GÖDEL’S THEOREM AS AN ARGUMENT TOWARDS THE COMPUTATIONAL THEORY OF THE MIND. AN HISTORICAL OVERVIEW OF THE DEBATE. CONCLUSIONS AND INQUIRIES FOR THE FUTURE.

Tzanis D. (MSc on Applied Mathematics, & Intelligent Systems)

The Computational Theory of the Mind (CTM), according to which the fundamental function of the human mind is computation, exists as one of the dominant models of cognitive process analysis, gaining ground in the Cognitive Science field. CTM, nevertheless, has been - and still is - a subject of intense dispute, which is generally confined in a continuum where “mind can not be simply a computer” or “certain cognitive processes may be in existence (intuition, perception, emotion, and more) which can not be deducted to simplified mechanical computations”.

A long-standing debate of such a kind stemmed from Gödel’s Incompleteness Theorem, which instantly caught attention due to the fact that it invoked a mathematical, and therefore, by principle, reliable result. In its basic form, Gödel’s Incompleteness Theorem refers to simple arithmetics, which declares as non complete, since at least one of its sentences is not decidable (the theorem does not infer to its true or false nature). What makes Gödel’s theorem unique is that the undecidable sentence, supposing G, is deduced from a true one of the kind “G is true, if and only if G has no proof in simple arithmetics”. Hence, one could suppose that G, being undecidable, is indeed true, not having a proof in the context of the theorem. The apparent simplicity of the truth of G (which Gödel himself also shared – Gödel, 1931) became the cause of an ongoing debate for the last 60 years.

The Gödelian argument against CTM was first formulated by Nagel & Newman (Putnam, 1960), gained its fame due to J.R. Lucas (Lucas, 1961) and was shaped into its contemporary form by R. Penrose (Penrose, 1989 - 1996). The argument is based on the fact that the Incompleteness Theorem is valid for every 'powerful enough' typical system and that a computer (Turing Machine) will undoubtedly contain such a system (Lucas) or it will be equal to it in terms of computation (Penrose). If we apply Gödel's Theorem on such a computational system, we may find its undecidable sentence (or principle, as we no longer have to do with simple arithmetics) G, which ourselves *can* conclude as true, but not the system itself. It is proved, therefore, that we are different from the specific machine - mind model (Lucas) or that our cognitive capacities overcome those of a computation through algorithm (Penrose).

Both scholars were heavily criticized for their propositions and for using Gödel's mathematical results. The historical course of the debate reveals that - apart from the changes observed in terminology once Penrose's argument appeared - arguments remained the same and Penrose's critics (Chalmers, 1996; Lindström, 2001; McCullogh, 1996; McDermott, 1996) were often, and sometimes absolutely, aligned with those of Lucas (Benacerraf, 1967; Chihara, 1972; Whiteley, 1962; George, 1962 respectively). The historical reappearance of the debate can be partially explained by the nature of the argument itself, which is based on the establishment of the truth of the sentence G. This establishment is up to the *human* and outside the computational system which G originates from.

Hence, supposing that someone believes that human cognitive abilities are not fully reduced to algorithmic processes, then the discovery will need to have come from that non-algorithmic part of his thought, confirming that the deduction power of the formal system or machine is not enough to validate the truth of G. On the other hand, if someone believes that the mind functions computationally, then the establishment of the truth of G needs to be attained inside the limits of a formal proof system. However, the attempt to formalize the entirety of the human proving capacities leads, through Gödel’s theorem, to the “best possible” result that “if we are computational systems, then we cannot be aware of our own function”; thus, we are not necessarily aware of something more than any other system (Benacerraf, 1967; Chihara, 1972). Observing the conclusions of both sides, each side seems to conclude exactly what it has presupposed as fact. Therefore, it seems that the debate’s problem is *structural* and, thus, any tension created is impossible to be resolved, in spite of arguments.

The above realization has been similarly made by other scholars researching the matter more holistically (Webb, 1968; King, 1996) and it seems that things are being led to a need of changing the fundamental intuition of what a computational process is and what a computer can do. When A. Turing attempted to formalize the notion of mechanical process (Turing, 1950), he invented the ‘digital computer’, the computational equivalent of the human computer that writes on infinite tape, following a strict set of typical rules. This formalization of the notion of computational function was rapidly accepted as the most successful attempt of formalizing the corresponding abstract idea, and proved to be equivalent to all other similar attempts, such as Gödel’s recursive functions or Church’s λ-calculus. If we further accept the Church – Turing Thesis, the models mentioned above will be proven to be equivalent to any other computational model, present or future. Such conditions would confine any computational system in the capacities of a

Turing Machine (TM), yet they would not legitimize the positions of CTM since they do not prevent any physical or artificial system from overpassing the abilities of a TM (Piccinini, 2007). Such systems exist, as of now, only in theory and the reason for being more powerful than a TM is the ability to solve non-computable problems!

What would it mean, though, for an artificial system - supposing a machine - to be able to solve a non-computable problem? It would simply mean that the machine would do an infinite number of calculations in finite time (Ord, 2002). That system would be able to respond to undecidable questions that concern Turing machines; would it, however, preserve the characteristics expected to be seen in a machine? (accuracy, absolutely predictable behavior, knowledge of function and more) And if not, would we equally and easily accept its characterization as a machine? Could it be, finally, that the mind is such a hyper-computer, and if this is true, what would the impact be on CTM? *Would it be preferable to have an analogy of mind and hyper-computer, under the knowledge that the former holds more than finite calculation abilities* ***or*** *a structural weakness of proving the equivalence of cognitive processes with the absolutely predetermined manipulation of logical symbols?*

So we arrive at a dilemma, one such as Gödel himself had, on his infamous Gibbs Lecture (Gödel, 1951). Gödel was arguing back then against intuitionism and for a platonic view of mathematics on the grounds of a disjunctive argument such as this: “either ... the human mind (even within the realm of pure mathematics) infinitely surpasses the power of any finite machine, or else there exist absolutely unsolvable diophantine problems”. Gödel’s intuition urged him to think of this as a very important argument for his view; and the question now is, is it possible, that the aforementioned *new* disjunctive argument be used effectively against CTM?

My personal (mathematical) intuition is yes, because adopting either thesis, leads to conclusions that the CTM cannot implement, in order to function as a theory. The first alternative clearly shows that, unless we forsake the TM computational model (along with the C-T Thesis), there will always be a barrier any computational system will never surpass, whereas the second shows that CTM will have to rely on faith (or intuition) about its fundamental thesis, since it will remain unprovable, something of a contradiction into the core of the theory.

P.S. One a final note, the above intuition may be taken as a choice of sides in the Gödelian argument, and the critics of the formal side may argue against it, in the grounds of absence of a formal proof. That’s understandable but please note and respect the disjunctive nature of the argument and its fundamental difference from the Lucas/Penrose argument, which is already established as one side on a fundamentally problematic debate.

References

Benacerraf, P. 1967. God, the Devil, and Gödel. *Monist* 51:9-32.

Chalmers, D.J. 1996. Minds, machines, and mathematics. *Psyche* 2:11-20.

Chihara, C. 1972. On alleged refutations of mechanism using Gödel's incompleteness results. *Journal of Philosophy* 64:507-26.

Davis, M. 1990. Is Mathematical Insight Algorithmic?

Davis, M. 1993. How Subtle is Gödel's Theorem. More on Roger Penrose.

Feferman, S. 1996. Penrose's Gödelian argument. *Psyche* 2:21-32.

George, F. 1962. Minds, machines and Gödel: Another reply to Mr. Lucas. *Philosophy* 37:62-63.

Gödel, K. 1931. On formally undecidable propositions of Principia Mathematica and related Systems I. *Collected Works vol. I.*

Gödel, K. 1946. Remarks Before the Princeton Bicentennial Conference. Gödel's note on mechanistic processes. Available at the tome '*The undecidable. Basic Papers on undecidable propositions, unsolvable problems and computable functions*'. Raven Presss, Hewlett, N.Y. 1965.

Gödel, K. 1951. Some basic theorems in the foundations of mathematics and their implications. Gibbs Lecture. *Collected Works vol. III.*

King, D. 1996. Is the human mind a Turing Machine? *Synthese,* 108: 379-389.

Lindström P. 2001. Penrose's New Argument. *Journal of Philosophical Logic* 30:241-250.

Lucas, J.R. 1961. Minds, machines and Gödel. *Philosophy* 36:112-127.

Lucas, J.R. 1968. Satan stultified: A rejoinder to Paul Benacerraf. *Monist* 52:145-58.

Lucas, J.R. 1971. Metamathematics and the philosophy of mind: A rejoinder. *Philosophy of Science* 38:310-13.

Lucas, J.R. 1996. The Gödelian Argument: Turn Over the Page. [a speech at the BSPS conference in Oxford, available at the URL: http://users.ox.ac.uk/~jrlucas/Gödel/turn.html].

McCullough, D. 1996. Can humans escape Gödel? *Psyche* 2:57-65.

McDermott, D. 1996. [Star] Penrose is wrong. *Psyche* 2:66-82.

Ord. T. 2002. Hypercomputation: computing more than the Turing machine, *Honours Thesis, University of Melbourne*.

Penrose, R. 1989. *The Emperor's New Mind*. Σελ. 129-146 & 538-541. Oxford University Press.

Penrose, R. 1996. Beyond the doubting of a shadow. *Psyche* 2:89-129.

Piccinini G. 2007. Computationalism, the church-turing thesis and the church-turing fallacy. *Synthese*, vol. 154, no1, pp. 97-120.

Putnam, H. 1960. Minds and Machines, (first published on) *Sidney Hook, ed., Dimensions of Mind. A Symposium*, New York, 1960. [Also available at ‘Mind, Language and Reality’ Philosophical Papers vol. 2 (CUP 1975), of the same].

Putnam, H. 1985. Reflexive reflections. *Erkenntnis* 22:143-153.

Turing, A.M. 1950. Computing Machinery and Intelligence. *Mind*: 433-460.

Webb, J. 1968. Metamathematics and the philosophy of mind. *Philosophy of Science* 35:156-78.

Whiteley, C. 1962. Minds, machines and Gödel: A reply to Mr. Lucas. *Philosophy* 37:61-62.