieves and is in process of refuting – is irrelevant; he ought to be able to see that ϕ_M is true without the accuser's aid, contrary to his insistence on the dialectical character of his argument.

How could he see that? Perhaps as follows. (I can see no other way.) By premise (1), Lucas's arithmetical output is true. If true, then *a fortiori* it is consistent. If the accusation that Lucas is M is true, it follows that the arithmetical output of M is consistent. Accordingly, a Gödel sentence expressing the consistency thereof is true – and ϕ_M is just such a sentence.

And so the supposition that Lucas is M has seemingly led to contradiction. On the one hand, θ contains ϕ_M and must therefore be inconsistent; on the other hand ϕ_M is true, so θ is true, so

θ is consistent. The mechanistic accusation stands refuted. Q.E.D.

Not quite! We must be more careful in saying what ϕ_M is. It is, we said, "a Gödel sentence expressing the consistency of M's arithmetical output". Does ϕ_M then express the consistency of O_M, M's arithmetical output when not accused of being any machine? Or of O_M^M, M's arithmetical output when accused of being M? After all, under the supposition that Lucas is M, M has in fact been accused of being M and M's arithmetical output may well have been modified thereby.

First case: ϕ_M is a Gödel sentence expressing the consistency of O_M, M's original arithmetical output unmodified by any accusation. Then we have a correct proof (given premise (1)) that if Lucas is M, then ϕ_M is true. But this ϕ_M does not express the consistency of O_M^M, so it may belong to θ although θ is true and hence consistent. In this case Lucas's *reductio* against the accusation that he is M fails.

Second case: ϕ_M is a Gödel sentence expressing the consistency of O_M^M, M's arithmetical output when accused of being M. Then, since ϕ_M also expresses the consistency of θ, ϕ_M cannot belong to θ unless θ is inconsistent and ϕ_M is therefore false. If Lucas is M, ϕ_M does belong to θ and is false. But so be it. In this case we have no good argument that ϕ_M is true. Even if Lucas is M, ϕ_M no longer expresses the consistency of the trustworthy O_L, but rather of O_L^M: that is, of O_L plus ϕ_M itself. If we tried to argue that ϕ_M is true (if Lucas is M) because it expresses the consistency of a set of truths, we would have to assume what is to be proved: the truth, *inter alia*, of ϕ_M. In this case also Lucas's *reductio* fails.

There are machines that respond to true mechanistic accusations by producing true Gödel sentences of the sort considered in the first case; for all we know, Lucas may be one of them. There are other machines that respond to true mechanistic accusations by producing false Gödel sentences of the sort considered in the second case; for all we know, Lucas may be one of them. Perhaps there also are non-machines, and for all we know Lucas may be one of them.

To confuse the two sorts of Gödel sentences is a mistake. It is part of the mistake of forgetting that the output of Lucas, or of a machine, may depend on the input. And that is the very mistake that Lucas has warned us against in insisting that we heed the dialectical character of his refutation of mechanism.