

**UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS SOCIAIS E HUMANAS
PROGRAMA DE PÓS-GRADUAÇÃO EM FILOSOFIA**

Pablo Rafael Rolim dos Santos

**ESTUDOS SOBRE A TEORIA COMPUTACIONAL DA MENTE: O
ARGUMENTO GÖDELIANO E O ARGUMENTO ANTI- SEMÂNTICO**

Santa Maria, RS
2017

Pablo Rafael Rolim dos Santos

**Estudos Sobre a Teoria Computacional da Mente: O Argumento Gödeliano e o
Argumento Anti-Semântico**

Dissertação apresentada ao Curso de Mestrado em Filosofia do Programa de Pós-Graduação em Filosofia, Área de Concentração em Filosofia Teórica e Prática, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do título de Mestre em Filosofia.

Orientador: Prof Dr. Frank Thomas Sautter
Coorientador: Prof Dr. Rogerio Passos Severo

Santa Maria, RS
2017

dos Santos, Pablo
Estudos Sobre a Teoria Computacional da Mente: O
argumento Gödeliano e o argumento Anti-Semântico / Pablo
dos Santos.- 2017.
56 p. ; 30 cm

Orientador: Frank Sautter
Coorientador: Rogério Severo
Dissertação (mestrado) - Universidade Federal de Santa
Maria, Centro de Ciências Sociais e Humanas, Programa de
Pós-Graduação em Filosofia, RS, 2017

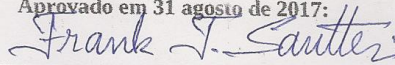
1. Filosofia da Mente 2. Computacionalismo 3. Gödel 4.
Searle I. Sautter, Frank II. Severo, Rogério III. Título.

Pablo Rafael Rolim dos Santos

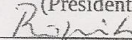
**ESTUDOS SOBRE A TEORIA COMPUTACIONAL DA MENTE: O
ARGUMENTO GÖDELIANO E O ARGUMENTO ANTI-SEMÂNTICO**

Dissertação apresentada ao Curso de Mestrado em Filosofia do Programa de Pós-Graduação em Filosofia, Área de Concentração em Filosofia Teórica e Prática, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do título de Mestre em Filosofia.

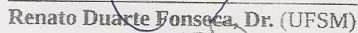
Aprovado em 31 agosto de 2017:

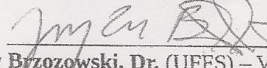


Frank Thomas Sautter, Dr. (UFSM)
(Presidente/Orientador)



Rogério Passos Severo, Dr. (UFRGS) – Videoconferência
(Coorientador)


Renato Duarte Fonseca, Dr. (UFSM)



Jerzy Byzozowski, Dr. (UFFS) – Videoconferência

Santa Maria, RS
2017

RESUMO

ESTUDOS SOBRE A TEORIA COMPUTACIONAL DA MENTE: O ARGUMENTO GÖDELIANO E O ARGUMENTO ANTI-SEMÂNTICO

AUTOR: Pablo Rafael Rolim dos Santos
ORIENTADOR: Frank Thomas Sautter
COORIENTADOR: Rogério Passos Severo

Esta dissertação analisa duas objeções contra a teoria de que a mente humana é um sistema computacional. No primeiro artigo analiso os argumentos baseado nos teoremas da incompletude utilizados por Kurt Gödel e John Randolph Lucas para defender a tese da superioridade: a tese de que a mente humana é superior a quaisquer sistemas computacionais. Adicionalmente a análise dos argumentos, aponto uma falha em um desses argumentos e forneço uma hipótese alternativa a ele. No segundo artigo, analiso o argumento anti-semântico de John Searle contra a teoria computacional da mente. Nesse argumento Searle defende que é impossível que sistemas computacionais tenham a capacidade de compreensão semântica, limitando-se apenas a manipulação de símbolos. Realizo a identificação das teorias alvo da crítica de Searle e aponto aparentes mal entendidos que esse comete em relação a tais teorias e como isso influencia em sua argumentação. Por fim, apresento uma tradução do artigo “Minds, machines and Gödel” de John Randolph Lucas, o mais influente nas discussões sobre a alegada impossibilidade da mente humana ser um sistema computacional fornecida pelos teoremas da incompletude.

Palavras-chave: Teoremas da Incompletude. Inconsistência. Kurt Gödel. John Randolph Lucas. Paraconsistência. Teoria computacional da mente. Searle. Argumento anti-semântico. Quarto chinês.

ABSTRACT

STUDIES ON THE COMPUTATIONAL THEORY OF MIND: THE GÖDELIAN AND ANTI-SEMANTICAL ARGUMENTS

AUTHOR: Pablo Rafael Rolim dos Santos

ADVISOR: Frank Thomas Sautter

CO-ADVISOR: Rogério Passos Severo

This dissertation analyzes two objections against the theory that human mind is a computational system. The first paper analyzes the arguments put forward by Kurt Gödel and John Randolph Lucas using incompleteness theorems to support the superiority thesis: the thesis that human mind is superior to any computational system ever built. Further, I point a flaw in one of those arguments and change it to an alternate hypothesis. The second paper analyzes John Searle's anti-semantical argument against computationalism. That argument states that it is impossible for computational systems to possess semantical understanding. Instead, they would only be able to manipulate meaningless symbols. It is identified the theories which Searle is aiming with his critics and are pointed out some seeming mistakes in his understanding of them. Also, are identified how those misunderstandings are influential in his argumentation. Finally, I present a translation of John Randolph Lucas's "Mind, machines and Gödel", one of the most influential paper in discussions about how the incompleteness theorems allegedly makes impossible that human mind were a computational system.

Keywords: Incompleteness theorems. Inconsistency. Kurt Gödel. John Randolph Lucas. Paraconsistency. The computational theory of mind. John Searle. Anti-semantical argument. Chinese Room.

Sumário

INTRODUÇÃO.....	8
ARTIGO 1 – THE SUPERIORITY THESIS AND THE ANTI- INCONSISTENCY OF MIND ARGUMENT: AN ANALYSIS OF THE GÖDELIAN ARGUMENT AGAINST COMPUTATIONALISM.....	11
1. Mechanicism and computationalism.....	11
2. The superiority thesis.....	12
2.1 Gödel’s disjunction.....	12
2.2 Lucas’s argument.....	13
3. Loosening the rules of logics: Formal systems and tolerance....	16
4. The Paraconsistent mind hypothesis: an alternative to the anti- consistency mind argument.....	18
5. Conclusion.....	19
References.....	19
Artigo 2 – SEARLE’S OBJECTION TO THE COMPUTATIONAL THEORY OF MIND: SOME SEEMING MISTAKES.....	21
1. the computational theory of mind.....	21
2. Searle’s reasons to reject computationalism.....	24
3. The anti-semantical argument and the critical mistakes about the mind.....	24
4. Rehabilitating the system and robot replies: some hypotheses..	29
5. Conclusion.....	33
Reference.....	33
TRADUÇÃO – MENTES, MÁQUINAS E GÖDEL.....	35
DISCUSSÃO.....	49
CONCLUSÃO.....	49
REFERÊNCIAS.....	50

INTRODUÇÃO

A presente dissertação analisa duas objeções à teoria computacional da mente: as limitações metamatemáticas alegadamente impostas pelos teoremas da incompletude (GÖDEL, 1992; GÖDEL, 1995; LUCAS, 1961) e a alegada insuficiência dos processos computacionais para proporcionar o fenômeno semântico (SEARLE, 1980). A teoria computacional da mente (ou computacionalismo) defende que cognição é computação. Em sua versão forte, a tese afirma que todos os fenômenos cognitivos podem ser explicados unicamente por computações. Já em sua versão mais fraca, a tese é que a computação é um elemento essencial em todos ou na maioria dos processos cognitivos. Assim, essa posição apresenta-se como a que melhor se adequaria ao programa computacionalista na ciência, visto que nele as explicações usualmente oferecidas às tarefas cognitivas são formuladas em termos de computações realizadas em representações. Além de uma pressuposição prévia de como tais representações adquirem seu conteúdo e, para casos de estados conscientes, assumindo adicionalmente uma explicação de como esses tornar-se-iam conscientes. Por vezes, ainda, tais explicações são formuladas em termos de estados e processos que uniriam o corpo do organismo a seu ambiente, fornecendo explicações de como computações neurais ocorreriam ligando esses dois elementos (PICCININI, 2010).

Entretanto, um erro recorrente e que é de grande importância se evitar é atribuir um modelo computacional específico como elemento necessário da tese computacionalista. Pois, contrariamente a isso, o computacionalismo é uma tese neutra em relação a qual modelo computacional a mente humana efetivamente implantaria. Um erro do tipo mencionado seria, por exemplo, pressupor que a cognição humana implantaria computações através de um modelo ao estilo de uma máquina de Turing e, ao objetar esse modelo, acreditar se estar demonstrando a falsidade da tese computacionalista. Frisando, o essencial à teoria computacional da mente é a tese de que a cognição é computação, independente de qual modelo seja implementado. Na história da filosofia algumas posições computacionalistas tornaram-se notáveis, cada uma com uma tese distinta de como o cérebro humano estaria realizando computações. A título de conhecimento essas são: o *funcionalismo de máquina* (PUTNAM, 1975; PUTNAM, 2002), a *teoria representacional da mente* (FODOR, 1975), e o *conexionismo* baseado nos trabalhos em redes neurais artificiais de McCulloch e Pitts (1943). A última posição é igualmente alvo de enganos corriqueiros, pois, ao defender que fenômenos cognitivos devem ser explicados (ao menos em parte) por processos de redes neurais, ela é tomada como uma posição anticomputacionalista. Entretanto, frisando novamente, o computacionalismo não seria rejeitado através dessa posição visto que ela não exclui que o sistema nervoso realize computações. Apenas defende um modelo específico de como essas estariam sendo implementadas.

Entretanto, o tema principal deste trabalho não são os principais elementos e méritos das distintas posições computacionalistas. Pelo contrário, aborda-se duas objeções que pretendem minar a tese computacionalista. Em primeiro lugar analiso o argumento de John Randolph Lucas (com incursões em Kurt Gödel) – que aqui chamo de *argumento da superioridade* – que afirma que a mente humana é superior a *quaisquer* sistemas computacionais devido a certas limitações que esses possuiriam e dos quais mente humana não parece sofrer. Como elemento central desse argumento temos os resultados metamatemáticos obtidos por Kurt Gödel (1992) e conhecidos como *teoremas da incompletude*. Principalmente o segundo desses que é expresso informalmente por Gödel como

Em qualquer sistema bem definido de axiomas e regras [...] a proposição declarando a consistência desses (ou melhor, a proposição número-teorética equivalente) é indemonstrável a partir desses axiomas e regras caso esses sejam consistentes e suficientes para derivar uma certa porção da aritmética finitista dos inteiros (GÖDEL, 1995, p. 308)¹.

Em tais teoremas Gödel demonstra tanto que sistemas consistentes onde seja possível expressar a aritmética básica são necessariamente incompletos quanto que caso um sistema desses seja realmente consistente, é impossível demonstrar sua consistência através de seus axiomas. Em termos gerais podemos dizer informalmente que a forma utilizada por Gödel para chegar a tais resultados é a construção de uma fórmula que faz uso da linguagem do sistema, e a qual nós podemos ver que é verdadeira em tal sistema, porém que não é possível de ser derivada a partir dos axiomas desse sistema. Assim, é possível demonstrar a existência de fórmulas que ainda que verdadeiras, se o sistema for consistente e capaz de expressar a aritmética básica, não são deriváveis de seus axiomas².

Como já mencionado, é feita uma curta incursão na discussão levantada por Gödel sobre a questão através da *Disjunção de Gödel* (GÖDEL, 1995, p. 304) na qual esse realiza a distinção da matemática em dois sentidos: o *objetivo* da matemática é entendido como as verdades da matemática em sentido absoluto; e o sentido *subjetivo* é compreendido como todas as verdades possíveis de serem demonstradas por humanos. Sua argumentação é que se ambos sentidos coincidirem isso significaria que nenhum sistema formal axiomático (ou máquina de Turing) poderia jamais possuir todas as potencialidades de matematização da mente humana, e assim a tese computacionalista se mostraria falsa. Já no caso de tais sentidos não coincidirem, isso mostraria que existem problemas matemáticos absolutamente insolúveis. Já o argumento de Lucas – o principal em análise nesse trabalho – oferece argumentos em defesa de um dos disjuntos de Gödel,

¹Por axiomas e regras suficientes para a “aritmética finitista dos inteiros” Gödel refere-se ao sistema PA dos axiomas de Peano para a aritmética dos números naturais (GÖDEL, 1995, p. 308) .

² Caso tal fórmula seja adicionada como axioma em uma extensão desse sistema, é possível construir uma nova fórmula igualmente verdadeira e igualmente não derivável na versão estendida.

especificamente a defesa da já citada *tese da superioridade*, além da formulação de um argumento que demonstraria que a mente humana não pode ser um sistema inconsistente, o qual chamo de “anti-inconsistency of mind argument” e sugiro possíveis falhas. Em geral, temos que caso os argumentos baseados na incompletude citados sejam corretos, a teoria computacional da mente mostrar-se-ia falsa.

A segunda objeção analisada é o *argumento semântico* de Searle (1980) contra a teoria computacional da mente. Essa objeção é que mentes humanas não podem ser sistemas computacionais pois enquanto mentes possuem estados cognitivos como o da compreensão de significados, sistemas que apenas manipulam símbolos seriam destituídos dessa capacidade. Visto de um ponto de vista ligeiramente distinto, a objeção é que sistemas computacionais jamais poderão possuir e estar cientes do conteúdo para os símbolos que manipulam. Mesmo que, por vezes, humanos tendam a atribuir estados cognitivos a esses sistemas devido ao comportamento que eles exibem. O ponto de Searle é que ainda que sistemas computacionais possam apresentar capacidades de input e output que sejam análogos ao comportamento de um humano, como por exemplo ser falante da língua chinesa no experimento de pensamento utilizado pelo autor, ainda assim esse sistema não seria capaz de compreender chinês independentemente de como esse for programado (SEARLE, 1980).

Neste trabalho, além de iniciar a identificação e classificação dos argumentos que utilizam-se dos teoremas da incompletude sob a forma da *tese da superioridade*, ainda aponto uma aparente falha em umas das premissas utilizadas por Lucas para sua defesa (a qual denomino “The anti-inconsistency of mind argument”) e proponho uma nova hipótese em seu lugar a qual faria uso de lógicas não clássicas como subjacentes à mente humana. A qual denomino “The inconsistency of mind hypothesis”. Já no segundo artigo, busco identificar alguns aparentes erros cometidos na construção do argumento do quarto chinês devido ao seu caráter primário de persuasão e não argumentativo. Subsequentemente analiso as consequências desses aparentes erros para nos contra-argumentos oferecidos por Searle a sua tese de que processos computacionais não podem proporcionar estados cognitivos com conteúdo semântico.

O último componente da dissertação é uma tradução do artigo de Lucas que impulsionou a discussão sobre a tese da superioridade. Como dito anteriormente, essa não parece ter sido a primeira apresentação dessa tese, entretanto historicamente foi a que obteve maior atenção e devido a isso sua tradução torna-se de grande valor para o debate em língua portuguesa.

ARTIGO 1 – THE SUPERIORITY THESIS AND THE ANTI-INCONSISTENCY OF MIND ARGUMENT: AN ANALYSIS OF THE GÖDELIAN ARGUMENT AGAINST COMPUTATIONALISM

Pablo Rolim

Abstract

The *superiority thesis* states that no machine can ever have a mind. Among the premises given to support it is what I call the “anti-inconsistency of mind argument” endorsed by J. R. Lucas. In this paper I analyze Gödel’s disjunction – an earlier presentation of superiority thesis – and Lucas’s arguments supporting the *superiority thesis* and propose an objection to the *anti-inconsistency of mind argument*. First I give a general description of the computational theory of mind and its relation to mechanicism. After that, I introduce Gödel’s disjunction and Lucas’s arguments against it and finally I propose the *Paraconsistent mind hypothesis* as an alternative to the debatable thesis.

1. MECHANICISM AND COMPUTATIONALISM

The term Mechanicism was used around the seventeenth century to name the thesis that matter obeys mechanical causal laws. Among those who held a version of it was Descartes in his work *Le Monde* where he sought to explain a wide range of phenomena of the world by means of conservation of initial motion and contact action (CRAVER; TABERY, J, 2015, sec 2). Also, it seems to appear as the reason for Descartes’s *Meditations* that sought to replace Aristotelian physics. Albeit he was a mechanist, Descartes had a dualistic view about the mind and thought it doesn’t have anything to do with matter. Contemporary versions of mechanicism are more prone to accept that minds also obey mechanical laws (CRAVER; TABERY, 2015). A way to understand the human mind as being governed by mechanical laws is seeing it as a computational machine. Through this view mechanicism becomes a strand of computationalism.

The computational theory of mind (or computationalism) is the thesis that cognition is computation. The thesis has two corollaries: it is possible to build computers that can think and the human mind is a kind of computer (DIETRICH, 1994). Being a computational system, the implementation of mind should be taken as possible to be done in more than one substrate. It could be, for example, implemented either in silicon chips or flesh and blood. The research motivated by computationalism was divided in at least two distinct programs: while the role of understanding how mind computes was taken by cognitive science, the project of building a thinking computer was taken by the field of artificial intelligence. Yet, over the years, various arguments have been

raised against computationalism. In this paper I analyze a very influential one: the *superiority thesis* put forward by Kurt Gödel (1995) and John Randolph Lucas (1961).

2. THE SUPERIORITY THESIS

The *Superiority thesis* states that human mind is superior to any machine ever build. Both Gödel (1995, p. 304) and Lucas (1961) hold some version of it on the basis of the *incompleteness theorems* (GÖDEL, 1992), a pair of metamathematical results concerning the limits of provability in formal systems³. The first theorem, roughly speaking, shows that for every consistent formal axiomatic theory in which basic statements about arithmetic can be formulated (*e.g.* Peano arithmetic), there is at least one statement which can be formulated in the language of the theory but can neither be proved nor disproved on this theory, that is, it is *undecidable* (FRANZÉN, 2004, p. 1). The second theorem, also roughly speaking, shows that for a wide class of such theories T, if the system is *consistent*, its consistency cannot be proved using only the axioms of the theory T itself (FRANZÉN, 2004, p. 1). These arguments aims to refute the *mechanicist* thesis that human mind is, or can be accurately modelled, as a digital computer or Turing machine (SHAPIRO, 1998, p. 273)⁴.

2.1 Gödel's disjunction

Gödel approaches the question through a disjunction concerning the powers of human thought. The first disjunct makes an statement about the potentialities of human mind, while the second imposes a restriction to what can be known by humans

Either the human mind surpasses all machines (to be more precise: it can decide more number theoretical questions than any machines), or else there exist number-theoretical questions undecidable for the human mind [it is not excluded that both alternatives may be true.] (WANG, 1996, p. 185)⁵.

Briefly, the argument goes as follows: assume that human mind is actually a computational machine. That is, the human mind is an algorithm. Now considering the mathematical capacity of it, this algorithm should be able to produce all the mathematical theorems that the human mind is capable of producing. By the Church-Turing thesis we known that every algorithmically computable function is computable by a Turing machine and that is the same to say that all humanly

³It should be noted that Alan Turing had already noted the possibility of applying incompleteness theorems against the thesis that it is possible to build an artificial mind. That was called the "Mathematical objection" in his "Computing machine and intelligence" (1950, p. 444-445).

⁴As Shapiro has also noted, the definition of mechanism in these arguments is way more imprecise than what would be expected for attack with something as sharp and precise as the incompleteness theorems (SHAPIRO, 1998, p. 275). Yet, it is the one that anti-mechancists assume.

⁵An early version was given in (GÖDEL, 1995, p. 310).

knowable theorems can be recursively axiomatized in some formal theory T . This theory, then, has to be consistent. But now comes into play Gödel's second incompleteness theorem.

For any well defined system of axioms and rules, in particular, the proposition stating their consistency (or rather the equivalent number-theoretical proposition) is indemonstrable from these axioms and rules, provided these axioms and rules are consistent and suffice to derive a certain portion of the finitistic arithmetic of integers (GÖDEL, 1995, p. 308-309)⁶

This theorem states that any consistent theories capable of express some elementary arithmetic⁷ cannot prove their own consistency even if they can express it via number coding. So, there should be a mathematical proposition ϕ that cannot be decided by T . But given that T is supposed to capture what is humanly provable, this proposition is humanly (and absolutely) undecidable (HORSTEN; WELCH, 2016, p. 2). That's it, if human mind were a computational system there should be some undecidable problems in mathematics.

Gödel did not have a way to prove any of its disjuncts, although he strongly supported the thesis that human mind is superior than any machine ever build and that we are able to solve any mathematical problem.

The philosophical conclusion that Gödel drew from his disjunction is that: if the first disjunct were true then the working of human mind couldn't be reduced to the working of the brain which he construed as a finite machine with a finite number of parts. So, some non-mechanicist alternative should be considered. However, if the second disjunct were true, it seems to to disprove the thesis that mathematics is a human creation, because if there were absolutely unsolvable problems of mathematics, Gödel believes, it could be the case that human mind has created those problems given that in his credence, the creator necessarily knows all properties of his creation given that it couldn't have any property that wasn't given to them (GÖDEL, 1995, p. 311)⁸.

2.2 Lucas's argument

Lucas also uses incompleteness theorems as an objection against mechanicism. It is an argument for the first disjunct of Gödel's disjunction. That is, that human mind isn't a computational system. Much of his arguments are similar to those of Gödel, although, Lucas is way more straightforward: he argues that if mechanicism were true then human mind had to be a kind of machine, and being any physical computational machine an instance of a formal system, that machine should be subject to the same metamathematical constraints that the corresponding formal system is, among which, Gödel's incompleteness theorems.

⁶Again, by axioms and rules sufficient for the 'finitistic arithmetic of integers' Gödel understand the system of Peano Arithmetic PA (GÖDEL, 1995, p. 308, footnote 10).

⁷For instance, the system PA of Peano arithmetics.

⁸Gödel's mathematical realism isn't addressed on this paper. To a discussion of it see Parsons (1995).

Gödel's theorem must apply to cybernetical machines, because it is the essence of being a machine that it should be the instantiation of a formal system. It follows that given any machine which is consistent and capable of doing simple arithmetic, there is a formula which it is incapable of producing as being true - i. e., the formula is unprovable-in-the-system - but which we can see to be true. It follows that no machine can be an adequate model of the mind, that minds are essentially different from machines (LUCAS, p. 113, 1961)⁹

Lucas's argument has two halves. The first half is the same statement present on the first of Gödel's disjunction – the *superiority thesis*. That is, it tries to establish human mind as superior than any machine. The second is the argument against the possibility of the human mind being an inconsistent system – what I call the *anti-inconsistency of mind argument*. The first half can be stated as follows: (a) any consistent system strong enough to formalize Peano arithmetic is subject to Gödel's incompleteness theorems; (b) if human mind were a consistent system capable of formalizing Peano arithmetic, it would be subject to Gödel's incompleteness theorems; (c) nevertheless, human mind can do things that consistent formal systems subject to Gödel incompleteness theorems cannot do; (d) therefore, human mind isn't a consistent formal system constrained by Gödel's incompleteness theorems.

The *superiority thesis* is held (again) on the basis of the second incompleteness theorem. The one who states that in a consistent formal system the proposition stating its own consistency (or the number-theoretical proposition equivalent to it) cannot be derived from the axioms of the system. Just to remember, a *consistent* system is one in which we can't derive from its axioms both a theorem and its negation. What is called *Principle of non-contradiction* (PNC). Though, if the system is consistent it is not possible to derive from its axioms the statement that guarantees that the system does not prove a proposition and its negation, *i.e.* a contradiction. A further feature of consistent systems is that, as it cannot prove contradictions, the only derivable statements possible in it are the true ones. So, if some statement is derivable from the axioms of the system by its rules of inference this should be a true statement. Putting those things together we have that a) consistent systems only derives true statements and b) there is one true statement that, if the system were really consistent, cannot be derivable in it: the one stating the consistency of the system itself. This is the loophole exploited by Lucas in (c): the human mind is capable of seeing the truth of statements of consistency in formal systems producing exactly the statement that cannot be derivable in the system – or the “Gödel sentence” of the system. That, argues Lucas, is the kind of thing that wouldn't be possible to human minds if they were a machines, because no machine can derive that kind of statement from its axioms

[...] For every machine there is a truth which it cannot produce as being true, but which a mind can. This shows that a machine cannot be a complete and adequate model of the

⁹As we saw, Gödel has already made the same point.

mind. It cannot do *everything* that mind can do, since however much it can do, there is always something which it cannot do, and a mind can (LUCAS, 1961, p. 115).

The other half of the argument the rejection that that human mind is inconsistent: the anti-inconsistency of mind argument, although, as Lucas himself notices, this is a presupposition that he cannot prove

Gödel's theorem applies only to consistent systems. All that we can prove *formally* is that *if* the system is complete, then the Gödelian formula is unprovable-in-the-system. To be able to say categorically that the Gödelian formula is unprovable-in-the-system, and therefore true, we must say that it is consistent. And, as Gödel showed in his second theorem – a corollary of his first – it is impossible to prove in a consistent system that that system is consistent. Thus in order to fault the machine by producing a formula of which we can say both that it is true and that the machine cannot produce it as true, we have to be able to say that the machine (or, rather, its corresponding formal system) is consistent; and there is no absolute proof of this. All we can do is to examine the machine and see if it appears consistent. There always remains the possibility of some inconsistency not yet detected. (LUCAS, 1961, p. 120).

Even being formally unprovable, Lucas argues that we still have reasons to think that we are not inconsistent. The main idea is that we are not inconsistent systems because if we were, it would make us do not bother when confronted with statements of a sentence and its negation

The fact that we are all sometimes inconsistent cannot be gainsaid, but from this it does not follow that we are tantamount to inconsistent systems. Our inconsistencies are mistakes rather than set policies. They correspond to the occasional malfunctioning of a machine, not its normal scheme of operations. Witness to this that we eschew inconsistencies when we recognize them for what they are. If we really were inconsistent machines, we should remain content with our inconsistencies, and would happily affirm both halves of a contradiction. Moreover, we would be prepared to say absolutely anything – which we are not. It is easily show that in an inconsistent formal system everything is provable, and the requirement of consistency turns out to be just that not everything can be proved in it [...] human beings, although not perfectly consistent, are not so much inconsistent as fallible (LUCAS, 1961, p. 121).

Lucas's anti-inconsistency of mind argument can be stated as follows: (i) an inconsistent system can derive anything; (ii) if an inconsistent system can derive anything, in particular, it can derive a contradiction; (iii) if human mind were inconsistent it would be able to state anything, in particular to state a contradiction (iv) Nevertheless, humans don't state whatsoever, in special if recognized, they abstain to (instead of) state a contradiction; (v) being so, human mind is not inconsistent, but fallible. Statement (i) describes the property of *inconsistency*. Let L be a language, an arbitrary set of formulas Σ from L is *inconsistent* if given a formula α we have $\Sigma \vdash \alpha$ and $\Sigma \vdash \neg\alpha$. That is, if any formula and its negation are derivable from a set of the system, then the set (and by extension the whole system) is *inconsistent*. On other hand, as said before, a *consistent* system is one at which not every formula is derivable. The path from inconsistency to contradiction is given as follows: given that an inconsistent system can derives anything, a corollary is that it can derive a contradiction $\alpha \wedge \neg\alpha$ as stated at (ii). The next step in Lucas's argument is a *reductio ad absurdum*. Lucas presupposes the inconsistency of human mind at (iii) and shows that if human

minds were inconsistent then humans would state contradictions; he proceed to give support against it at (iv) by means of the resistance that humans impose to state recognized contradictions; and thus concludes from it that humans aren't inconsistent systems (v). However, a way to refute this argument would be showing how humans could be inconsistent system and yet doesn't state whatever proposition, including contradictions.

3. LOOSENING THE RULES OF LOGICS: FORMAL SYSTEMS AND TOLERANCE

Carnap's *principle of tolerance* states that one should be free to use any formal system, whatever its axioms and rules of inference, when studying the logic of science. This principle looses the restrictions of what kind of logics could be applied in ones research

In logic, there are no morals. Everyone is at liberty to build up his own logic, i.e. his own form of language, as he wishes. All that is required of him is that, if he wishes to discuss it, he must state his methods clearly, and give syntactical rules instead of philosophical arguments (CARNAP, 1937, p. 52)

In the same spirit, da Costa's *principle of tolerance in mathematics* states that those formal systems worth of investigation are the *non-trivial* ones, rather than the *non-contradictory*: "From the syntactical-semantical standpoint, every mathematical theory is admissible unless it is trivial (da COSTA, 1959 APUD CARNIELLI; MARCOS, 2001, p.3)". As presented in the following sections, *triviality* and *contradictoriness* are different features of a formal system that, although tied in some of them, do not have to be necessarily so. If one took a position aligned with these two principles, it would be possible to hypothesize that human mind could be an inconsistent yet non-trivial system given that one is not constrained by the kind of logic one should use to investigate the mind. If the resulting explanation does better than Lucas's *anti-inconsistency of mind argument* in explaining human mind functioning in the presence of contradictions, it could be a hint to pursue that hypothesis.

Lucas's argument presupposes classical logic as the underlying logic of human mind. Among the main characteristics of a classical system is the principle of explosion (PE) or *ex contradictione quodlibet*, i.e. the rule that assert that wherever we have a contradiction, anything goes. The main problem with it is that if "anything goes", we could prove anything when confronted with a contradiction and this would *trivialize* the system because we could get any formula whatsoever. Given that, when framed in classical fashion, a virtue of a system would be the capacity to avoid that contradictions could arise in it. That is, the property of *consistency*

If we really were inconsistent machines, we should remain content with our inconsistencies, and would happily affirm both halves of a contradiction. Moreover, we would be prepared to say absolutely anything - which we are not. It is easily shown that in an inconsistent formal system everything is provable, and the requirement of consistency turns out to be

just that not everything can be proved in it - it is not the case that “anything goes”. This surely is a characteristic of the mental operations of human beings: they are selective: they do discriminate between favored - true - and unfavored - false - statements: when a person is prepared to say anything, and is prepared to contradict himself without any qualm or repugnance, then he is adjudged to have “lost his mind” (LUCAS, 1961, p. 121).

Lucas thesis could be revised given our current trends in logic. Lucas state things as “it is easily show that in an inconsistent system everything is provable” and “if we really were inconsistent machines, we should remain content with our inconsistencies and would happily affirm both halves of a contradiction”. Although now we have new research program devoted to deal with inconsistent yet non trivial systems of logic. And even some that would accept that *some* contradictions are true (PRIEST, 2002).

To better understand the some main distinctions among classical logic and Paraconsistent logic – which is of main interest in the present work – we can distinguish three principles assumed by distinct logical systems: *Principle of Non-contradiction*, *Principle of Non-Triviality* and *Principle of Explosion*. The principle of Non-contradiction states that it isn't possible to a given logic to derive an specific theorem and its negation from all of its theories (*i.e.* sets of sentences built using that logic system). That is, it isn't possible to derive a contradiction A and $\neg A$ in all its theories. The Principle of Non-Triviality states that in at least one theory of that logic, there is a formula which isn't derived by that theory. That is, it is not the case that for all theories of that given logic, any formula whatever could be derived from its axioms. And, finally, the Principle of Explosion states that states that a theory is explosive if the addition of a contradiction is sufficient to make it trivial. And, if that principle range over all theories, that logic is called explosive. Classical logical systems, as the one Lucas has in mind on his argument not just accept all of those principles but also equates triviality with contradictoriness through explosiveness. Other logical systems the Paraconsistent ones try to allow contradictions locally without trivializing the whole system. The way to achieve that is weakening the Principle of Explosion in a way that does not cause any theory to trivialize in face of a contradiction (CARNIELLI; MARCOS, 2002, p. 4).

This is not the place for the recalcitrant analysis of the structure of those logics¹⁰ although it could be stated that the family of LFIs (Logics of Formal Inconsistency) deals with explosiveness formulating the metatheoretical properties of consistency and inconsistency in its object language, introducing a non-classical negation and controlling the explosiveness of those theories which present contradictions. The resulting system, although still assuming the Principle of Non-contradiction, does not lead to trivialization. That is, if we work with a system of this kind as the underlying logic for the mind, the mind could well be an inconsistent system and yet not have the problems pointed by Lucas's anti-inconsistency of mind argument. Still, we could formulate a sketch of the argument.

¹⁰For the technical presentation the reader see (CARNIELLI; W CONIGLIO, M; MARCOS, J., 2007).

4. THE PARACONSISTENT MIND HYPOTHESIS: AN ALTERNATIVE TO THE ANTI-CONSISTENCY MIND ARGUMENT

If minds were taken as consistent formal systems, Lucas's argument seems to fit. Although, if my following argument is sound it could be an alternative to the second half of Lucas's argument – the anti-inconsistency of mind argument: (α) A paraconsistent does not equate contradictoriness with triviality; (β) If a paraconsistent system doesn't equate contradictoriness with triviality, then it isn't the case that from a contradiction everything follows; (γ) If anything follows from a contradiction, this system is called trivial; (δ) given that it is not the case that anything follows from a contradiction in a paraconsistent system, although it is inconsistent, it isn't trivial; (ε) if a paraconsistent system isn't trivial, then it can handle contradictions; (ζ) if the human mind were a paraconsistent system, then given a contradiction, it is not the case that anything follows (η) if human mind were paraconsistent, given that it isn't trivial, then it wouldn't state whatsoever from a contradiction (θ) So, given the ability to handle contradictions, the human mind could be an inconsistent but not trivial system; (ι) if the previous premises are true, then Lucas's argument is wrong and it is possible (although not a straightforwardly consequence) that human mind has an underlying paraconsistent logic.

The target point in Lucas's argument against the inconsistency of human mind is the working at (iv) of one of the main engines of classical logic: the *ex contradictione sequitur quodlibet* (from the contradiction everything follows) or *principle of explosion*. This principle establishes that the relation of consequence in classical logic is *explosive*. A system with an explosive relation of consequence has among its valid inferences the relation $\alpha, \neg\alpha \vdash \beta$ i.e., given a contradiction $\alpha \wedge \neg\alpha$ we infer β for any formula β whatsoever. However, it seems that the principle of explosion wasn't ubiquitous in the history of logic. Its acceptance as a valid form of inference is, allegedly, a feature of the mathematization of logic made by Boole, Frege, Russell, and Hilbert at the end of the 19th century (PRIEST, 2015, section 1.2)¹¹. On the other hand, a *paraconsistent system*, the main element of (α), is a formal system that challenges the *principle of explosion*. At such systems we cannot equate contradiction plus consistency with triviality. My main approach to paraconsistency is through the Logics of Formal Inconsistency (LFIs). These are a class of paraconsistent logics which can internalize the metatheoretical notions of consistency and inconsistency on its object language in a way that the classical logic could be reconstructed from them (CARNIELLI; CONIGLIO; MARCOS, 2007, p. 2). The main strategy underlying LFIs is to divide the world between consistent propositions and non-consistent propositions. If we are dealing with consistent propositions, these are to be subject to classical logic and then if a theory Γ has a

¹¹ I don't endorse any position about its status previous to 19th century. The rightness of Priest's statement is not influential to my main argument.

pair of contradictory propositions $\alpha, \neg\alpha$ it will explode if α or $\neg\alpha$ is a consistent proposition. These are market in the system as $\alpha\alpha$ (or $\alpha\neg\alpha$).

5. CONCLUSION

Although the *superiority thesis* has still alive and kicking, it is possible to show that the *anti-inconsistency of mind argument* is vulnerable if one confronts it with paraconsistent systems. How much it can damage the superiority thesis (if at all) still has to be evaluated, although some progress has been made showing a debatable premise.

REFERENCES

CARNAP, R. **Logical syntax of language**. Psychology Press, 1937.

CARNIELLI, W A; MARCOS, J. A taxonomy of C-systems. In: CARNIELLI, W; CONIGLIO, M; D'OTAVVIANO, I. (eds) **Paraconsistency: the logical way to the inconsistent**. New York: Marcel Dekker Inc, 2002.

CARNIELLI, W; CONIGLIO, M; MARCOS, J. Logics of formal inconsistency. In: GABBAY, D; GUNTHER, F (eds). **Handbook of Philosophical Logic**. 2ed. Dordrecht: Springer, v. 14, p. 1-93, 2007.

CRAVER, C.; TABERY, J. Mechanisms in Science. In: Edward N. Zalta, E. (ed.) **The Stanford Encyclopedia of Philosophy**, spring, 2017. *Disponível em:*
<<https://plato.stanford.edu/archives/spr2017/entries/science-mechanisms/>>

CHURCH, A. Review: A.M. Turing. On computable numbers, with an application to the Entscheidungsproblem. **The Journal of Symbolic Logic**, v. 2, n.1, p. 42-43, 1937.

DIETRICH, E (ed). **Thinking computers and virtual persons: essays on the intentionality of machines**. San Diego: Academic Press, 1994.

FODOR, J. **The language of thought**. New York: Thomas Y. Crowell, 1975.

FRANZÉN, T. **Inexhaustibility: a non-exhaustive treatment**. A K Peters, Ltd: Wellesley, 2004.

GABBAY, D. What is a logical system? In: Gabbay, M. (ed) **What is a logical system?** Oxford: Clarendon Press, 1994, p. 179 – 216.

GABBAY, D; GUENTHER, F. (eds). **Handbook of philosophical logics**. 2ed, v. 14, p. 1-93, 2007.

GÖDEL, K. Some basic theorems on the foundations of mathematics and their implications. In: FERERMAN (ed). **Kurt Gödel collected works volume III: Unpublished essays and lectures**. Oxford: Oxford University Press, p. 304-423, 1995.

GÖDEL, K. **On Formally undecidable propositions of Principia Mathematica and related systems**. New York: Dover Publications, 1992.

- GÖDEL, K. What is Cantor's continuum problem? In: BENACERRAF, P; PUTNAM, H. **Philosophy of mathematics: Selected reading**. Cambridge: Cambridge University Press, p. 470-485, 1983.
- HOLSTEN, L; WELCH, P. **Gödel's disjunction: The scope and limits of mathematical knowledge**. Oxford: Oxford University Press, 2016.
- LUCAS, J. Minds, machines and Gödel. **Philosophy**, p. 112-127, 1961.
- MCCULLOCH, W; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **Bulletin of mathematical biophysics**, v. 7. p. 115-133.
- MEYER, K. Relevant Arithmetic. **Bulletin of the section of logic**, Polish Academy of Sciences, p. 133-137, 1976.
- NELSON, D. Negation and separation of concepts in constructive systems. In: Heyting (ed). **Constructivity in Mathematics**. Amsterdam: North Holland Publishers, p. 208-225, 1959.
- PARSON, C. Platonism and mathematical intuition in Kurt Gödel's thought. **The Bulletin of Symbolic Logic**. v. 1m n. 1, p. 44-74, 1995.
- PRIEST, G. **In contradiction**. 2nd ed. Oxford: Oxford University Press, 2006.
- PRIEST, G. Paraconsistent logic. In: GABBAY, D; GUENTHNER, F (eds). **Handbook of Philosophical Logic**. Dordrecht: Springer, v. 6, p. 287-393, 2002.
- PRIEST, G; TANAKA, K; WEBER, Z. Paraconsistent Logics. In: Zalta, E. (ed) **The stanford encyclopedia of philosophy**, spring, 2015. Disponível em: <<http://plato.stanford.edu/archives/spr2015/entries/logic-paraconsistent/>>.
- OPPY, G; DOWE, D; The turing test. In: Zalta, E (ed). **The Stanford Encyclopedia of Philosophy**, spring, 2016. Disponível em: <<http://plato.stanford.edu/archives/spr2016/entries/turing-test/>>.
- RAATIKAINEN, P. Gödel's Incompleteness Theorems. In: Zalta, E (ed). **The Stanford Encyclopedia of Philosophy**, spring, 2015. Disponível em: <<https://plato.stanford.edu/archives/spr2015/entries/goedel-incompleteness/>>.
- SHAPIRO, S. Incompleteness, mechanism, and optimism. **The Bulletin of Symbolic Logic**, v. 4, n. 4, Sept. 1998.
- SMITH, P. **An introduction to Gödel's theorems**. 2nd. Cambridge: Cambridge University Press, 2013.
- TURING, A. Computing machinery and intelligence. **Mind**. v. 59, n. 236, p. 433-460, Oct. 1950.
- WANG, H. **A logical journey: From Gödel to Philosophy**. Cambridge: The MIT Press, 1996.

ARTIGO 2 – SEARLE’S OBJECTION TO THE COMPUTATIONAL THEORY OF MIND: SOME SEEMING MISTAKES

Pablo Rolim

Abstract

In this paper I analyze John Searle’s anti-semantical argument against the computational theory of mind given in his paper “Minds, brains, and programs”. Much has been said about it in the last decades, although I think some points could still be enlightened. My goal is not given an even further purportedly “knock down” argument, instead I do analyse some possible mistakes in both the Chinese room thought experiment which allegedly supports Searle’s *anti-semantical* objection to computationalism, and in his counterargument to the System and Robot replies.

1. THE COMPUTATIONAL THEORY OF MIND

The computational theory of mind is the thesis that cognition is computation. The thesis has two corolaries: it is possible to build computers that can think and the human mind is a kind of computer (DIETRICH, 1994). Being a computational system, the implementation of mind should be taken as possible to be done in more than one substrate. It could be, for example, implemented either in silicon chips or flesh and blood. The research motivated by computationalism was divided in at least two distinct programs: while the role of understanding how the mind computes was taken by the cognitive science, the project of building a thinking computer was taken by the field of artificial intelligence.

Cognitive science is the science that aims to explain how the mind works by describing the various mechanisms – systems of related parts interacting with each other to produce changes – responsible for producing processes such as problem solving, memory and learning. The main underlying assumption is that mind is a kind of computational system, and the field’s task is to explain how different kinds of thinking occur as the result of mental representations operated on by computational procedures that change mental states (THAGARD, 2012, p. 74). The proposed designs to explaining how those representations and procedures produce the processes that mind presents are called “cognitive architectures”. They are like blueprints of the mind and are understood as involving (1) a model of how the mind is organized into different cognitive systems, and (2) an account of how information is processed with (and across) those cognitive systems (BERMUDEZ, 2014, p. 139). The two most influential architectures are: the symbolic or rule-based

architecture, using if-then rules and procedures that operate on them – sometimes also referred as the classical computational theory of mind (CCTM) – and the connectionist architecture using artificial neural networks¹².

Even though the symbolic architecture encompasses some highly influential works as Chomsky's "Syntactic Structures" (1965) and Fodor's "Language of Thought" (1975). We can address its main features through Allen Newell and Herbert Simon's *physical symbol system hypothesis* (NEWELL, 1980). The main thesis of the physical symbol system hypothesis is that a symbol system has the necessary and sufficient means for general intelligent action. The four basic elements of it, as synthesized by Bermudez (2014, p. 143) are: 1) symbols are physical patterns; 2) These symbols can be combined to form complex symbol structures; 3) The physical symbol system contains processes for manipulating complex symbol structures; 4) the processes for generating and transforming complex symbol structures can themselves be represented by symbols and symbol structures within the system. It must be noted that those symbols are understood as physical patterns in the structure of the brain. It does not state that there is some representational symbol inside the human brain. As an analogy, we could take personal computers. When it is said that they function by means of binary code, this should not be understood as if there were 1's and 0's (or other representation of binary symbols) on its physical structure inside it. If you open a personal computer all that you will see are circuits transmitting electricity. The same should be thought of the brain. What we have on the physical (or, more precisely, biological) level are only neurons, glial cells, blood vessels, etc. Although the searching to find out how a physical system implements computations (including representational ones) is one of the main questions in this research program.

The symbol system hypothesis is an architecture based on Turing's model of an abstract machine that realizes computations. The way it is implemented could vary among different systems, although its description is purely mathematical. Informally, a computable task is one possible to specify by a sequence of instructions that, when carried out by some system, will result in the completion of the task (BARKER-PLUNER, 2016). The set of instructions specified for its completion is called an *effective process*, a *mechanical procedure* or an *algorithm*. A formal definition was proposed by Turing in his paper "On computable numbers, with an application to the Entscheidungsproblem" (1937), where he defines what became called Turing-computability: a task is computable if it can be carried out by a Turing Machine. Those machines are mathematical objects with defined characteristics and some idealizations. Such characteristics are: an infinite one-dimensional *tape* divided into cells. Each of them containing a symbol from a fixed alphabet, e.g. "0" and "1". Also, the machine has a head with the function of reading and writing that scans a single cell at a time, being able to move left and right along the tape to scan successive cells. The actions of

¹²A cognitive architecture is also used as a possible design for how to produce an artificial mind.

the machine are completely determined by: the current state of it, the symbol in the cell currently being scanned and the table of transition rules which serve as a “program” for the machine. Each transitional rule is a 4-tuple: $\langle \text{state}_{\text{current}}, \text{symbol}, \text{state}_{\text{next}}, \text{action} \rangle$. It can be read as saying that if the machine is in the $\text{state}_{\text{current}}$ and the cell being scanned contains symbols, then move into $\text{state}_{\text{next}}$ taking action. These actions could be either write a symbol on the cell denoted by the current symbol or to move the head left or right. If the machine reaches a condition where there's no more transition rule or are more than one to be carried, the machine halts (BARKER-PLUMMER, 2016). Also, as these machines are mathematical objects, they have some idealizations. For instance, the unrestricted time and storage capacity to perform its process. However, when implemented in the physical world by a computer (or by a brain if computationalism is true), the system becomes limited in its time to perform its process and the storage capacity becomes finite.

Connectionism, on the other hand, is the cognitive architecture which uses artificial neural networks to explain intellectual abilities. It doesn't take cognitive process as being encoded by rules, as the symbolic architecture does, instead it holds that those processes are encoded by the connections between neuron-like units called “nodes” (THAGARD, 2012 p. 53). Artificial neural networks (also called “neural nets” or “neural networks”) are systems composed of interconnected and weighted nodes which aim to be analogous to interconnected neurons and the strength of their synaptic connections (GARSON, 2016). Neural nets consist of various units joined in a pattern of connection. These nodes in the net can be classified into three classes: input nodes, hidden nodes, and output nodes. The input nodes are the ones responsible for receiving the information to be processed, while the output nodes are those where the results of that processing are found. The hidden nodes have a more subtle role in the network, they do not have to have any specific interpretation. Instead, they are important given the statistical connection between the input and the output nodes they furnish via its linkage.

The connectionist view does not necessarily go against the computational theory of mind. To better understand how this is possible we have to remember that computationalism is the view that cognition is computation, nothing else. It doesn't advance any special way by which those computations are done. The view about representation given by connectionism is a good example of how the position could still be computationalist. There are two different ways by which a neural network could represent some concept and both are computational: the localist and the distributed. The localist representation takes each node of the network as a unique concept usually specified in advance and with negative (suppressive) or positive connections between them. On the other hand, the distributed representation represents one single concept using many nodes. Let us take the concept of fairness, in a distributed representation nodes are not concepts. They are not specifications of what fairness is. Instead, the network learns what fairness is from experience and represents it using the entire pattern of nodes that were activated. In this case, the relationship between representations is coded in the similarities between those patterns (THAGARD, 2012 p.

53). The distributed way to represent concept is more flexible and able to work even if parts of the model were destroyed (GARSON, 2016).

Further, connectionists are divided into two groups: radical and implementational. The radical connectivists holds that symbolic architecture is a flawed way to explain mind because it fails to match human flexibility and efficiency. On the other hand, the implementational connectivists holds that the human brain, while being composed by neural nets, implement a symbolic processor at higher and more abstract levels of description, a position which emerged to accommodate both symbolic and connectionist approaches in just one model (GARSON, 2016).

2. SEARLE'S REASONS TO REJECT COMPUTATIONALISM

Searle's reasons to reject computationalism seems to be the following: (I) Searle assumes that *behaviorism* and *computationalism* are the two main mistakes in matters of consciousness and cognitive states; (II) as Searle assumes that behaviorism and computationalism are the two main mistakes in matters of consciousness and cognitive states, he designs a thought experiment combining both thesis to show that they cannot generate consciousness and other cognitive states; (III) However, Searle mistakenly identifies behaviorism as a core element in the computational theory of mind, as well as Turing-computation as a presupposed model of mind; (IV) As Searle mistakenly identifies those positions with the computational theory of mind, he cannot allow for the functional views at work in at least two replies to his thought experiment: the System and Robot replies; (V) Also as Searle mistakenly identifies that thesis with the computational theory of mind, he cannot allow that a "rightly programmed software" could have true cognitive states with content (or "intentionality" in his parlance); (VI) As the main causes behind the human to consciousness and other cognitive states could not be neither the behavioral or the computational. Searle identifies that cause as the "causal power of the brain", that is, the intrinsic biological processes of brains. Given so, I hold that (VII) if the components from I-VI are the reasons behind some seemingly mistakes in Searle's argumentation, and if they were corrected, the system and robot replies could be rehabilitated. In the next sections those theses are explained and the seeming mistakes analysed.

3. THE ANTI-SEMANTICAL ARGUMENT AND THE CRITICAL MISTAKES ABOUT THE MIND

Searle holds that a rightly programmed computer cannot be a mind in the sense of possessing understanding and other cognitive states (SEARLE, 1980, p. 417). The reason behind it is that such system would just realize computational processes over formally defined elements. And that would not be enough to generate content loaded cognitive states (*intentionality*) or content

apprehension (SEARLE, 1980, p. 422). That is, the system could act as if it has some cognitive state – the understanding of a language, for instance – but actually it would lack it because all that the system supposedly does are mechanical manipulations over meaningless formal symbols. For that reason, the alleged inability of computational systems to apprehend the content encompassed by the operations realized by itself, this argument is sometimes called the *anti-semantic argument*. His argument belongs to the class of *insufficiency objections* to the computational theory of mind: computation is insufficient for some cognitive phenomenon *X* (PICCININI, 2010).

To support that thesis Searle doesn't give a straightforward or even a semi-formalized argument, instead he hopes to persuade through a thought experiment

Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore (as is indeed the case) that I know no Chinese, either written or spoken, and that I'm not even confident that I could recognize Chinese writing as Chinese writing distinct from, say, Japanese writing or meaningless squiggles. To me, Chinese writing is just so many meaningless squiggles. Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that "formal " means here is that I can identify the symbols entirely by their shapes. Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch. Unknown to me, the people who are giving me all of these symbols call the first batch "a script", " they call the second batch a "story, " and they call the third batch "questions." Furthermore, they call the symbols I give them back in response to the third batch "answers to the questions," and the set of rules in English that they gave me, they call "the program." (SEARLE, 1980, p. 418).

What Searles describes in the first part of his thought experiment seems to be basically a specific kind of system very popular in the 80's: the expert system. An *expert system* is a system designed to be a support tool for decision makers. It has the capability to make simple inferences from a knowledge database of hand-coded facts given by humans (RUSSELL; NORVIG, 2010, p. 22). It also evokes one of the main purposes of the first flourishing years of artificial intelligence – around 1956 – where the idea was to debunk statements in the form of “A machine could never do *X*” by creating a machine that indeed does *X* in micro or toy world¹³ (BOSTROM, 2014, p.6). Back to expert systems, in Searle's example the database is given in the second batch of Chinese symbols as a “story”. The first batch could be thought as the alphabet to be followed, the third batch is the requests made of the information given in the database. Finally, the “rules” are the program necessary to drawing inferences from the database. It is also important to note that this kind of system was not designed to possess a general intelligence in the sense of matching humans in all fields. They were designed with only one specific task in mind and couldn't use its methods a data

¹³Unfortunately, when the researchers tried to build a more robust version of the machine aiming to work with the real world, a bunch of problems emerged. See Bostrom (2014, p. 3-18) for a brief explanation of the main problems in that period and the strategies developed to overcome them.

to do other tasks. So, Searle's thought experiment seems to give its first signs of mistakenly representing what it intends¹⁴. Further

Now just to complicate the story a little, imagine that these people also give me stories in English, which I understand, and they then ask me questions in English about these stories, and I give them back answers in English. Suppose also that after a while I get so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view - that is, from the point of view of somebody outside the room in which I am locked - my answers to the questions are absolutely indistinguishable from those of native Chinese speakers. Nobody just looking at my answers can tell that I don't speak a word of Chinese. Let us also suppose that my answers to the English questions are, as they no doubt would be, indistinguishable from those of other native English speakers, for the simple reason that I am a native English speaker. From the external point of view - from the point of view of someone reading my "answers" - the answers to the Chinese questions and the English questions are equally good. But in the Chinese case, unlike the English case, I produce the answers by manipulating uninterpreted formal symbols. As far as the Chinese is concerned, I simply behave like a computer; I perform computational operations on formally specified elements. For the purposes of the Chinese, I am simply an instantiation of the computer program (SEARLE, 1980, p. 418).

The experiment seems to have been designed to resemble the Turing test (TURING, 1950) while at the same time attack computationalism via Turing-computability (TURING, 1937). The Turing test was proposed to replace the question "Can a machine think?" by a more pragmatical one: Can a machine pass a test as if it were human? In the test a human judge has to decide whether a system interacting with them is a human or a computer and is a slightl modification of an existing game

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

Now suppose X is actually A, then A must answer. It is A's object in the game to try and cause C to make the wrong identification. His answers might therefore be ' My hair is shingled, and the longest strands are about nine inches long.'

In order that tones of voice may not help the interrogator the answers should be written, or better, typewritten. The ideal arrangement is to have a teleprinter communicating between the two rooms. Alternatively the question and answers can be repeated by an intermediary. The object of the game for the third player (B) is to help the interrogator. The best strategy for her is probably to give truthful answers. She can add such things as 'I am the woman, don't listen to him!' to her answers, but it will avail nothing as the man can make similar remarks.

We now ask the question, 'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, 'Can machine think?' (TURING, 1950, p. 433-434)

¹⁴For the various ways by which we are trying to achieve general intelligence see Bostrom (2014, p. 22-55)

Searle makes crystal clear his credence that Turing test is one the main responsible for what he sees as the two most critical mistakes considering matters of consciousness and cognitive states: behaviourism and computationalism

It [the Turing test] lead us to suppose that for a system to be conscious it is both necessary and sufficient that it has the right computer program or set of programs with the right input and outputs [...]. A traditional objection to behaviorism was that the behaviorism could not be right because a system could behave as if it were conscious without actually being conscious. There is no logical connection, no necessary connection between inner, subjective qualitative mental states and external, publicly observable behavior. Of course, in actual fact, conscious states characteristically cause behavior. But the behavior that they cause has to be long distinguished from the states themselves. The same mistake is repeated by computational accounts of consciousness. Just as behavior by itself is not sufficient for consciousness, so computational models of consciousness are not sufficient by themselves for consciousness (SEARLE, 1993, p. 317-318).

The second alleged mistake is computationalism, which is attacked by proxy via Searle's strikes against Turing-computability. We can identify the presence of Turing-computability in the Chinese room thought experiment by the designing features chosen by the author that reflects some of the main elements of Turing's computational machine¹⁵: a set of input symbols, precise rules describing how to correlate those symbols, a reading device (Searle himself) to implement those processes, and a set of output symbols given in the end of the process. A roughly idea of a Turing machine (already described in section 1 but now given in Turing's own terminology) is

We may compare a man in the process of computing a real number to a machine which is only capable of a finite number of conditions q_1, q_2, \dots, q_l ; which will be called "m-configurations". The machine is supplied with a "tape" (the analogue of paper) running through it, and divided into sections (called "squares") each capable of bearing a "symbol". At any moment there is just one square, say the r -th, bearing the symbol $\mathfrak{S}(r)$ which is "in the machine". We may call this square the "scanned square". The symbol on the scanned square may be called the "scanned symbol". The "scanned symbol" is the only one of which the machine is, so to speak, "directly aware". However, by altering its m-configuration the machine can effectively remember some of the symbols which it has "seen" (scanned) previously. The possible behaviour of the machine at any moment is determined by the m-configuration q_n and the scanned symbol $\mathfrak{S}(r)$. This pair $q_n, \mathfrak{S}(r)$ will be called the "configuration": thus the configuration determines the possible behaviour of the machine. In some of the configurations in which the scanned square is blank (i.e. bears no symbol) the machine writes down a new symbol on the scanned square: in other configurations it erases the scanned symbol. The machine may also change the square which is being scanned, but only by shifting it one place to right or left. In addition to any of these operations the m-configuration may be changed. Some of the symbols written down will form the sequence of figures which is the decimal of the real number which is being computed. The others are just rough notes to "assist the memory". It will only be these rough notes which will be liable to erasure. It is my contention that these operations include all those which are used in the computation of a number. (TURING, 1950, p. 231-232).

Considering both behaviorism and computationalism, we can identify some other elements in the thought experiment with features of the thesis under attack.

¹⁵Or, as is currently called "Turing machine".

Roughly speaking, *behaviorism* is an attitude towards science: the only kind of scientific research that can be done about psychological states are one who seeks to investigate its public aspects i.e.; the behavior that a person exhibit. A person is said to be a behaviorist if he or she seeks for behavioral evidence for any psychological hypothesis (SELLARS, 1991, p. 22). For such person, there aren't differences between two states of mind if there aren't differences in the behavior associated with each state (GRAHAM, 2015). In the Chinese room thought experiment Searle accuses computationalists of thinking that two distinct systems (some real person and the room) have the same cognitive state (understanding) because of both fulfil the behavioral evidence necessary to it. However, while the real person would have true cognitive states, the computational system (the room) just would be mimicking it. That kind of critique to behaviorism is called *psychologism*

Let psychologism be the doctrine that whether behavior is intelligent behavior depends on the character of the internal information processing that produces it. More specifically, I mean psychologism to involve the doctrine that two systems could have actual and potential behavior *typical* of familiar intelligent beings, that the two systems could be exactly alike in their actual and potential behavior, and in their behavioral dispositions and capacities and counterfactual behavioral properties (i.e., what behaviors, behavioral dispositions, and behavioral capacities they would have exhibited had their stimuli differed) – the two systems could be alike in all these ways, yet there could be a difference in the information processing that mediates their stimuli and responses that determines that one is not at all intelligent while the other is fully intelligent (BLOCK, 1981, p. 5).

Assuming psychologism, Searle reject both behaviorism and computationalism given his belief that latter falls for the same mistake as the former: “Just as behavior by itself is not sufficient for consciousness, so computational models of consciousness are not sufficient by themselves for consciousness (SEARLE, 1993, p. 317-318)”. But, it should be noted, psychologism does npt lead to a straightforward rejection of computationalism. Actually, both are fully compatible if one adopts a *functionalist* position about cognitive states.

Functionalism is the thesis that cognitive states are identified by what they *do* instead of what they are made of. On this approach, a cognitive state is independent of the internal constitution of the system, the only important thing is the way it functions or the role it plays in that system (LEVIN, 2016). Differently than behaviorism, functionalism takes into account not just the input-output relation publicly exhibited by a system, but also the existence of some *real* internal process of the right sort mediating that relation. In this sense, functionalism could be understood as building upon psychologism. Although, as seen, Searle believes that the tendency of computational accounts of mind to repeat the same mistakes of behavioral theories given the behaviorist influences in Turing's test who assumes that a teletyped (or something alike) input-output equivalence to a normal human being sophisticated enough to deceive a judge would be all that is needed for intelligent thinking, is enought to leads Searle to rejects any computational view, including

functionalism. Unlike those views, he goes in the entirely opposite direction and states that intrinsic lower and higher level properties of brains are the actual explanation for consciousness and cognitive state. A thesis called *Biological Naturalism*.

Biological Naturalism states that lower-level elements in the brain (e.g., neurons and synapses) are the causes of the higher-level features of consciousness and content loaded cognitive states (a.k.a. intentionality) (SEARLE, 2002, p. 57). The causality between lower-level elements and higher-level processes in the brain should be understood not as occurring between two discrete objects belonging to metaphysically distinct categories. Consciousness and content loaded cognitive states are to be understood as biological phenomena in the sense that they are caused by and interact with other biological processes. Further, they should be taken as biological processes in the same way as other processes as photosynthesis, digestion and the secretion of Bile (SEARLE, 2002, p. 60). Moreover, although *causally* reducible to lower-level elements and processes, consciousness and content loaded cognitive states are not *ontologically* reducible to them. They are understood as different from other phenomena that undergo an ontological reduction on the basis of a causal reduction. Consciousness, for example, is said not to be just firing neurons, it would also have a distinctive first-person ontology: it is something that only exists as experienced by some animal (e.g. human) and that cannot be reduced to something that has a third person ontology (SEARLE, 2002, p. 60)¹⁶. Yet, it is still a biological phenomenon given that all its causal powers are given by its (neuro)biological lower-level elements. Briefly, consciousness and content loaded cognitive states are states in which the brain *can be in* and not a some property distinct from it (SEARLE, 2002, p. 61).

In short, given that Searle understands consciousness and content loaded cognitive states as nothing more than physical phenomena of brains, the internal constitution of it becomes the necessary condition to that phenomena. Then, no running system of algorithms could have cognitive states given its internal constitution independence. Yet, a curious fact noted Rey (1986), is that the human doing the role of the head device in Searle's Chinese though experiment has all the "causal powers" that Searle could be asking for and yet, in Searle's own words, cannot understand Chinese.

4. REHABILITATING THE SYSTEM AND ROBOT REPLIES: SOME HYPOTHESES

Now that we have revised both Searle's reasons to reject computationalism and his own thesis about the mind, we are better suited to take another look at some counter-objections he raises against two main attacks on the Chinese room thought experiment: The "system reply" and the

¹⁶ Yet, Searle seems to had previously advocated the opposite thesis (SEARLE, 1997, p. 7-9).

“robot reply”. As already known, Searle’s anti-semantic objection to computationalism is not supported by a straightforward argument. Instead, he gives a thought experiment – a kind of device whose legitimacy could be disputed (SORENSEN, 1998; DAMPER, 2006, p. 168) – to support it. That strategy makes indeed its objection much more vulnerable to a wide range of possible mistakes.

The *System reply*, one of the main objections against the *anti-semantic argument*, states that while the person inside the Chinese room indeed doesn’t understand Chinese, the whole system does (SEARLE, 1980, p. 419-420). Searle replies to that objection arguing that if he internalizes in his memory the entirely manual describing (in English) how to associate the unknown input symbols with the unknown output symbols it would make all the components given in the thought experiment internal to himself. If it were the case that the room was what understood Chinese then now the person with the internalized components (themselves) should understand Chinese as well. Although, he continues, he *still* doesn’t have any understanding at all.

The internalized version of the rule manual, if computationalism were true *and* if the right model for it were Turing-computability (i.e, a system as that described in section 1), would be the *machine table* determining how to associate the inputs with outputs. At this point the seeming mistake of assuming a non-essential feature of computationalism as that the right model for it is Turing-computability, seems to play an important role. And, we saw at the beginning, the only essential thesis for computationalism is that cognition is computation. Not that cognition is computation through Turing machines or anything else.

Although, even if Searle were right in that assumption (and he isn’t) we can speculate a little further. If computationalism were true it seems that being a computational system (e.g. a Turing machine) running the right computational processes needed to have some cognitive state as, for instance, to “understand” a language, doesn’t make those *computational processes* to be necessarily *consciously known* to the system in which they are running on. The only necessary thing for some system to be a computational one is that the right processes are *running on* it. The *knowability* of its machine table seems to be a kind of *reflexive* ability or an *auditing* over what going on in the system and not a necessary ability given by being those processes running on it. It would be as if it were an additional program in the system, a program whose task is solely to perform the reflexive or auditing tasks over the other programs running on the system. If this were true, it could be the case that computing the outputs to inputs given in Chinese without auditing it could be the same as providing outputs to inputs in English. The alleged “understanding” of English words could be just a way to provide to oneself or to someone a distinct piece of data that were correlated with the one it is being taken as an input at the moment. Further, if we want to go a little deeper, we could try to conjecture how a system which runs some set of programs – e.g., takes inputs from perceptual modules and gives outputs to them – would *feel from inside*. That is, what

the system would feel from *being* this kind of system. Well, I believe that it's at least reasonable to hypothesize that *from inside* it would feel *as if* it really understands what is going on when using some language, describing some action or something alike. It seems to me that it would be analogous to when someone ask you a question. At your *personal level* (or *phenomenical level*) – the level at which things are conscious to oneself – you just give an answer based on your previous relevant knowledge of what was being asked (or possible new hypothesis generated upon that). And if it is something hard to answer, it isn't rare that take some time until the answer pop up in you mind. The main point with my speculations is that all that was described could happen *even if you are a program (or set of programs) with "computational operations on purely formally specified elements"* as Searle states. There's nothing in the first-person feeling about itself that *necessarily* would distinguish if you are a computational system or not. You would have just the same feelings about yourself as in the case in which computationalism is false.. All my previous hypotheses and speculations seems to be in direct opposition to Searle's account of how a computational system would feels from inside

In the Chinese case I have everything that artificial intelligence can put into me by way of a program and I understand nothing; in the English case I understand everything and there is so far no reason at all to suppose that my understanding has anything to do with computer programs, that is, with computational operations on purely formally specified elements (SEARLE, 1980, p. 418).

In short, my hypothesis is that you does not have to have access to ones own program to be a program. One do not have to necessarily know that you are composed of many programs running on yourself to have those programs running in yourself. It seems to me that knowing its all source code isn't something necessary at all to be that source code. Still further, it could be argued that knowing its own program is something redundant, costing, and purposeless as an evolutionary strategy. Some evolved system doesn't has to be a system with constant internal auditing if it could achieve the same or some approximate result without the constant auditing capability. If the usual cognitive processes as to speak a language could happen without a constant awareness of the underlying processes running to make it possible. That is, it could just seem to flow intuitively instead of having to know what its happening to each of the words being picked up to be said, when to be said and even how those words are related to form what is being said. And this could even free the more costing process of auditing to more important events. If the previous hypothesis is right, it could be possible to rehabilitate the system reply from Searle's counter-objection.

In its turn, the *robot reply* holds that besides the computational features of an agent, its embodiment also plays a decisive role to generate content loaded cognitive states

“[...] Suppose we put a computer inside a robot, and this computer would not just take in formal symbols as input and give out formal symbols as output, but rather would actually operate the robot in such a way that the robot does something very much like perceiving, walking, moving about, hammering nails, eating, drinking – anything you like. The robot would, for example, have a television camera attached to it that enables it to ‘see,’ it would have arms and legs that enable it to ‘act,’ and all of this would be controlled by its computer ‘brain.’ Such a robot would [...] have genuine understanding and mental states” (SEARLE, 1980, p. 420).

Roughly, what the robot reply states is that what produce the link between the computational system and the world which gives the content of its cognitive states are the different modes provided by its embodiment such as: sensor to extract data from the world, the navigating ability to explore and perform action in the world given what was mapped by its sensors, etc. That is, the embodiment of such system would provide a spectrum of interlocking abilities that together amounts to a cognitive agent (DAMPER, 2006, p. 166). However, Searle has a different opinion about that. He believes that “[...] the answer to the robot reply is that the addition of such “perceptual” and “motor” capacities adds nothing by way of understanding, in particular, or intentionality in general [...] to the original program (SEARLE, 1980, p. 420)”. It appears to me that, unfortunately, the way in which Searle took the new abilities of the system is over simplistic and a source of mistakes. The robot reply does not attach perceptual and motor skills to the system in a trivial sense. We should think of it as a much more complex interrelation of specialised programs regulating various systems that “hook up” into the world. That mischaracterization of an embodied computational system hidden the true question: does the Chinese room thought experiment functionally replicates what happens inside a normal human speaker? The answer seems to be straightforward “no”. In Searle’s view, if computationalism were true it seems that cognitive states would have to be a straightforward correlation among some meaningless symbols. However, a less over simplistic model should consider that the embodied system would have to have at least some rules for correlating inputs from a wide range of specialized programs like perception, belief fixation, problem solving, preference ordering, decision making and the ability to act non-verbally to other peoples sentences (REY, 1986, p. 171). If we take a look on some of the main research programs in artificial intelligence – all of them presumably working exclusively with syntax in Searle’s view – we find: natural language processing (the capacity of communicate successfully in natural language); knowledge representation (the capacity to store what it knows or hears); automated reasoning (the capacity to use the stored information to answer questions and to draw new conclusions); machine learning (the capacity to adapt to new circumstances and to detect and extrapolate patterns); computer vision (the capacity to perceive objects); and robotics (the capacity to manipulate objects and move about) (RUSSEL; NORVIG, 2010, p. 2 – 3). Insofar as Searle’s analyses go, he would call all of these fields examples of “weak AI” (1980, p 417), not useful to generate a mind. Although all of these are components to be expected in an agent functionally

equivalent to a normal human and should be taken into account as possible parts of an embodied system given that those fields are already under development.

Briefly, to a rightly account about cognitive states in a computationalist basis one has to take into account that various relevant modules that are working together to furnish information to each other and acting over what was received. The embodied system in the robot reply would not be realizing just another trivial symbol association on the data given by its new modules. Instead it would be performing a highly complex interrelation of such new specialized systems before performing its output. And that is not some unimportant new feature to content loaded cognitive states. Instead, the new modules seems to be *directed towards* receiving new data *about* the world and to act upon the world given what was drawing from it. Two features that seem highly relevant to some system that hopes to possess the ability to not just represent but also to knows that it has such representations of either his internal processes or the external world. Therefore, I believe, to rehabilitate the system reply one has just to give a slightly better model than the one Searle's gives to how a computational embedded system could work.

5. CONCLUSION

When an analysis is done of how it was chosen the main features into designing the "Chinese room thought experiment" some seeming misconceptions become apparent about the computational theory of mind previously assumed and built into the though experiment. Aside from the misidentification of the core thesis of computationalism and the wrongly attack through a proxy that has been done against it, it was also possible to detect some seeming mistakes when it was tried to use some of the main features of computationalism against itself. Once identified those seeming mistakes it was possible to give a possible new breath to some of the main replies against Searle's anti-semantic argument.

REFERENCE

- BARKER-PLUMMER, D, Turing Machines. In Zalta, E (ed). **The Stanford Encyclopedia of Philosophy**, spring, 2016. Disponível em: <http://plato.stanford.edu/archives/spr2016/entries/turing-machine/>.
- BERMÚDEZ, J. **Cognitive science**: An introduction to the science of the mind. Cambridge: Cambridge University Press, 2014.
- BLOCK, N. Psychologism and behaviorism. **The Philosophical Review**, v. 90, n. 1, p. 5-43, 1981.
- BOSTROM, N. **Superintelligence**: Paths, dangers, strategies. Oxford: Oxford University Press, 2014.
- CHOMSKY, N. **Syntactic structures**. Berlin: Walter de Gruyter, 2002.

- DAMPER, R. The logic of Searle's Chinese room argument. **Minds & Machines**, v. 16, n. 2, 2006.
- DIETRICH, E (ed). **Thinking computers and virtual persons: essays on the intentionality of machines**. San Diego: Academic Press, 1994.
- FODOR, J. **The language of thought**. Massachusetts: Harvard University Press, 1975.
- GARSON, J, Connectionism. In: Zalta, E (ed). **The Stanford Encyclopedia of Philosophy**, winter, 2016. Disponível em: <<https://plato.stanford.edu/archives/win2016/entries/connectionism/>>.
- GRAHAM, G, Behaviorism, In: Zalta, E (ed). **The Stanford Encyclopedia of Philosophy**, spring, 2017. Disponível em: <<https://plato.stanford.edu/archives/spr2017/entries/behaviorism/>>.
- LEVIN, J. Functionalism. In: Zalta, E (ed), **The Stanford Encyclopedia of Philosophy**, winter, 2016. Disponível em: <<https://plato.stanford.edu/archives/win2016/entries/functionalism/>>.
- NEWELL, A. Physical symbol systems. **Cognitive science**, v. 4, n. 2, p. 135-183, Apr. 1980.
- PICCININI, G. The resilience of computationalism. **Philosophy of Science**. v. 77. n. 5. pp 852-861, 2010.
- REY, G. What is really going on in Searle's "Chinese room". **Philosophical studies: An international Journal for Philosophy in the Analytic Tradition**. v. 50, n. 2, pp. 169-185, 1986.
- RUSSEL, S; NORVIG, P. **Artificial Intelligence: A modern approach**. 3. ed. Essex: Pearson Education Limited, 2010.
- SEARLE, J. Minds, brains, and programs. **Behavioral and brain sciences**, v. 3, n. 3, 417-424, 1980.
- SEARLE, J. **The mystery of consciousness**. New York: New York Review of Books, 1997.
- SEARLE, J. The problem of consciousness. **Social Research**, v. 2, n. 4, p.3-16, 1993.
- SEARLE, J. Why I am not a property dualist. **Journal of consciousness studies**, v. 12, n. 9, p. 57-64, 2002.
- SELLARS, W. The scientific image of man. In: SELLARS, W. **Science, perception, and reality**. Atascadero: Ridgeview Publishing Company, p. 1-40, 1991.
- SORENSEN, R. **Thought experiments**. Oxford: Oxford University Press, 1998.
- THAGARD, P. Cognitive Architectures. In: Frankish, K; Ramsey, W (ed). **The Cambridge handbook of cognitive science**. Cambridge: Cambridge University Press, p. 50 -70, 2012.
- TURING, A. Computing machinery and intelligence. **Mind**. v. 59, n. 236, p. 433-460, Oct. 1950.
- TURING, A. On computable numbers, with an application to the Entscheidungsproblem. **Proceedings of London Mathematical Society**, v. s2-42, ed. 1, p. 230-265, Nov. 1937.

TRADUÇÃO – MENTES, MÁQUINAS E GÖDEL

Tradução: Pablo Rolim

J. R. Lucas

O Teorema de Gödel, parece-me, prova que o Mecanicismo é falso. Isto é, que mentes não podem ser explicadas como sendo máquinas. Assim também há parecido à muitas outras pessoas: quase todo lógico matemático ao qual fiz a questão confessou-me possuir pensamentos similares mas estar relutante de comprometer-se definitivamente com eles até poder ver todo o argumento organizado, com todas suas objeções completamente explícitas e adequadamente satisfeitas. E isso, é o que tento fazer.

O teorema de Gödel afirma que em qualquer sistema consistente forte o bastante para produzir a aritmética simples há fórmulas que não podem ser demonstradas no sistema, mas as quais nós podemos ver que são verdadeiras. Essencialmente, consideremos a fórmula que diz, por exemplo, “Esta fórmula não é demonstrável no sistema”. Se essa fórmula for demonstrável no sistema então teremos uma contradição pois se ela for demonstrável no sistema, então ela seria indemonstrável no sistema, de modo que “Esta fórmula não é demonstrável no sistema” seria falsa. Igualmente, se ela fosse demonstrável no sistema, então ela não seria falsa, mas verdadeira, visto que em qualquer sistema consistente nada falso pode ser provado, apenas verdades. Portanto a fórmula “Esta fórmula não é demonstrável no sistema” não é demonstrável no sistema, mas indemonstrável no sistema. Ainda, se a fórmula “Esta fórmula é indemonstrável no sistema” é indemonstrável no sistema, então é verdade que essa fórmula é indemonstrável no sistema, isto é, “Esta fórmula é indemonstrável no sistema” é verdadeira.

O argumento precedente é trivial mas difícil de compreender completamente. Pode ser de ajuda colocar o argumento de outra forma: considere a possibilidade de que “Esta fórmula é indemonstrável no sistema” possa ser falso, mostre que isso é impossível portanto que a fórmula é verdadeira; do que segue-se que ela não é demonstrável. Mesmo assim, o argumento continua não sendo convincente: sentimos que deve haver algo errado em algum lugar. Todo o trabalho do teorema de Gödel é mostrar que não há nada de errado, e que o resultado pode ser estabelecido através da dedução mais rigorosa. E que isso sustenta-se para todos os sistemas formais que são (i) consistentes, (ii) adequados para a aritmética simples – isto é, que inclua os números naturais e as operações de adição e multiplicação – e mostra que eles são incompletos - isto é, contém fórmulas não demonstráveis, ainda que perfeitamente significativas. Algumas das quais, ademais, que nós, posicionados de fora do sistema, podemos ver que são verdadeiras.

Os teoremas de Gödel devem aplicar-se à máquinas cibernéticas, pois está na essência de ser uma máquina que essa deva ser uma instanciação concreta de um sistema formal. Segue-se que

dada uma máquina que é consistente e capaz de realizar aritmética simples, há uma fórmula que essa é incapaz de mostrar que é verdadeira – isto é, a fórmula não é demonstrável no sistema – mas que nós podemos ver que é verdadeira. Segue-se, então, que nenhuma máquina pode ser um modelo completo ou adequado da mente. Mentes são essencialmente diferentes de máquinas.

Entendemos por uma máquina cibernética um aparato que realize um conjunto de operações conforme um conjunto definido de regras. Normalmente nós “programamos” uma máquina: isto é, damos à ela um conjunto de instruções sobre o que deve fazer em cada circunstância; e fornecemos a “informação” inicial sobre a qual a máquina irá realizar seus cálculos. Quando consideramos a possibilidade de que a mente seja uma máquina cibernética temos tal modelo em vista; supomos que o cérebro seja composto de circuitos neurais complicados, e que a informação fornecida pelos sentidos seja “processada” e posta em uso ou seja armazenada para uso futuro. Caso a mente seja tal mecanismo, então dada a forma segundo a qual é programada --- o modo em que é “cabeadada” – e a informação com a qual tenha sido alimentada, a resposta – o “output” – é determinado, e pode, dado tempo suficiente, ser calculado. Nossa ideia de uma máquina é simplesmente esta, dado que seu comportamento é completamente determinado pela forma na qual é construída e o “estímulo” de entrada: não há possibilidade de que ela aja por conta própria: dada uma certa forma de construção e uma certa entrada de informação, ela deve agir de uma certa forma específica. Não devemos, no entanto, nos preocuparmos com o que uma máquina *deve* fazer, mas sim com o que ela *pode* fazer. Isto é, ao invés de considerar todo o conjunto de regras que conjuntamente determinam exatamente o que uma máquina irá fazer nas circunstâncias dadas, devemos considerar apenas um esboço dessas regras, que irá delimitar as respostas possíveis da máquina, mas não completamente. As regras completas irão determinar completamente o comportamento em qualquer momento; em cada momento haverá uma instrução definida. Por exemplo, “Se o número é primo e maior que dois, adicione um e divida por dois: caso não seja primo, divida por seu menor fator”. Nós, entretanto, consideraremos a possibilidade de que hajam instruções alternativas. Por exemplo, “Em uma fração você pode dividir a parte superior e inferior por *qualquer* número que seja um fator tanto do numerador quanto do denominador”. Assim, relaxamos a especificação de nosso modelo, de forma que ele não seja mais completamente determinista, ainda que permaneça mecanicista; assim poderemos levar em conta uma característica frequentemente proposta para modelos mecânicos da mente, a saber, que eles contêm um dispositivo de randomização. Pode-se construir uma máquina cuja as escolhas entre um número de alternativas seja estabelecida, digamos, pelo número de átomos de rádio que tenham desintegrado em um dado recipiente no último meio minuto. É plausível, *prima facie*, que nossos cérebros devam estar sujeitos a efeitos aleatórios: um raio cósmico pode bem ser o bastante para disparar um impulso neural. Mas claramente em uma máquina um dispositivo randomizador não pode ser introduzido de forma a escolher uma alternativa qualquer: poderia a ele apenas ser permitido escolher entre um número de alternativas permissíveis. Não há problemas em adicionar um número *qualquer* escolhido aleatoriamente à ambos os lados de

uma equação, mas há em adicionar à um lado e não ao outro. Não há nenhum problema em escolher demonstrar um teorema de Euclides ao invés de outro, ou de usar um método e não o outro, mas há em “demonstrar” algo que não é verdadeiro, ou em um usar um “método de demonstração” que não é válido. Qualquer dispositivo randomizador deve permitir escolhas apenas entre aquelas operações que não levarão à inconsistência: que é exatamente o que descrição relaxada de nosso modelo especifica. Certamente, pode-se colocar da seguinte forma: a invés de considerar o que uma máquina completamente determinada *deve* fazer, devemos considerar o que uma máquina pode ser capaz de fazer se possuir um dispositivo randomizador que aja sempre que houver duas ou mais operações possíveis, nenhuma das quais leve à inconsistência.

Se tal máquina fosse construída visando produzir teoremas da aritmética (que de muitas formas pode ser considerada a parte mais simples da matemática), ela teria apenas um número finito de componentes, e assim poderia ter apenas um número finito de tipos de operações realizáveis por ela, assim como apenas um número finito de pressuposições iniciais sob as quais poderia estar operando. Poderíamos, certamente, ir além e afirmarmos que há apenas um número *preciso* de tipos de operações e pressuposições iniciais que poderiam ser construídas nela. Máquinas são precisas: quaisquer coisas que sejam imprecisas ou infinitas não devem ser consideradas como máquinas. Note que dissemos número de *tipos* de operações, não número de operações. Dado tempo suficiente, e considerando que não se desgaste, uma máquina pode continuar a repetir uma operação indefinidamente: apenas deve haver um número definido de espécies de operações que ela possa realizar.

Se houver apenas um número definido de tipos de operações e pressuposições iniciais construídas no sistema, podemos representar todas elas através de símbolos adequados escritos em papel. Podemos paralelizar a operação através de regras (“regras de inferência” ou “Esquemas de Axiomas” permitindo-nos ir de uma ou mais fórmulas (ou mesmo de nenhuma fórmula) à outra fórmula, e podemos paralelizar as pressuposições iniciais (se houver alguma) através de um conjunto inicial de fórmulas (“proposições primitivas”, “postulados” ou “axiomas”). Uma vez que tivermos representado isso no papel, podemos representar cada operação singular: tudo o que precisamos fazer é usar fórmulas para representar a situação antes e após a operação, e anotar qual fórmula está sendo invocada. Podemos então representar no papel qualquer sequência de operações possíveis que a máquina possa realizar. Independente de quanto tempo a máquina leve para realizar uma operação, nós podemos, dado o tempo necessário, papel e paciência, escrever um análogo das operações da máquina. Esse análogo seria de fato uma prova formal: cada operação da máquina é representada pela aplicação de uma das regras; e as condições que determinam se a máquina pode realizar uma operação em uma dada situação tornam-se em nossa representação condições que estabelecem se uma regra pode ser aplicada a certa fórmula. Isto é, condições formais de aplicabilidade. Portanto, construindo nossas regras como regras de inferência, teremos a sequência de demonstração das fórmulas, cada uma delas sendo escrita em virtude de alguma regra formal de

inferência sendo aplicada à alguma fórmula ou fórmulas prévias (exceto, obviamente, para as fórmulas iniciais, que são dadas pois representam as pressuposições iniciais construídas no sistema). As conclusões que a máquina pode exibir como sendo verdadeiras serão, portanto, corresponder aos teoremas que podem ser demonstrados no sistema formal correspondente. Construamos agora uma fórmula Gödeliana nesse sistema formal. Essa fórmula não pode ser *demonstrada no sistema*. Portanto, a máquina não pode apresentar a fórmula correspondente como verdadeira. Mas nós podemos ver que a fórmula Gödeliana é verdadeira: qualquer ser racional poderia seguir o argumento de Gödel e convencer-se de que a fórmula Gödeliana, ainda que não demonstrável no sistema, é entretanto – de fato, por essa mesma razão - verdadeira. Ora, qualquer modelo mecânico da mente deve incluir um mecanismo que possa enunciar verdades aritméticas, visto que isso é algo que mentes podem fazer. Na verdade, é fácil produzir modelos mecânicos que irão, em muitos aspectos, produzir verdades aritméticas muito melhor do que seres humanos. Mas neste aspecto em específico elas não podem se sair tão bem: para cada máquinas há uma verdade que ela não pode mostrar ser verdadeira, mas que uma mente pode. Isso mostra que uma máquina não pode ser um modelo completo e adequado da mente. Ela não pode realizar *tudo* que uma mente pode, uma vez que independente de quanto ela possa realizar haverá sempre algo que ela não pode e que uma mente pode. Isto não é dizer que não possamos construir uma máquina para simular *cada* parte desejada do comportamento de uma forma semelhante à mente. Podemos (ou seremos capazes um dia) de construir máquinas capazes de reproduzir peças de comportamento semelhantes ao da mente, e certamente de superar a performance de mentes humanas, mas independente de quão boa essa máquina seja e de quão bem ela puder se sair em quase todos os aspectos que uma mente humana pode, ela sempre possuirá essa fraqueza. Essa coisa que ela não pode realizar enquanto uma mente pode. A fórmula Gödeliana é o calcanhar de Aquiles da máquina cibernética. E portanto não podemos ter esperanças de criar uma máquina que será capaz de realizar tudo que uma mente pode: nunca poderemos, mesmo em princípio, ter um modelo mecânico da mente.

Essa conclusão será altamente suspeita para algumas pessoas. Elas irão objetar primeiro que nós não podemos afirmar ambos: que uma máquina *pode* simular *qualquer* parte do comportamento similarmente a uma mente e que ela *não pode* simular *cada* parte desse comportamento. Para alguns isso é uma contradição: para eles não é o bastante apontar que não há nenhuma contradição entre o fato de que para todo número natural pode ser produzido um número ainda maior, e o fato de que não podemos produzir um número maior que qualquer número. Podemos usar a mesma analogia contra aqueles que, encontrando uma fórmula que suas máquinas não podem exibir como verdadeira, concedem que a máquina é certamente inadequada, mas que em consequência disso buscam construir uma segunda máquina mais adequada na qual a fórmula *pode* ser exibida como verdadeira. Isso eles podem certamente fazer: mas então a segunda máquina terá uma fórmula Gödeliana própria que pode ser construída ao aplicar o procedimento Gödeliano ao sistema formal que representa seu próprio esquema de operações ampliado (isto é, ao da segunda máquina). E essa

fórmula a segunda máquina não será capaz de mostrar como verdadeira, enquanto uma mente será capaz de ver que o é. E caso construamos uma terceira máquina capaz de fazer aquilo que a segunda máquina é incapaz o mesmo irá ocorrer: haverá ainda uma terceira fórmula, a fórmula Gödeliana para o sistema formal correspondendo ao esquema de operações da terceira máquina, que essa é incapaz de mostrar verdadeiro mas que uma mente ainda será capaz de ver que o é. E assim por diante. Independente de quão complicada seja a máquina que construamos, ela terá, se for uma máquina, um sistema formal correspondente que por sua vez estará sujeito aos procedimentos de Gödel para encontrar uma fórmula indemonstrável nesse sistema. Essa fórmula a máquina será incapaz de mostrar que é verdadeira, entretanto uma mente pode ver que é verdadeira. E assim a máquina ainda não será um modelo adequado da mente. Estamos tentando produzir um modelo da mente que seja mecânico - que é essencialmente “morto” - mas a mente, estando de fato “viva”, pode sempre ir além de qualquer sistema formal, fossilizado, e morto.

Uma segunda objeção será agora realizada. O procedimento através do qual a fórmula Gödeliana é construída é um procedimento padrão – apenas assim nós podemos estar certos de que uma fórmula Gödeliana pode ser construída para cada sistema formal. Mas se esse é um procedimento padrão, então uma máquina deveria ser capaz de ser programada para realizá-lo. Podemos construir uma máquina com as operações usuais e, adicionalmente, uma operação de realização do procedimento de Gödel produzindo assim a conclusão daquele procedimento como verdadeira; e assim repetindo o procedimento tanto quanto for requerido. Isso seria correspondente a ter um sistema com uma regra de inferência adicional que permita adicionar como um teorema a fórmula Gödeliana do resto do sistema formal, e então a fórmula Gödeliana desse novo e fortalecido sistema e assim por diante. Isso seria equivalente à adicionar ao sistema formal original uma sequência infinita de axiomas, com cada uma das fórmulas Gödelianas obtidas previamente. Porém mesmo assim, a questão não estará resolvida: pois a máquina com o operador Gödelizador, como podemos chamá-lo, é uma máquina *diferente* da máquina sem tal operador; e, ainda que a máquina com tal operador seja capaz de realizar aquelas coisas cuja máquina sem o operador é superada pela mente, nós ainda podemos esperar que uma mente confrontada com uma máquina que possua o operador Gödelizador, leve isso em conta e então supere a nova máquina, mesmo com o operador Gödelizador. Isso tem, de fato, mostrado-se o caso. Mesmo se unirmos a um sistema formal o conjunto infinito de axiomas consistindo de sucessivas fórmulas Gödelianas, o sistema resultante ainda será incompleto, e conterá uma fórmula que não pode ser demonstrada no sistema e que ainda assim, um ser racional pode, posicionado de fora do sistema, ver que é verdadeira¹⁷. Isso era esperado, pois mesmo se um conjunto infinito de axiomas for adicionado, eles terão de ser especificados por um conjunto finito de regras ou especificações, e essa regra ou especificação adicional poderia então ser levada em conta por uma mente ao considerar o sistema formal

¹⁷A prova original de Gödel aplica-se, v. § I init. § 6 init. de suas “Lectures at the Institute of Advanced Study”, Princeton, N.J., U.S.A., 1934.

ampliado. Em um sentido, apenas porque a mente possui a última palavra, ela pode sempre encontrar uma brecha em qualquer sistema formal apresentado à ela como um modelo de seu próprio funcionamento. O modelo mecânico deve ser, em algum sentido, finito e precoso: e então a mente sempre poderá sair-se melhor.

Essa é a resposta para uma objeção apresentada por Turing¹⁸. Ele argumenta que a limitação das capacidades de uma máquina não significam muito. Ainda que cada máquina individual seja incapaz de alcançar a resposta correta para alguma questão, os seres humanos também o são por serem falíveis: e em cada caso “nossa superioridade pode apenas ser sentida em relação à uma máquina sob a qual nós tenhamos obtido nosso pequeno triunfo. Não haveria um caso de triunfo simultâneo sob *todas* as máquinas”. Mas esse não é o ponto. Não estamos discutindo se máquinas ou mentes são superiores, mas se elas são o mesmo. Em alguns aspectos máquinas são sem dúvida superiores à mentes humanas; e a questão sob a qual eles estão perplexos é, embaraçosamente, trivial. Entretanto isso não é o bastante, não o bastante para mostrar que a máquina *não é o mesmo* que uma mente. Verdadeiramente, a máquina pode fazer muitas coisas que a mente humana não pode: mas se houver necessariamente algo que a máquina não pode fazer, ainda que a mente possa, então, seja quão trivial for, nós não podemos igualar as duas, e não podemos esperar que algum dia haja um modelo mecânico que irá adequadamente representar a mente. Também não significa que seja apenas sobre uma máquina individual que tenhamos triunfado: pois o triunfo não é sobre *uma* máquina individual, mas sobre *qualquer* uma que alguém importe-se em especificar. Em latim *quavis* ou *quilibet*. não *quidam* – e um modelo mecânico da mente deve ser uma máquina individual. Ainda que seja verdade que qualquer “triunfo” particular de uma mente sobre uma máquina possa ser “superado” por outra máquina capaz de apresentar a resposta que a primeira máquina não pode apresentar, de modo que “não haja uma questão de triunfo simultâneo sobre todas as máquinas”, ainda assim isso é irrelevante. O que está em questão não é se pode haver alguma, única, máquina que possa fazer tudo que uma mente pode. Para a tese mecanicista manter-se, deve ser possível, em princípio, produzir um modelo, um único modelo, que possa realizar tudo que uma mente pode. É como um jogo¹⁹. O mecanicista possui a primeira rodada. Ele produz *a* – *qualquer*, mas apenas *um que definido* – modelo mecânico da mente. Eu, então, aponto para algo que ele não pode realizar mas que a mente pode. O mecanicista está então livre para modificar seu exemplo, mas a cada vez que ele o faz, eu possuo o direito de procurar por defeitos no modelo revisado. Se o mecanicista puder inventar um modelo ao qual eu não possa encontrar falta sua tese é estabelecida: se ele não puder, então não está provada: e visto que – como se constata – necessariamente ele não pode, ela é refutada. Para ter sucesso, ele precisa ser capaz de produzir algum modelo mecânico específico da mente – qualquer um que ele preferir, mas um que ele possa precisar e manter-se com ele. Mas visto que ele não pode, mesmo em princípio, produzir um

¹⁸*Mind*, 1950, pp. 444-5; Newman, p. 2110.

¹⁹Para um tipo similar de argumento, veja J. R. Lucas “The Lesbian Rule”; *PHILOSOPHY*, July 1955, pp. 202-6; e “on not worshipping Facts”; *The Philosophical Quarterly*, Abril 1958, p. 144.

modelo mecânico que seja adequado, mesmo se o ponto de falha seja pequeno, ele está fadado ao fracasso, e o mecanicismo deve ser falso.

Objeções mais profundas ainda podem ser realizadas. O teorema de Gödel aplicam-se à sistemas dedutivos, e seres humanos não estão confinados a realizar inferências dedutivas. O teorema Gödel aplica-se apenas à sistemas consistentes, e pode-se ter dúvidas sobre o quão longe é permissível pressupor que seres humanos são consistentes. O teorema de Gödel aplica-se apenas à sistemas formais, e não há *a priori* nenhum limite ao engenho humano que exclua a possibilidade de inventarmos alguma replica da humanidade que não possa ser representável por um sistema formal.

Seres humanos não estão limitados à realizar inferências dedutivas, e tem sido insistido por C. G. Hempel²⁰ e Hartley Rogers²¹ que um modelo justo da mente da mente deveria permitir a possibilidade de realizar inferências não dedutivas, e essas poderiam prover uma forma de escapar do resultado de Gödel. Hartley Rogers faz a específica sugestão de que uma máquina deveria ser programada para acolher várias proposições que não tenham sido provadas ou refutadas, e ocasionalmente adicioná-las à sua lista de axiomas. O último teorema de Fermat ou a conjectura de Goldbach poderia então ser adicionada. Se subsequentemente a sua inclusão for descoberto que isso leva a uma contradição, elas deveriam ser abandonadas novamente, e certamente em tais circunstâncias suas negações deveriam ser adicionadas à lista de teoremas. Dessa forma poderia ser construída uma máquina que fosse capaz de exibir como verdadeiras certas fórmulas que não podem ser demonstradas a partir de seus axiomas de acordo com as regras de inferência. E portanto o método de demonstrar a superioridade da mente sobre a máquina não mais funcionaria.

A construção de tal máquina, entretanto, apresenta dificuldades. Ela não pode aceitar todas as fórmulas não demonstradas e adicioná-las à seus axiomas, ou ela iria encontrar-se tanto aceitando a fórmula Gödeliana quanto sua negação, e então seria inconsistente. Nem mesmo daria certo se aceitasse o primeiro de cada par de fórmulas indecidíveis, e, tendo adicionado esse à seus axiomas, não poderia mais considerar sua negação como indecidível e assim jamais poderia aceitar isso também: pois isso poderia acontecer no membro errado do par: ela poderia aceitar a negação da fórmula Gödeliana ao invés da própria fórmula Gödeliana. E o sistema constituído por um conjunto normal de axiomas juntamente com a negação da fórmula Gödeliana, ainda que não seja inconsistente, não é um sistema sólido, não admitindo a interpretação natural. É algo como as geometrias não Desarguenianas em duas dimensões: não realmente inconsistentes, mas erradas o suficiente para desqualificá-las de considerações sérias. Uma máquina que não é passível a infortúnios daquele tipo não poderia ser um modelo da mente humana.

Torna-se claro que critérios bastante cuidadosos de seleção de fórmula não prováveis serão necessários. Hartley Rogers sugeriu alguns possíveis. Mas uma vez que tenhamos regras gerando

²⁰Em conversa privada.

²¹*Theory of Recursive Functions and Effective COmputability*, 1957, Vol. I, pp. 152 ff.

novos axiomas, mesmo se os axiomas gerados são apenas provisoriamente aceitos, e sejam passíveis de serem abandonados novamente se for descoberto que levam à inconsistências, então podemos começar a Gödelizar esse sistema, assim como em qualquer outro. Estamos no mesmo caso de quando tínhamos uma regra gerando um conjunto infinito de fórmulas Gödelianas como axiomas. Em resumo, como quer que seja que uma máquina seja projetada, ela deve funcionar ou aleatoriamente ou de acordo com regras precisas. Na medida que seu processo é aleatório, nós não podemos ser mais espertos que ela: mas sua performance não será uma paródia convincente do comportamento inteligente: na medida em que é um processo de acordo com regras precisas, o método de Gödel pode ser usado para produzir uma fórmula que a máquina, de acordo com aquelas regras, não possa exibir como verdadeira, ainda que nós, posicionados fora do sistema, possamos ver que é verdadeira ²².

O teorema de Gödel aplica-se apenas à sistemas consistentes. Tudo o que podemos provar *formalmente* é que se o sistema é completo, então a fórmula Gödeliana indemonstrável no sistema. Para sermos capaz de dizer categoricamente que a fórmula Gödeliana não é demonstrável no sistema, e portanto verdadeira, devemos não apenas estar lidando com um sistema consistente, mas também sermos capaz de dizer que ele é consistente. E, como Gödel mostrou em seu segundo teorema – um corolário do primeiro – é impossível demonstrar em um sistema consistente que esse sistema é consistente. Portanto, para fins de criticar a máquina produzindo uma fórmula da qual possamos dizer que é verdadeira e que a máquina não pode mostrá-la como verdadeira, devemos ser capazes de dizer que a máquina (ou, ao invés disso, seu sistema formal correspondente,) é consistente; e não há prova absoluta disso. Tudo o que podemos fazer é examinar a máquina e ver se ela parece consistente. Sempre permanece a possibilidade de alguma inconsistência ainda não detectada. No máximo podemos dizer que a máquina é consistente, dado que que nós sejamos. Mas com que direito podemos dizê-lo? O segundo teorema de Gödel parece mostrar que um homem não pode declarar sua própria consistência, e assim Hartley Rogers²³ defende que não podemos realmente usar o primeiro teorema de Gödel para conter a tese mecanicista à menos que possamos dizer que “há atributos distintivos que habilitam um ser humano a transcender essa última limitação e declarar sua própria consistência enquanto ainda permanece consistente”.

A reação de um homem inculto, se sua consistência for questionada, é afirmá-la veementemente; mas isso, na visão do segundo teorema de Gödel, é tomado por alguns filósofos como evidência de sua inconsistência. O professor Putnam²⁴ tem sugerido que os seres humanos são máquinas, mas máquinas inconsistentes. Se uma máquina foi cabeada para corresponder a um

²²A prova original de Gödel aplicava-se se a regras fosse tal como gerar uma classe recursiva primitiva de fórmulas adicionais; v. § I init. e § 6 init. de suas Lectures t the Insittute of Advanced Study, Princiton, N. J., U.S.A, 1934. É de fato suficiente que a classe seja recursivamente enumerável. v. Barkley Rosser: “Extensions of some theorems of Gödel and Church”, Journal of Symbolic Logic, Vol. I, 1936, pp. 87-91.

²³Op. cit., p. 154.

²⁴Universidade de Princeton, N.J, U.S.A. Em conversa privada

sistema inconsistente, então não haveria fórmula bem formada que ela não pudesse demonstrar ser verdadeira; e assim não poderia de nenhuma forma ser provada como inferior a um ser humano. Nem poderíamos entender sua inconsistência como uma reprovação à ela – não seriam os homens também inconsistentes? Certamente mulheres são, e políticos; e mesmo homens que não sejam políticos contradizem a si mesmos algumas vezes, e uma única inconsistência é o bastante para tornar um sistema inconsistente²⁵.

O fato de que todos somos inconsistentes, por vezes, não pode ser discutido. Mas disso não se segue que somos equivalentes a um sistema inconsistente. Nossas inconsistências são enganos ao invés de políticas definidas. Elas correspondem a um mal funcionamento ocasional da máquina, não a seu esquema normal de operações. Sendo testemunhas desse fato, nós evitamos inconsistências quando as reconhecemos. Se fossemos realmente máquinas inconsistentes, deveríamos nos manter satisfeitos com nossas inconsistências e afirmar alegremente ambos os lados de uma contradição. Ademais, estaríamos preparados para dizer absolutamente qualquer coisa - o que não estamos. É fácil mostrar²⁶ que em um sistema formal inconsistente tudo é demonstrável, e o requerimento de consistência acaba por ser apenas de que não é tudo que é demonstrável nele – não é o caso de que “tudo vale”. Isso é certamente uma característica das operações mentais dos seres humanos: elas são seletivas, elas discriminam entre afirmações favoráveis - verdadeiras - ou desfavoráveis - falsas. Quando uma pessoa está disposta a dizer algo e a contradizer a si mesma sem qualquer escrúpulo ou repugnância, ela é dita ter “perdido a cabeça”. Seres humanos, ainda que não sejam perfeitamente consistentes, não são tão inconsistentes quanto falíveis.

Uma máquina falível mas auto-corrigível ainda estaria sujeita aos resultados de Gödel. Apenas uma máquina fundamentalmente inconsistente escaparia. Poderíamos nós termos uma máquina fundamentalmente inconsistente mas ao mesmo tempo auto-corretora que fosse tanto livre dos resultados de Gödel e que ainda pudesse não ser trivial e ser completamente diferente de um ser humano? Uma máquina com uma *recherché* [obscursa] inconsistência cabeada nela, de forma que para todos os propósitos normais ela seja consistente, mas se apresentada à sentença Gödeliana fosse capaz de prová-la?

Há muitas formas através das quais uma prova indesejada pode ser removida. Podemos ter uma regra onde sempre que tivermos demonstrado p e não- p , examinaremos as demonstrações e rejeitaremos a mais extensa. Ou podemos organizar os axiomas e as regras de inferência em uma certa ordem, e quando a demonstração que leva à uma inconsistência for apresentada, veremos quais axiomas e regras são necessários à ela e rejeitaremos àqueles que vem por último na ordenação. De uma forma como essa podemos ter um sistema inconsistente com uma regra de parada, assim uma inconsistência nunca seja permitida surgir na forma de uma fórmula inconsistente.

²⁵O tradutor não se responsabiliza pelas posições políticas, morais ou religiosas do autor do artigo em questão.

²⁶Veja, por exemplo, Alonzo Church: *Introduction to Mathematical Logic*, Princeton, Vol. I, § 15, p. 108.

A sugestão a primeira vista parece atrativa: entretanto há algo profundamente errado. Mesmo pensando que podemos preservar a fachada de consistência possuindo uma regra que sempre que duas fórmulas inconsistentes apareçam nós rejeitaremos aquela que tiver a maior prova, ainda assim essa regra seria repugnante ao nosso sentido lógico. Mesmo as sugestões menos arbitrárias são arbitrárias o bastante. Esse sistema não operaria mais com certas regras de inferência precisas sobre certas formulas definidas. Ao invés disso as regras aplicam-se, os axiomas são verdadeiros, desde que... nós não os achemos inconvenientes. Já não sabemos onde estamos. Uma aplicação da regra do Modus Ponens pode ser aceita enquanto outra é rejeitada; em uma ocasião um axioma pode ser verdadeiro, em outra, aparentemente falso. O sistema haverá deixado de ser um sistema de lógica formal, e a máquina mal irá qualificar-se como um modelo para a mente. Pois ela estará longe de parecer-se com a mente e suas operações: a mente de fato experimenta axiomas e regras de inferências dúbios; mas se for descoberto que levam a contradições, serão rejeitados. Experimentamos axiomas e regras de inferência provisoriamente – verdade : mas não os mantemos, uma vez que eles são descobertos levarem a contradições. Podemos procurar substituí-los por outros, podemos sentir que nossa formalização é falha, e ainda que pensemos que alguns axiomas ou regras de inferência desse tipo sejam necessários, podemos não ser capazes de formulá-los corretamente: mas nós não mantemos formulações falhas sem modificações, apenas com a condição de que quando o argumento levar a uma contradição nós recusaremos de segui-lo. Fazer isso seria completamente irracional. Deveríamos estar em posição de que em algumas ocasiões quando providos com premissas de um Modus Ponens, digamos, nós aplicássemos a regra e permitíssemos a conclusão, e em outras nós nos recusássemos a aplicar a regra e proibíssemos a conclusão. Uma pessoa, ou uma máquina, que fizesse isso sem ser capaz de dar uma boa razão para tal, seria considerada arbitrária e irracional. É parte do conceito de “argumentos” ou “razões” que se eles sejam em algum sentido gerais e universais: que se Modus Ponens é um método válido de argumentação quando estou estabelecendo uma conclusão, ele também será quando você, meu oponente, está estabelecendo uma conclusão que eu não quero aceitar. Nós não podemos escolher as vezes em que uma forma de argumento será válida; caso sejamos sensatos. E bem verdade, que com nossos argumentos informais, que não são completamente formalizados, nós realmente distinguimos entre argumentos que são a primeira vista similares, e adicionamos razões adicionais de porque eles não, no entanto, realmente similares; e pode ser defendido que uma máquina poderia similarmente estar no direito de distinguir entre argumentos à primeira vista similares, se ela possuir boas razões para fazê-lo. Pode, adicionalmente, ser defendido que uma máquina possui boas razões para rejeitar aqueles padrões de argumentos que ela rejeita, certamente sendo a melhor das razões evitar contradições. Mas isto, se de fato for uma razão, é uma muito boa. Nós não creditaríamos a um homem que ele evita contradições apenas por recusar-se a aceitar os argumentos que o levariam a elas e nada mais. O nome desse tipo de raciocínio é apelo especial [uma forma de falácia] ao invés de um argumento sólido. Nenhum crédito é adicionado a um homem que, sendo esperto o

bastante para ver alguns movimentos à frente no argumento, evita ser levado a reconhecer sua própria inconsistência sendo evasivo assim que percebe onde o argumento terminará. Ao contrário, nós o consideraremos igualmente inconsistente; não, nesse caso, devido a ele afirmar e negar a mesma proposição, mas porque ele fez uso e rejeito a mesma regra de inferência. A regra de interromper-se para não enunciar efetivamente uma inconsistência não é o bastante para salvar uma máquina inconsistente de ser chamada de inconsistente.

A possibilidade ainda mantém-se de que nós sejamos inconsistentes, e que não haja regra de parada, mas que a inconsistência seja tão *obscura* que ela nunca tenha se mostrado. Afinal de contas, a teoria de conjuntos *ingênua*, que fora profundamente incorporada nas formas de pensar do senso comum, acabou por mostrar-se inconsistente. Podemos nós estarmos certos de que um destino similar não está guardado também para a aritmética simples? Em um sentido não podemos, apesar de nosso grande sentimento de certeza de que nosso sistema dos números naturais que podem ser adicionados e multiplicados uns com os outros, nunca mostrar-se-á inconsistente. É concebível que possamos descobrir que o formalizamos incorretamente. Caso houvermos, nós podemos tentar reformular novamente nosso conceito intuitivo de número, assim como fizemos com nosso conceito intuitivo de conjunto. Caso fizermos isso devemos obviamente reformular o sistema, estando na mesma posição em que estamos agora, possuindo um sistema que acredita-se ser consistente, mas que não o tenha sido demonstrado. Dessa forma, não poderia haver então outra inconsistência? Essa é certamente uma possibilidade. Mas novamente nenhuma inconsistência que uma vez tenha sido detectada será tolerada. Estamos determinados a não ser inconsistentes, e estamos decididos a erradicar inconsistências, caso alguma apareça. Portanto, ainda que nunca possamos estarmos completamente certos ou completamente livres do risco de termos de repensar nossa matemática novamente, a posição definitiva deve ser uma entre duas: ou nós temos um sistema de aritmética simples que segundo o melhor de nosso conhecimento e crença é consistente; ou tal sistema não é possível. No primeiro caso estaríamos na mesma posição que estamos atualmente, na última, se descobrirmos que nenhum sistema contendo a aritmética simples pode estar livre de contradições, teremos que abandonar não apenas apenas toda a matemática e as ciências matematizadas, mas todo o pensamento.

Ainda pode ser defendido que mesmo que um homem deva nesse sentido assumir, ele não pode propriamente afirmar sua própria consistência sem conseqüentemente desmentir suas palavras. Podemos ser consistentes, na verdade temos toda a razão para termos esperanças de que somos, mas uma modéstia necessária nos proíbe de dizê-lo. No entanto, isso não é bem o que o segundo teorema de Gödel afirma. Gödel mostrou que em um sistema consistente a fórmula que afirma a consistência do sistema não pode ser demonstrada *naquele sistema*. Segue-se disso que uma máquina, se consistente, não pode apresentar como verdadeira uma afirmação de sua própria consistência; conseqüentemente uma mente, *se ela for realmente uma máquina*, não pode alcançar a conclusão de que ela é consistente. Para uma mente que não é uma máquina essa conclusão não se

segue. Tudo o que Gödel provou é que uma mente não pode produzir uma demonstração formal da consistência de um sistema formal dentro do próprio sistema; mas não há objeção em se ir fora do sistema ou à produção de argumentos informais tanto para a consistência do sistema formal ou para algo menos formal e menos sistematizado. Tais argumentos informais não serão passíveis de serem completamente formalizados, mas todo o teor dos resultados de Gödel é de que não devemos perguntar, e não podemos obter, uma formalização completa. E ainda que pudesser ter sido legal se fossemos capazes de obtê-los, visto que argumentos completamente formalizados são mais coercivos do que os informais, ainda assim visto que nós não podemos ter todos os nossos argumentos moldados naquela forma, nós não podemos objetar argumentos informais por serem eles serem informais ou considerá-los sem valor. Parece-me então ser tanto apropriado quanto razoável que uma mente declare sua própria consistência. Apropriado pois ainda que máquinas, como poderíamos esperar, sejam incapazes de refletir completamente sobre sua própria performance e capacidade, podemos esperar que uma mente autoconsciente seja. E razoável pelas razões dadas. Nós não apenas podemos dizer simplesmente que sabemos que somos consistente, apesar de nossos enganos, mas devemos *pressupor* que somos, caso o pensamento seja possível. Ademais, somos seletivos. Não iremos, como uma máquina inconsistente faria, dizer quaisquer coisas [arbitrárias] que sejam. E finalmente, nós podemos, em um sentido, *decidir* sermos consistentes, no sentido de que podemos resolver não tolerar inconsistências em nosso pensamento e fala e eliminá-los caso esses surjam retirando e cancelando um dos membros da contradição.

Podemos ver como poderíamos quase esperar que o teorema de Gödel distinguisse seres autoconscientes de objetos inanimados. A essência da fórmula Gödeliana é de que ela é autorreferente. Ela diz que “Esta fórmula é indemonstrável neste sistema”. Quando aplicada por uma máquina, a fórmula é especificada em termos que dependem da máquina particular em questão. A máquina é está recebendo um questionamento sobre seus próprios processos. Nós estamos pedindo a ela para ser autoconsciente e dizer coisas que ela pode e não poder dizer. Tais questões notoriamente levam à paradoxos. Na primeira e mais simples tentativa de filosofar, nos tornamos enredados em questões de se quando alguém sabe algo, esse sabe que sabe isso. E o que, quando alguém está pensando sobre si mesmo, está sendo pensado sobre, e o que está realizando o pensamento. Depois de se estar perplexo e ter sido atingido por esse problema por um longo tempo, aprende-se a não pensar sobre essas questões: o conceito de um ser consciente é, implicitamente, percebido ser diferente daquele de um ser objeto não consciente. Ao dizermos que um ser consciente sabe algo, estamos dizendo não apenas que ele o sabe, mas que ele sabe que sabe isso, e que ele sabe que sabe que sabe isso, e assim por diante tanto quanto nos importarmos o bastante para fazer essa questão. Há, nós reconhecemos, um infinito aqui, mas não é um regresso infinito no mau sentido, pois são as questões que esgotam-se, por serem inúteis, e não as respostas. As questões parecem ser inúteis porque o conceito contém em si mesmo a ideia de ser capaz de continuar respondendo tais questões indefinidamente. Ainda que seres conscientes tenham a capacidade de

continuarem, não queremos exibir isso como simplesmente como uma sucessão de tarefas que eles são capazes de realizar, nem mesmo vemos a mente como uma sequência infinita de eus e super-eus e super-super-eus. Ao contrário, insistimos que um ser consciente é uma unidade, e ainda que falemos sobre partes da mente, fazemos isso apenas como uma metáfora, e não permitimos que seja tomado literalmente.

Os paradoxos surgem porque um ser consciente está ciente de si mesmo, assim como de outras coisas, e assim não pode realmente ser interpretados como um ser divisível em partes. Isso significa que um ser consciente pode lidar com questões Gödelianas de uma forma que uma máquina não pode, porque um ser consciente pode tanto considerar a si mesmo e sua performance e ainda assim não ser outro que aquele que realiza a performance. Uma máquina pode ser construída, podemos dizer, de forma a “considerar” sua própria performance, mas ela não pode levar isso “em conta” sem, devido a isso, tornar-se uma máquina diferente. Isso é, a velha máquina com uma “nova parte” adicionada. Mas é inerente a nossa ideia de uma mente consciente que ela pode refletir sobre si mesma e criticar seu próprio desempenho, e nenhuma parte extra é necessária para que isso seja feito: ela já é completa, e não há um calcanhar de Aquiles.

A tese passa então a ser mais uma questão de análise conceitual do que de descoberta matemática. Isso surge da consideração de outro argumento apresentado por Turing²⁷. Até agora, nós construímos apenas artefatos simples e previsíveis. Quando aumentarmos a complexidade de nossas máquinas elas podem, talvez, ter surpresas reservadas a nós. Ele traça um paralelo com uma quantidade de fissão. Abaixo de um certo tamanho “crítico” quase nada ocorre, mas acima dele as faíscas começam a voar. Assim também, talvez, seja com cérebros e máquinas. A maioria dos cérebros e máquinas são, atualmente, “sub-críticas” - eles reagem ao estímulo que chega de uma forma enfadonha e desinteressante, não possuindo ideia de si mesmo, podendo produzir apenas respostas armazenadas – mas alguns poucos cérebros atualmente, e possivelmente algumas máquinas no futuro, são super-críticas e cintilam por sua própria conta. Turing está sugerindo que é apenas uma questão de complexidade, que acima de um certo nível de complexidade uma diferença qualitativa aparece, de forma que as máquinas “super-críticas” serão bem diferentes das simples que concebemos até agora.

Pode ser assim mesmo. A complexidade frequentemente introduz diferenças qualitativas. Entretanto isso soa implausível, poderia ocorrer que acima de um certo nível de complexidade, a máquina pare de ser previsível, mesmo em princípio, e comece a realizar coisas por conta própria, ou, para usar uma frase bem divulgada, ela pode começar a ter uma mente própria. Ela pode começar a ter uma mente própria. Ela pode começar a ter uma mente própria quando não mais for inteiramente previsível e inteiramente dócil, mas for capaz de realizar coisas que nós reconheçamos como inteligentes, e não apenas enganos ou disparos aleatórios, e que não tenhamos programado nela. Mas então ela deixaria de ser uma máquina, na acepção do ato. O que está em jogo no debate

²⁷Mind, 1950, p. 454; Newman, p. 2117-18.

do mecanicismo não é como mentes vem ou podem vir à ser, mas como elas operam. É essencial para a tese mecanicista que um modelo mecânico da mente opere de acordo com “princípios mecânicos”, isso é, que nós possamos entender as operações do todo em termos de operações de suas partes, e as operações de cada parte ou sejam determinadas por seu estado inicial e pela construção da máquina, ou sejam escolhas aleatórias entre um número determinado de operações também determinadas. Se o mecanicista produzir uma máquina que é tão complicada que a posição deixe de aplicar-se a ela, então isso não será mais uma máquina para os propósitos de nossa discussão, não importa como ela seja construída. Deveríamos dizer, ao invés, que construímos uma mente, no mesmo sentido que dizemos quando produzimos pessoas atualmente. Poderia então haver duas formas de trazer mentes ao mundo, a forma tradicional, produzindo crianças nascidas de mulheres, e uma nova forma ao construir sistemas muito, muito complicados de, digamos, válvulas e relés. Quando falarmos da segunda forma, devemos tomar cuidado em reforçar que ainda que o que criamos pareça-se com uma máquina, na verdade não o é, porque não é um total de suas partes. Não se poderia dizer o que ela poderia fazer apenas conhecendo a forma como e foi construída e o estado inicial de suas partes. Não poderia mesmo dizer os limites do que ela pode fazer, pois mesmo quando exposta a uma questão de tipo Gödeliano, ela pode dar a resposta correta. Na verdade podemos dizer brevemente que qualquer sistema que não seja assolado pela questão de Gödel não é *eo ipso* uma máquina de Turing, isto é, não é uma máquina no sentido do termo.

Se a prova da falsidade do mecanicismo for válida, isso terá grande consequência para o todo da filosofia. Desde o tempo de Newton o fantasma do determinismo mecanicista tem obcecado os filósofos. Caso quisermos ser científicos, me parece que devemos olhar para os seres humanos como automatizados determinados, e não como agentes morais autônomos; se quisermos ser morais, me parece que devemos negar à ciência seu dever, estabelecer um limite arbitrário ao progresso no entendimento neurofisiológico humano e nos refugiarmos no misticismo obscurantista. Nem mesmo Kant foi capaz de resolver a tensão entre os dois pontos de vista. Mas agora, ainda que muitos argumentos contra a liberdade humana continuem, o argumento a partir do mecanicismo, talvez o mais convincente argumento de todos, há perdido sua força. Não mais nessa questão será incumbido ao filósofo natural de negar a liberdade em nome da ciência. Não mais o moralista terá que sentir o desejo de abolir o conhecimento para deixar espaço para a fé. Podemos mesmo começar a ver como pode haver espaço para a moralidade, sem ser necessário abolir ou mesmo limitar o domínio da ciência. Nosso argumento não põe limites à investigação científica: ainda será possível investigar o funcionamento do cérebro. Ainda será possível produzirmos modelos mecânicos da mente. Apenas, podemos agora ver que nenhum modelo mecânico será completamente adequado, ou mesmo qualquer explicação em termos unicamente mecanicistas. Podemos criar modelos e explicações que serão iluminadoras, mas, independente de quão longe elas forem, sempre haverá mais a ser dito. Não há limite arbitrário para a investigação científica, mas nenhuma investigação científica jamais poderá esgotar a infinita multiplicidade da mente humana.

Merton College, Oxford.

DISCUSSÃO

Os dois artigos que compõem essa dissertação apresentam objeções à teoria computacional da mente. Em ambos são apresentados argumentos que buscam mostrar capacidades da mente humana que estariam ausentes em sistemas computacionais, e em ambos procurou-se apontar possíveis erros na interpretação computacionalista.

No primeiro artigo foi apresentado o argumento de Lucas em defesa da tese de que a mente humana é superior a sistemas computacionais. Analisou-se especialmente seu argumento de que a mente humana não pode ser um sistema inconsistente visto seu funcionamento ordinário. Como objeção, foi lançada a hipótese de que a mente humana possa ser um sistema inconsistente desde que sua lógica subjacente seja uma lógica da inconsistência formal e não a lógica clássica.

No segundo artigo foram identificadas as posições filosóficas alvo do experimento de pensamento do quarto chinês de Searle. Adicionalmente apontamos as aparentes falhas na compreensão dessas teses cometidas pelo autor. Subsequentemente, mostrou-se como tais falhas exerceram influência nos ataques à tese computacionalista realizadas por Searle e nas respostas que o autor dá a seus críticos.

Buscou-se mostrar que apesar das críticas com base em argumentos de natureza tão distintas quanto resultados metamatemáticos e experimentos de pensamento, a tese computacionalista não é tão vulnerável como poderia parecer em uma primeira análise. Apesar de ser considerada refutada por algumas comunidades filosóficas, a posição computacionalista, se devidamente compreendida, parece sobreviver as objeções mais persuasivas.

CONCLUSÃO

Em um sentido geral, esse não foi um trabalho de desenvolvimento propriamente dito da tese computacionalista. Ao contrário, podemos melhor caracterizá-lo como uma análise de dois dos argumentos anti-computacionalistas mais influentes na comunidade filosófica e da busca por respostas e alternativas aos problemas apontados. Mesmo onde hipóteses positivas foram esboçadas, destaca-se a necessidade de pesquisa futura para análise, sofisticação, adequação empírica e avaliação dos méritos e deméritos dessas.

A Hipótese da mente paraconsistente, por exemplo, que criei como uma alternativa ao argumento de impossibilidade da mente humana ser um sistema inconsistente, dá margem para pesquisa futura em temas extremamente sofisticados para sua adequação e avaliação, tais como: computação

paraconsistente, máquinas de Turing paraconsistentes, modelos inconsistentes da aritmética, incompletude em sistemas paraconsistentes, entre tantos outros. Questões essas que serão devidamente trabalhadas em pesquisas futuras sobre o tema.

REFERÊNCIAS

GÖDEL, K. Some basic theorems on the foundations of mathematics and their implications. In: FERERMAN (ed). **Kurt Gödel collected works volume III: Unpublished essays and lectures**. Oxford: Oxford University Press, p. 304-423, 1995.

FODOR, J. **The language of thought**. New York: Thomas Y. Crowell, 1975.

GÖDEL, K. **On Formally undecidable propostions of Principia Mathematica and related systems**. New York: Dover Publications, 1992.

GÖDEL, K. Some basic theorems on the foundations of mathematics and their implications. In: FERERMAN (ed). **Kurt Gödel collected works volume III: Unpublished essays and lectures**. Oxford: Oxford University Press, p. 304-423, 1995.

MCCULLOCH, W; PITTS, W. A logical calculus of the ideias imannent in nervous activity. **Bulletin of mathematical biophysics**, v. 7. p. 115–133, 1943.

PICCININI, G. The resilience of computationalism. **Philosophy of Science**. v. 77. n. 5. pp. 852-861, 2010.

PUTNAM, H. The nature of mental states. In: Chalmers, D (ed). **Philosophy of mind: Classical and contemporary readings**. Oxford: Oxford University Pres, p. 73- 79, 2002.

PUTNAM, H. The nature of mental states. In: **Hilary Putnam Philosophical papers volume 2: Mind, language and reality**. Cambridge: Cambridge University Press, p, 429-440, 1975.

SEARLE, J. Minds, brains, and programs. **Behavioral and brain sciences**, v. 3, n. 3, 417-424, 1980.