Etica & Politica / Ethics & Politics, 2003,
http://www.units.it/~dipfilo/etica_e_politica/2003_1/1_monographica.html

Guest Editor's Preface

Introduction to Lucas's argument against Mechanism by means of Gödel's Incompleteness Theorem (*)

Barbara Giolito

Dipartimento di Filosofia
Università Piemonte Orientale 'A. Avogadro', Vercelli

With this issue, Etica & Politica / Ethics & Politics *intends to offer homage to John Randolph Lucas, one of the most significant contemporary philosophers, by presenting the debate about Mechanism raised in the second half of the Nineteenth Century as a consequence of some very relevant articles proposed by Lucas. Lucas suggested for the first time in "Minds, Machines, and Gödel" - an article published in Philosophy, XXXVI during 1961 - the possibility of proving the fallacy of any Mechanist position, by showing the existence, for every machine, of a true sentence recognizable by human beings but not by the machine itself: this article was followed by many replies and it has therefore produced interesting disputes between Lucas himself and other philosophers.*
Every article proposed in this issue has been already published in other reviews: with regard to this we would like, first of all, to thank authors and reviews that have given us the authorization to reproduce such articles, in particular the British Journal for the Philosophy of Science *and I. J. Good for* Human and Machine Logic*,* The Monist *and P. Benacerraf for* God, the Devil, and Gödel*,* Philosophia *and A. Hutton for* This Gödel is killing me*,* Philosophy *and D. Coder for* Gödel's Theorem and Mechanism*, and again* Philosophy *and D. Lewis for* Lucas against Mechanism*. As underlined, this presentation does not aim to propose new papers but to make previously published papers more easily accessible by means of a unitary issue. Moreover we present a recent interview with Lucas, whereby we try to clarify some interesting but controversial questions concerning the use of Gödel's Incompleteness Theorem against Mechanism; with reference to this interview, we would like to conclude whit a particular thanks to Lucas himself for his kindness and for his permission to reproduce the articles here presented.*
My own thanks to those who have given me the possibility of contributing to this project: in particular Pierpaolo Marrone who offered me the opportunity of writing for Etica & Politica / Ethics & Politics*, Diego Marconi for his organization of my PhD course, and Michele Di Francesco, Paolo Casalegno and Piergiorgio Odifreddi who helped me during the study of this subject.*

The nineteenth century sees a revival of the debate about the possibility of analysing the human mind's operations with the models of Mechanicism and, thanks to the evolution of computing science, of interpreting human thought by means of the analogy with particular kinds of software. In the attempt to contrast these aims, very different arguments have been used; the most interesting ones have been developed thanks to the significant results of contemporary logics.

In this context we can find a good number of remarkable and important articles written by Professor John Randolph Lucas: in his opinion it is possible to find interesting answers to ethical problems – such as the impossibility of reducing the human mind to a sort of computer's program – and also to different subjects such as mathematics and logics. He tries indeed to maintain the difference between human thought and computers' work by means of the Incompleteness Theorem, one of the most important results of contemporary logics, demonstrated by Kurt Gödel in 1931 and presented in 'Über formal unentscheidbare Sätze per Principia Mathematica und verwandter Systeme I'(1).
In the Lucas's opinion, there is at least one sentence whose truth a rational human being can see and a computer is on the other hand incapable of finding and such a fact can be pointed out by means of the Gödel's first theorem: this theorem proves – for every formal system consistent and powerful enough – the existence of a formula that can be expressed in the system and recognised as true by human beings, but that can't be demonstrated by the system itself. Considering that a computer is constituted by nothing else than an implementation of a formal system by a physical machine and that the work of every machine can be expressed by a formal system, it is possible to maintain – in Lucas's view – that for every computer and, in general, for every machine there is a sentence that can be formulated by the language of that computer or machine but that can't be proved by it: at the same time we can see the truth of this sentence by following the argument outlined by Gödel in his theorem. If, for every machine, a sentence exists about which we (human beings) can state something that the machine can't, it is possible to conclude that no machine and no computer can be able to reproduce human thought and therefore that human thought is something different from the work of any machine or computer.
A formal demonstration of the Incompleteness Theorem is far too complicated and technical for this context, but we can try to understand the general method of its development by means of an informal argument that can be compared whit that theorem: the liar's paradox. Lets suppose that someone states 'What I am asserting is false': if what he's saying is really false, it is false that what he is saying is false and so what he's asserting is true; on the other hand, if what he's asserting is true, it is true that what he's saying is false; in any case the statement produces a contradictory conclusion. The contradiction is due to the fact that the statement is auto-referencial: one of the most interesting aspects of the argument developed by Gödel consists in reproducing an argument from a point of view which is very similar to this one, although he substitutes the concept of 'truth' with the concept of 'provability'. As for the possibility of expressing meta-mathematical assertions inside arithmetic itself, Gödel creates an arithmetical sentence asserting its non-provability within the system (lets call it 'gödelian sentence' of the system); the demonstration of the non-provability of this sentence is carried out with logical instruments, thus proving that if this sentence can be demonstrated inside the system the system itself is not consistent (2), however the truth of the sentence - that can't be proved in the system as it is showed by the demonstration just mentioned - is intuitively derivable: the sentence asserts indeed its own non-provability and, because its not-provable, is exactly what it has already verified: if a human being can understand such a demonstration he can also recognise the truth of the sentence.
Therefore, according to Lucas, for every formal system - and for every possible machine (since a machine is nothing else than the concrete realisation of a formal system) - we can find a sentence not provable inside the system (and therefore not provable by the corresponding machine), but whose truth we are able to recognise if we use the argument conceived by Gödel. In Lucas opinion such a use of Gödel's first theorem can represent a good argument against the attempts of maintaining a Mechanist position: Lucas does not intend to suggest that his argument must to be interpreted as a complete and invincible demonstration against every form of Mechanism, but that it can be a good scheme of argument against those who maintain the truth of Mechanism. Mechanism, in all of its various forms, affirms that there is a machine at

least capable of reproducing what human beings are able of doing with their reasoning: nevertheless human beings are in principle able of recognising the truth of the gödelian sentence for any formal system and therefore they are able of recognising the truth of the gödelian sentence for any machine, truth that the corresponding formal systems and machines cannot reach. Therefore, for every machine, we can identify a sentence whose truth we are able of recognising but that the machine itself cannot recognise, thus – for every machine – there is a task that a human being can do, at least in principle, but that a machine cannot: for this reason we can maintain that, if for any machine there is a task impossible for it but possible for a human being, human thought cannot be reduced to the operations of a machine and therefore Mechanism is wrong.

In reply to the defenders of Mechanism, who suggest that for every machine M there is a machine more powerful than M capable of proving the gödelian sentence for M, Lucas responds that also for this new machine there is a sentence – the goedelian sentence for this new machine – that can be recognised as true by human reasoning but not by the machine itself, and so for any more and more powerful machines, capable of demonstrating the gödelian sentence for less powerful machines but not for themselves. On the other hand, a representative of Mechanism could suggest that neither machines nor human beings are capable of seeing the truth of the goedelian sentence for the most interesting formal systems and related machines, because this truth depends on their consistency and on another important theorem - Gödel's second theorem - which shows that for every formal system, powerful and consistent enough, it is not possible to prove its consistency within the system itself. Therefore, for systems of this kind, we can only suppose that their goedelian sentence is true but we can never confirm this hypothesis. Nevertheless, in Lucas opinion, it is not correct to maintain that - in general - we are not able of proving the consistency of a formal system: the only requirement of Gödel's second theorem is that such a consistency can't be demonstrated within the system itself, not that it can't be formally demonstrated in any way. A convincing proof of Peano's arithmetic – for instance – is formulated by Gentzen by using transfinite induction: thanks to Gödel's second theorem we know that we can't prove the consistency of Peano's arithmetic by means of its axioms and rules, but such a theorem does not prevent us from obtaining a demonstration of consistency for this systems by applying to principles external to the system itself. Moreover, although Gödel's second theorem shows that we cannot demonstrate the consistency of a formal system within that same formal system, we can in many cases argue for its consistency by means of wider and more general considerations (if we really suppose that formal systems expressing mathematics and arithmetic are inconsistent, we must indeed abandon not only these fields but also all the scientific subjects correlated to them, that is the great majority of our scientific knowledge).

In Lucas's opinion, from the a. m. analysis of Goedel's Incompleteness Theorem we obtain a new and powerful tool against Mechanism: the importance and peculiarity of the argument suggested by Lucas consist in developing from the analysis of rational mental faculties, that is mathematical operations, a characteristic that differentiates Lucas's proposal from the already diffused arguments against Mechanism that refer to other human ability such as fantasy, creativity, artistic ability and so on, hardly reducible to some formal and mechanic process.

Note

(*) For the copyright of the papers of J.R. Lucas see http://users.ox.ac.uk/~jrlucas/ back

(1) K. Gödel, *Über formal unentscheidbare Sätze per Principia Mathematica und verwandter Systeme I* , in 'Monatshefte für Mathematik und Phisyk', 38, 1931. back

(2) Lets imagine an arithmetical formula G representing the meta-mathematical sentence 'G is not provable': by means of the possibility of expressing meta-mathematical expressions in arithmetical terms, such a formula can correspond to a fixed number h and be equivalent to the sentence 'the formula associated to h is not provable'; it is possible to prove that G is demonstrable if and only if ~G is demonstrable. We can take a formula ~Dim(x,y), representing the meta-mathematical sentence 'the formulas' sequence associated to x is not a demonstration for the formula associated to z' (where x and z are numbers); we obtain, by adding (x) – that means 'for every x' – at the beginning of this formula, a new formula (x)~Dim(x,z), corresponding to the sentence 'for every x, the formulas' sequence associated to x is not a proof of the formula associated to z', that is 'the formula associated to z is not provable'. Lets consider (x)~Dim[x,sost(y,13,y)], representing 'the formula associated to the number sost(y,13,y) is not provable', where sost(y,13,y) means 'the number of the formula obtained from the formula associated to y by substituting the variable associated to 13 whit the number corresponding to y'. We can suppose that the number associated to (x)~Dim[x,sost(y,13,y)] is n: lets substitute, in this formula, the variable corresponding to 13 - that is y - with n; we obtain a new formula (x)~Dim[x,sost(n,13,n)], that we may call G. The number corresponding to this formula is sost(n,13,n): for G means 'the formula associated to sost(n,13,n) is not provable', (x)~Dim[x,sost(n,13,n)] represents the meta-mathematical sentence '(x)~Dim[x,sost(n,13,n)] is not provable'. We obtain in this way a formula G asserting its non-provability: G is not formally provable, otherwise its negation – that is ~(x)~Dim[x,sost(n,13,n)] – would be equally provable, and so it is true. In fact, if G was provable, there would be a formulas' sequence proving G within arithmetic: lets imagine a number k corresponding to such a proof. The arithmetical relation Dim(x,z) would have to link k and sost(n,13,n), so Dim[k,sost(n.13,n)] would have to be not only true but also formally provable and from this formula we could deduce – by logic's transformational rules - ~(x)~Dim[x,sost(n,13,n)]. back