ISSN 2965-1557
http://doi.org/10.1590/2965-1557.037.e202532317

Artigo científico

Revista de Filosofia Aurora

periodicos.pucpr.br/aurora

PUCPRESS

Argumento da Sala Chinesa revisitado: considerações a partir da IA generativa

Chinese room argument revisited: considerations from generative AI

Kleber Bez Birolo Candiotto [a]
Curitiba, PR, Brasil
[a] Pontifícia Universidade Católica do Paraná (PUCPR)

Como citar: CANDIOTTO, Kleber Bez Birolo. Argumento da Sala Chinesa revisitado: considerações a partir da IA generativa. *Revista de Filosofia Aurora*, Curitiba: Editora PUCPRESS, v. 37, e202532317, 2025. DOI: https://doi.org/10.1590/2965-1557.037.e202532317

Resumo

Neste artigo, revisitamos o debate instigado por John Searle em seu argumento da Sala Chinesa, no que tange à viabilidade da inteligência artificial forte, à luz do cenário contemporâneo de desenvolvimento da Inteligência Artificial Generativa, com os denominados Grandes Modelos de Linguagem (LLMs), especificamente o ChatGPT da OpenAI. Reexaminamos elementos cruciais do argumento searlano, os quais têm evoluído ao longo das últimas quatro décadas, em resposta às críticas recebidas. Nosso objetivo é investigar os limites do que pode ser alcançado por entidades, sejam elas de natureza humana ou mecânica, que se destacam no domínio linguístico.

Palavras-chave: Linguagem. Semântica. Inteligência Artificial. LLMs.

Abstract

In this article, we revisit the debate instigated by John Searle in his Chinese Room argument, regarding the viability of strong artificial intelligence, in light of the contemporary scenario of development of Generative

[a] Doutor em Filosofia pela Universidade Federal de São Carlos, e-mail: kleber.c@pucpr.br

Artificial Intelligence, with the so-called Large Language Models (LLMs), specifically OpenAI's ChatGPT. We re-examine crucial elements of Searlan's argument, which have evolved over the past four decades in response to criticism. Our objective is to investigate the limits of what can be achieved by entities, whether human or mechanical in nature, that stand out in the linguistic domain.

***Keywords**: Language. Semantics. Artificial intelligence. LLMs.*

Introdução

A "Sala Chinesa" é um experimento mental elaborado por John Searle em *Minds, brains and programs* (searle, 1980) com o intuito de criticar equivocadas conclusões decorrente do Teste de Turing. Sua crítica é dirigida ao que ele denomina por inteligência artificial forte, a teoria da mente que afirma que "a mente é apenas um programa de computador" (Searle, 1997, p. 09), ou melhor, a visão de que "um computador digital adequadamente programado com as entradas e saídas corretas, que satisfaça o teste de Turing, necessariamente teria uma mente" (Searle, 1999, p. 116). Neste clássico artigo, Searle procura explicitar a inabilidade das teorias computacionais da mente para compreender verdadeiramente os processos psicológicos, devido à sua natureza puramente sintática, incapaz de levar à semântica. Conforme explicado por Steven Harnad (1989, p. 05), essa questão central "*turns out to be an empirical question about the scope and limits of the purely symbolic (computational) model of the mind*".

Searle admite a existência da inteligência artificial (IA) numa perspectiva fraca e apresenta exemplos disso ao operarmos uma calculadora, ligarmos um computador ou utilizarmos qualquer maquinário com propósito específico. Entretanto, o que Searle procura rechaçar é a existência da IA forte, principalmente baseado em sua posição de que a maquinaria sintética carece dos poderes causais do cérebro biológico necessários para a cognição legítima.

Sobre a viabilidade da inteligência artificial em uma perspectiva fraca, Searle afirma: "*the computer is a useful tool, in doing simulations of the mind, as it is useful in doing simulations of just about anything we can describe precisely, such as weather patterns or the flow of money in the economy*" (Searle, 1997, p. 09). Uma ferramenta é definida como um instrumento ou meio empregado para atingir um resultado específico ou desejado. Analogamente, uma ferramenta, como uma caneta, não possui a capacidade de escrever por si só, mas é empregada como meio para registrar algo em um papel, que constitui o produto final almejado. Nesse sentido, a ferramenta é subordinada ao propósito de alcançar o fim desejado, carecendo de habilidade intrínseca, porém, desempenhando um papel crucial na consecução do resultado almejado.

O interesse de Searle, portanto, não repousa na chamada "IA fraca". Ele considera que essa abordagem da inteligência artificial não representa um desafio significativo que justifique uma exploração mais aprofundada. Em vez disso, o foco de sua análise recai sobre a abordagem da "IA forte", a ideia de uma inteligência artificial que transcende a mera simulação de comportamentos ou respostas humanas, adquirindo a capacidade de compreensão e consciência autênticas. Searle se propõe a desafiar a noção de que a "IA forte" poderia, de fato, exibir consciência genuína, além de questionar a validade das alegações que sustentam essa possibilidade. Diante disso, Searle encampa uma postura crítica, buscando refletir acerca dos limites e fundamentos subjacentes às concepções de inteligência artificial, especialmente no que tange à viabilidade da emergência de uma consciência verdadeira em sistemas computacionais.

Searle argumenta que, quando uma máquina é aprovada no Teste de Turing, isso não comprova que ela possua uma consciência. Ao invés disso, a máquina apenas demonstra a habilidade de simular uma conversa com um ser humano, podendo dar a impressão de compreender o uso da linguagem humana, mas, na realidade, apenas produzindo essa impressão. A máquina, de fato, não compreende o uso da linguagem humana; ela simplesmente a menciona. O motivo para tal situação reside no fato de que o uso da linguagem humana demanda o conhecimento de como se referir ao mundo externo. Portanto, o que Searle quer dizer é que a máquina não é dotada de Intencionalidade.

A recente repercussão da ascensão dos grandes modelos de linguagem (*Large Language Models*, LLMs) tem suscitado a conjectura de que, se uma entidade for dotada de proficiência linguística, então deve, por extensão, possuir a capacidade de raciocínio e, consequentemente, de compreensão. Alguns exemplares desses modelos, como o ChatGPT da OpenAI, conseguem gerar textos tão excepcionais que desafia a distinção entre a produção humana e a gerada por máquinas. Em decorrência dessa revolução linguística, têm ressurgido antigas questões tanto no âmbito da mídia popular quanto na esfera acadêmica, insinuando que os LLMs não apenas representam uma notável inovação no processamento de linguagem, mas, de maneira mais ampla, apontam para um avanço significativo em direção à realização da IA Forte (a aludida IA Geral), o que sinalizaria um passo concreto em direção a uma "máquina pensante".

O presente artigo procura retomar o debate promovido Searle com seu argumento da Sala Chinesa sobre a viabilidade da IA forte, considerando o contexto contemporâneo do desenvolvimento da denominada IA generativa, com escopo nos LLMs. Serão retomados alguns pontos fundamentais do argumento que foram aprimorados mediante as críticas recebidas ao longo das quatro décadas que sucederam sua formulação original, com a intenção de analisar os limites do que é possível estabelecer para uma entidade, seja ela humana ou mecânica, que se destaca na esfera linguística.

O argumento e suas consequências

Em seu clássico artigo *Minds, Brains and Programs*, John Searle (1980) apresenta seu famoso experimento mental da "Sala Chinesa", que tem desempenhado um papel central na discussão sobre inteligência artificial. A motivação original de Searle para a formulação desse experimento foi a crítica das conclusões frequentemente mal interpretadas do Teste de Turing[1]. Especificamente, ele concentrou sua crítica naquilo que ele denomina como "inteligência artificial forte", uma teoria da mente que propõe que a mente é apenas um programa de computador.

No seu experimento da "Sala Chinesa", Searle imagina-se trancado em uma sala com um manual altamente detalhado e extenso para construir frases em chinês corretas. Todas as regras necessárias para construir adequadamente tais frases estão disponíveis nesse manual, que opera como um computador com lógica de entrada e saída. Este manual é tão abrangente que, se um falante nativo de chinês estivesse fora da sala e lhe enviasse perguntas, John Searle, trancado em sua sala com este manual, seria capaz de responder a essas perguntas utilizando padrões estabelecidos[2]. Ou seja, ele domina o chinês em um nível sintático, sabendo como compor palavras, organizá-las e qual estrutura de frase construir em resposta a determinado tipo de estrutura de frase. As orientações de Searle referem-se à manipulação dos caracteres chineses. Uma vez que Searle não possui proficiência no idioma chinês, os caracteres apresentam-se, como ele mesmo diz,

[1] A proposta original do teste de Turing foi apresentada no clássico artigo de Alan Turing "Computing Machinery and Intelligence", publicado em 1950 na revista "Mind". Nesse artigo, Turing detalha sua ideia do que agora é conhecido como o "Teste de Turing", que foi uma tentativa de responder à pergunta: "As máquinas podem pensar?" Trata-se de uma simulação de conversação entre um ser humano e uma máquina, sem que o examinador saiba qual é qual. Se o examinador não puder distinguir, com base nas respostas, qual interlocutor é humano e qual é a máquina, a máquina é considerada "inteligente". Com o objetivo de explorar a capacidade de uma máquina em imitar o comportamento humano de tal forma que seja indistinguível de uma conversa com um ser humano real, o teste levanta questões filosóficas relevantes sobre a natureza da inteligência e da consciência. "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 (C) 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" (TURING, 1050, p.433).

[2] Como mencionado por Searle, "partisans of strong AI claim that in this question and answer sequence the machine is not only simulating a human ability but also 1. that the machine can literally be said to understand the story and provide the answers to questions, and 2. that what the machine and its program do explains the human ability to understand the story and answer questions about it" (1980, p.417).

como meros "rabiscos sem sentido"[3]. Em realidade, esses caracteres constituem-se em textos e questionamentos em chinês, provenientes da entrada do sistema. Empregando as regras estabelecidas, Searle procede à manipulação dos caracteres e produz cada resposta correspondente. Após obter uma resposta, ele a direciona por meio da abertura de saída. É notável que as respostas emitidas são em chinês e são atribuíveis à Searle, sendo estas de alta qualidade. Com efeito, as respostas são tão proficientes que ninguém pode distinguir Searle como não sendo um falante nativo de chinês.

Embora Searle domine o chinês em um nível sintático, nada comprova que ele compreende o significado do que está dizendo, visto que lhe falta o entendimento semântico da linguagem. Ele não sabe a que as palavras se referem. Em outros termos, nada evidencia que Searle pense o que diz: "*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*" (1980, p. 418). Para de fato expressar-se, para falar uma língua no sentido pleno, segundo Searle, é imprescindível possuir a capacidade de refletir sobre o que está sendo dito. Isso não implica que a pessoa precise sempre refletir sobre suas palavras ao falar uma língua, porém, é necessário ser capaz de ponderar sobre o que está sendo dito.

As reivindicações fundamentais expressas no artigo de 1980 de Searle são contundentes. A afirmativa postulada é que um dispositivo é considerado um computador digital devido à sua capacidade de realizar cálculos. Entretanto, sustenta-se que as computações *per se* não podem, em princípio, dar origem à cognição autêntica. Uma vez que a computação constitui meramente a manipulação de símbolos conforme regras puramente formais ou sintáticas, alega-se que um dispositivo restrito à computação "*cannot be said to have access to or know or understand the 'content', the semantic properties (meaning, interpretation) of the symbols it happens to be manipulating*" (Preston, 2002, p. 19). A computação não pode, portanto, explicar a cognição. Embora algo possa simular desempenhos inteligentes unicamente em virtude de realizar cálculos, não possui a capacidade de duplicá-los. Independentemente da qualidade da tecnologia ou da rapidez com que máquinas realizam cálculos, "*if it really is a computer, its operations have to be defined syntactically, whereas consciousness, thoughts, feelings, emotions, and all the rest of it involve more than syntax*" (Searle, 1984, p. 37).

O experimento da "Sala Chinesa" apresenta-se como um cenário emblemático para a discussão e questionamento das capacidades de processamento linguístico de um sistema, uma vez que, segundo Searle, é desprovido de compreensão semântica e consciência. As diretrizes delineadas pelo experimento permitem a manipulação e produção de respostas coerentes em chinês, ressaltando a natureza sintática do processamento da linguagem e evidenciando a ausência de compreensão genuína por parte do sistema. A partir da análise proposta por Searle, emergem reflexões críticas quanto aos limites inerentes a sistemas computacionais, no que concerne à verdadeira compreensão e experiência subjetiva.

The Robot Reply: crítica ou fortalecimento do argumento da Sala Chinesa?

Na publicação de 1980, Searle foi alvo direto de seis réplicas provenientes de universidades americanas, às quais ele prontamente respondeu. No âmbito deste artigo, destaca-se a *Robot Reply* da Universidade de Yale, a qual possui maior relevância para o presente texto, visto que contempla a possibilidade de um computador alojado em um robô, com capacidade para emular atividades humanas como "perceber, andar, mover-se, martelar pregos, comer, beber..." (1980, p. 420), bem como interagir, tal

[3] "To me, Chinese writing is just so many meaningless squiggles" (Searle, 1980, p.418)

qual os *chatbots* de inteligência artificial generativa. Segundo essa réplica, um computador capaz de interagir com o ambiente externo seria dotado de compreensão genuína e outros estados mentais. Dessa forma, a substituição do programa de inteligência artificial desencarnado por um robô equipado com sensores e mecanismos análogos aos membros humanos permitiria que tal robô fosse considerado "inteligente" de maneira comparável a um ser humano. Nessa perspectiva, a *Robot Reply* de Yale procura expandir o debate e aprofundar a reflexão sobre a viabilidade de atribuir compreensão e estados mentais a sistemas computacionais complexos, à medida que interagem com o mundo externo. A introdução de sensores e mecanismos robóticos realça a ideia de que a experiência e a interação com o ambiente são fundamentais para a emergência de inteligência.

Searle rechaça esta possibilidade valendo-se do centro do próprio argumento da réplica que "*tacitly concedes that cognition is not soley a matter of formal symbol manipulation, since this reply adds a set of causal relation with the outside world*" (Searle, 1980, p. 420). Mais uma vez, Searle emprega a Sala Chinesa, agora devidamente modificada, com o intuito de demonstrar que a incorporação de capacidades robóticas "perceptivas" e "motoras" não adiciona compreensão ao programa original. É importante ressaltar que, mesmo quando uma pessoa é inserida em um computador que, por sua vez, está contido em um robô, ela ainda se depara apenas com símbolos a serem manipulados, sem possibilidade de compreender o significado intrínseco desses símbolos. Embora o resultado desse arranjo apresente-se de maneira notavelmente impressionante em seu comportamento, em relação a um computador estático e inerte, isso meramente comprova sua habilidade em transpor um nível mais sofisticado de Teste de Turing. Contudo, a ênfase de Searle reside no fato de que o Teste de Turing é sempre inadequado, uma vez que tais avaliações frequentemente fracassam ao estabelecer a existência de fenômenos psicológicos genuínos.

A aplicação da Sala Chinesa de forma adaptada, que incorpora elementos robóticos de percepção e movimentação, visa aprofundar a análise da problemática da compreensão genuína em sistemas computacionais complexos. A inserção de uma pessoa no contexto, embora possa parecer uma abordagem mais avançada, não altera o cerne do desafio enfrentado. Ainda se trata, em última instância, de manipulação de símbolos sem que se estabeleça a verdadeira compreensão semântica. A aparente proeza do arranjo em sua execução sugere uma sofisticação no desempenho do Teste de Turing, entretanto, Searle alerta para os limites dessa abordagem, enfatizando sua inadequação em comprovar a presença de processos psicológicos autênticos. Essa investigação instiga reflexões sobre a natureza da cognição e das relações entre processamento computacional e compreensão consciente. A controvérsia em torno dos testes e das possibilidades da inteligência artificial se aprofunda com o desafio proposto por Searle, direcionando a atenção para aspectos cruciais do entendimento humano que transcendem a mera capacidade de simulação em computadores.

Críticas oriundas da réplica do robô, como a de Moravec (1999) [4], por exemplo, fundamentam-se na suposição de que robôs cada vez mais inteligentes tendem a invalidar o argumento da "Sala Chinesa". Por meio de um novo experimento mental intitulado "*the missing thought-experime*", Brigsjorn e Noel (2002) buscam reforçar a resposta de Searle e, consequentemente, afirmam que o progresso da inteligência artificial

[4] O paradoxo de Moravec, cunhado pelo pesquisador de robótica Hans Moravec, expõe uma complexa dissonância na capacidade das máquinas em comparação com a dos seres humanos. Essa dissonância se manifesta notavelmente nas tarefas que, embora triviais para os humanos, se mostram extremamente complexas (ou até impraticáveis) para as máquinas, ao passo que atividades que são naturalmente desafiadoras para os seres humanos são executadas com relativa facilidade por essas mesmas máquinas. Por exemplo, a capacidade humana de interagir habilmente no mundo físico e de compreender os matizes das interações sociais representa uma destreza que as máquinas ainda não conseguem replicar de forma adequada. Em contrapartida, tarefas que exigem análise de dados em larga escala ou cálculos matemáticos complexos são realizadas de maneira eficiente pelas máquinas, superando as capacidades humanas. Essa ideia foi apresentada em seu conhecido livro *Mind Children: The Future of Robot and Human Intelligence* de 1988.

na construção de robôs é simplesmente um avanço na concepção de "animais zumbis". Com a concepção de "animais zumbis", eles questionam a capacidade dos sistemas de inteligência artificial de realmente alcançar um nível autêntico de compreensão e interação com o ambiente.

Esse novo argumento emerge como uma réplica à atualização do "*robot replay*" proposto por Harnad (1991), conhecido como o Teste Total de Turing. Tal teste estabelece que, para ser considerada aprovada, uma máquina precisa não apenas exibir um comportamento linguístico convincente, mas também demonstrar um comportamento sensório-motor convincente. Searle, em sua busca por ampliar a complexidade de seu experimento mental, argumenta que a hipótese formulada por Harnad, embora possa parecer adicionar alguns elementos à "*reply robot*" de Yale, não altera em nada a essência do que o experimento original da Sala Chinesa tinha por objetivo apresentar.

Nesta oportunidade, Searle (Brigsjorn; Noel, 2002, p. 156) sugere imaginar um robô, cujo cérebro consiste em um sofisticado computador localizado em uma Sala Chinesa dentro de sua estrutura craniana. No entanto, em vez do computador comercial, Searle pede para que ele assuma o papel do cérebro do robô, operando seus programas e coordenando suas ações. Esse robô está equipado com todos os transdutores sensoriais e motores necessários para coordenar sua entrada e saída no mundo ao seu redor. Enquanto Searle está na Sala Chinesa, conduzindo essa coordenação, paradoxalmente, ele nada sabe sobre o que está acontecendo. Searle convida a imaginar também que, entre os transdutores do robô, existam dispositivos capazes de converter estímulos visuais em símbolos chineses. Esses símbolos chineses são então enviados para a Sala Chinesa, onde está Searle, atuando nesses símbolos de acordo com um conjunto de regras pré-estabelecido: o programa. Por meio dessa operação, envia símbolos aos transdutores que acionam a saída motora do robô. A saída motora, por sua vez, é uma declaração em chinês, afirmando, por exemplo, "acabei de ver uma foto de um Buda grande e gordo". É crucial destacar que, apesar do aparente processamento linguístico sofisticado e da declaração em chinês, não houve qualquer experiência consciente de um Buda ou qualquer outro objeto na situação. Nada disso foi vivenciado conscientemente, nem pelo robô e nem por Searle na Sala Chinesa. O que de fato ocorreu foi a conversão dos estímulos luminosos em símbolos, o processamento desses símbolos por Searle e a posterior transformação em um enunciado auditivo. Em outras palavras, não houve experiência visual ou consciência envolvida na sequência de eventos. Portanto, ainda que seja possível dispor de uma infinidade de transdutores e submeter o robô ao Teste Total de Turing repetidas vezes, não se pode garantir a emergência de experiências autênticas, compreensão genuína ou outros estados mentais relevantes.

Há como escapar do "Sala Chinesa"?

Quanto à questão da linguagem, presente no argumento de Searle, umas das críticas mais conhecidas e relevantes para esta investigação foi de Margaret Boden, em *Escaping from the Chinese room* (1988). Com o objetivo de contrapor a posição de Searle, de que a inteligência artificial não pode ser verdadeiramente inteligente, pois não possui compreensão ou intencionalidade, Boden procura sustentar que a inteligência artificial pode sim ter intencionalidade e que a compreensão é uma questão de grau. Para isso, entre as principais questões do seu texto, Boden sugere a "*English Reply*", que é baseada na ideia de que a compreensão é uma questão de grau e que a intencionalidade pode ser alcançada por meio de sistemas computacionais que representam o mundo de maneira adequada. A resposta propõe que, se aceitarmos essa

premissa, podemos rejeitar a ideia de que a compreensão e a intencionalidade não podem ser alcançadas por meio de sistemas computacionais.

A posição central da "*English Reply*" sustenta que, de qualquer forma que possa se manifestar, a instanciação de um programa de computador,

> whether by man or by manufactured machine, does involve understanding-at least of the rule-book. Searle's initial example depends critically on Searle-in-theroom's being able to understand the language in which the rules are written, namely English (Boden, 1988, p. 260).

O personagem *Searle-in-the-room* no experimento mental depende de sua compreensão do inglês (ou do "programa", que é o alvo da crítica de Searle) para que a operação funcione adequadamente. Sob a perspectiva de Boden, é necessário que haja uma compreensão da língua inglesa, que é a língua na qual as regras estão formuladas, para que as operações possam ser executadas de maneira adequada. Sem esse nível de compreensão, as operações na sala não poderiam ser instanciadas, levando à ausência de manipulação de símbolos formais na sala.

De acordo com Boden (1988), para que o exemplo proposto por Searle possa funcionar, seriam necessárias alterações significativas no vocabulário tanto do inglês quanto do chinês (*Searle-in-the-robot*). Isso se deve ao fato de que uma língua desconhecida não seria tratada apenas como um conjunto de símbolos desprovidos de significado, mas, no máximo, "*as an aesthetic object or a set of systematically related forms*". Por outro lado, "*artificial languages can be designed and studied, by the logician or the pure mathematician, with only their structural properties in mind*" (1988, p. 260), diferentemente do que ocorre com um falante nativo em relação aos símbolos de sua língua materna, os quais seriam dificilmente ignorados em termos de significado familiar.

O entendimento da língua, no caso específico do inglês, pode ser inferido, segundo Boden, a partir do comportamento do *Searle-in-the-room*. Observa-se que, para desempenhar as atividades dentro da sala, que consistem basicamente em aplicar as regras descritas no livro, *Searle-in-the-room* precisaria dominar apenas um subconjunto bastante limitado do vocabulário do inglês[5]. Mesmo que ele soubesse apenas 1% do inglês dominado pelo Searle real, isso não importaria significativamente, pois bastaria possuir 1% desse vocabulário para interpretar o conjunto de regras e executar as operações requeridas. Essa abordagem também se aplica ao *Searle-in-the-robot*, que precisaria entender apenas um subconjunto do inglês equivalente à linguagem de programação compreendida por um computador para gerar o mesmo comportamento de entrada e saída de respostas a perguntas na janela. O mesmo raciocínio se aplicaria ao *Searle-in-the-robot*, que necessitaria compreender um subconjunto do inglês equivalente à linguagem de programação compreendida por um robô visual-motor totalmente computadorizado.

É importante destacar que os lógicos e matemáticos mencionados já possuem o conhecimento das estruturas linguísticas necessárias para elaborar uma "linguagem artificial", o que mostra certa inconsistência do argumento do Boden contrário à Searle. Entretanto, Boden negligencia esse ponto,

[5] Boden apresenta um exemplo para mostrar o motivo da exigência de um restrito vocabulário por Searle-in-the-room: "Unlike Searle, Searle-in-the-room does not require words like 'catalyse', 'beer-can', chlorophyll', and 'restaurant'. But he may need 'find', 'compare', 'two', 'triangular', and 'window' (although his understanding of these words could be much less full than Searle's). He must understand conditional sentences, if any rule states that if he sees a squoggle he should give out a squiggle. Very likely, he must understand some way of expressing negation, temporal ordering, and (especially if he is to learn to do his job faster) generalization. If the rules he uses include some which parse the Chinese sentences, then he will need words for grammatical categories too. (He will not need explicit rules for parsing English sentences, such as the parsing procedures employed in Al programs for language-processing, because he already understands English.)" (BODEN, 1988, p.261)

focando-se na complexidade da aprendizagem de uma linguagem natural. Segundo Boden, aprender uma língua envolve estabelecer conexões causais relevantes não apenas entre as palavras e o mundo, mas também entre as palavras e os diversos processos não introspectivos envolvidos em sua interpretação.

A tentativa de Boden, por meio da "*English Reply*", de estabelecer a presença de intencionalidade em *Searle-in-the-room* ou *Searle-in-the-robot* através da suposta compreensão do inglês (ou de qualquer outra língua) parece apresentar inconsistências. Isso ocorre devido à falta de uma exigência explicitamente identificada no âmbito do argumento indicando que o protagonista do experimento deva realizar algo além de simplesmente operar de acordo com as regras (o programa). No cenário mental delineado por Searle, a adoção da língua inglesa como meio de comunicação para o personagem *Searle-in-the-room* (ou *Searle-in-the-robot*) se dá em razão da analogia de Searle, na qual um ser humano é empregado como ponto de referência.

É plausível que Boden tenha exagerado nas características humanas atribuídas ao experimento mental de Searle ao afirmar a presença de compreensão e intencionalidade no personagem *Searle-in-the-room*. Todavia, essa antropomorfização imputada ao cenário mental delineado por Searle não autoriza a inferência de uma compreensão dos computadores, pois estes são dispositivos de processamento de informações que convertem uma entrada em uma saída. Os computadores operam com base na manipulação de dígitos binários, cumprindo sua funcionalidade a partir de dois estados fundamentais: "ligado" e "desligado". A concepção de uma "linguagem de programação" formal sustenta a formulação de uma "linguagem de computadores" para fins de identificação, mas essa abordagem não deixa de ser uma extrapolação, tal como sugerida no argumento da "Sala Chinesa". Na verdade, o funcionamento interno de um computador é regido por um conjunto de regras específicas; no entanto, isso não equivale a afirmar que o computador compreenda efetivamente tais regras.

Cabe enfatizar que a função primordial dos computadores é aderir às regras estipuladas, o que não implica, e tampouco requer para sua eficácia funcional, uma verdadeira compreensão das regras que estão sendo seguidas, ao contrário do que ocorreria no caso de *Searle-in-the-room*. O cerne da problemática do argumento de Searle reside no fato de ele ter personificado a si mesmo no contexto do experimento mental, numa tentativa de ilustrar que seres humanos também podem desempenhar funções de máquinas computacionais que obedecem a regras para executar tarefas específicas. Contudo, um *Searle-in-the-room* atuando no papel de um computador estaria executando a tarefa sem compreender por que está realizando tal ação, na medida em que ele sequer teria consciência de que está realizando uma ação. A questão crucial reside no fato de que, devido à natureza humana do *Searle-in-the-room*, seria necessário que ele lesse, interpretasse e executasse as regras, o que perpetua uma equivocada antropomorfização do experimento mental proposto por Searle.

Nesse contexto, a tentativa de escapar da "Sala Chinesa", conforme defendido por Boden, apenas poderia se concretizar mediante uma excessiva consideração do comportamento de *Searle-in-the-room*, tendo em vista a sua caracterização como um ser humano. Ao que tudo indica, é a tendência inerente à antropomorfização que instiga a problemática no cerne do argumento.

Large language models e o *Searle-in-the-room* sofisticado

A emergência acentuada de grandes modelos de linguagem (LLMs), notadamente o Generative Pre-trained Transformer 4, o GPT-4, desenvolvido pela OpenAI, tem engendrado a suposição de que uma entidade, seja ela de natureza humana ou mecânica, dotada de proficiência linguística, deve igualmente ser

apta ao exercício do pensamento e, portanto, compreensão. Determinados exemplares destes modelos são capazes de gerar texto que desafia a discernibilidade em relação à produção humana, e inclusive superar indivíduos humanos em determinadas tarefas concernentes à compreensão textual. Por conta disso, têm surgido assertivas, tanto no âmbito da mídia popular[6] como na esfera acadêmica[7], insinuando que os LLMs não apenas representam uma significativa inovação no processamento linguístico, mas também, de maneira mais abrangente, indicam um progresso rumo à materialização da inteligência artificial geral (ou a IA forte), delineando assim um passo em direção a uma "máquina pensante".

O GPT, que está preste a ter sua versão 5.0, constitui um modelo de aprendizado profundo que adota a arquitetura Transformer[8] para a geração de texto em linguagem natural. Este modelo é pré-treinado em uma vasta base de dados textuais, e pode ser ajustado para a realização de tarefas específicas, como tradução entre idiomas, preenchimento de lacunas textuais e análise de contextos. A arquitetura subjacente ao GPT, à semelhança de outros modelos baseados em transformadores, apresenta distintas características intrínsecas que contribuem para seu êxito. Seu funcionamento é realizado em três etapas. Abordaremos o GPT-3, pois foi dele que surgiu o mais popular LLM, o ChatGPT.

Primeira etapa. O modelo é composto por diversas camadas, cada uma das quais capaz de acessar grandes quantidades de informações provenientes das camadas anteriores. Esse arranjo propicia ao modelo a aprendizagem de propriedades tanto de ordem mais elementar quanto mais elevada, possibilitando a apreensão de nuances da entrada em diferentes níveis. De acordo com Imamguluyev (2023), a etapa inicial chamada de "pré-treinamento", durante a qual é exposto a um conjunto extenso de dados textuais, inclui obras literárias, conteúdo de *websites* e artigos diversos. Durante este período, o modelo é treinado para antecipar a próxima palavra em uma sentença, levando em conta o contexto das palavras anteriores. Este processo de pré-treinamento utiliza uma abordagem autossupervisionada, na qual o modelo é treinado com base em um vasto conjunto de dados textuais não rotulados, eliminando a necessidade de rótulos explícitos.

[6] Dale (2021) compilou uma série de matérias que demonstra o impacto do lançamento do modelo de linguagem GPT-3 da OpenAI, em meados de 2020 em diante, com sua capacidade de gerar textos em linguagem natural, difíceis de se distinguir do conteúdo de autoria humana. A partir de suas manchetes, a cobertura da grande mídia apresenta uma tendência de exageros, seja por admiração ou por ansiedade em relação às capacidades do GPT-3. Como exemplo, um dos principais jornais do mundo, o New York Times, em 29th July de 2020, apresenta a seguinte matéria: "How do you know a human wrote this? Machines are gaining the ability to write, and they are getting terrifyingly good at it".

[7] David Chalmers (2022), em *Could a Large Language Model be Conscious?,* discute a hipótese de uma consciência a partir das LLMs, mas ele não chega a uma conclusão definitiva. Chalmers apresenta argumentos a favor e contra a ideia de que um LLM pode ser consciente e discute as implicações éticas e filosóficas dessa possibilidade. No entanto, ele conclui que, dadas as suposições atuais sobre a consciência, é razoável ter uma baixa credibilidade de que os LLMs paradigmáticos atuais, como os sistemas GPT, sejam conscientes. Ele sugere que, no futuro, com o desenvolvimento de sistemas mais avançados com recursos como sentidos, modelos de mundo e modelos de si, pode ser possível que um LLM seja consciente. No entanto, ele enfatiza que essa é uma questão complexa e controversa que requer mais pesquisa e discussão. Chalmers também menciona que a consciência tem muitas dimensões diferentes, como experiência sensorial, experiência afetiva, experiência cognitiva, experiência agente e autoconsciência. Chalmers também menciona o famoso artigo de Thomas Nagel "What is it like to be a bat?" para ilustrar a ideia de que a consciência é uma experiência subjetiva que é difícil de entender completamente. Por outro lado, importantes nomes da filosofia, como Noam Chomsky (2023), são extremamente críticos em relação ao desenvolvimento de LLMs como o ChatGPT. Em seu recente artigo ao New York Times, Chomsky destaca diferenças profundas entre essas ferramentas de IA e as capacidades humanas. Para ele, a mente humana se dissocia notavelmente de entidades como o ChatGPT e suas congêneres, que encarnam dispositivos estatísticos poderosos em sua função de corresponder a padrões, caracterizados por uma voracidade em consumir enormes volumes de dados na ordem de centenas de terabytes, a fim de extrapolar a resposta mais provável em interações conversacionais ou mesmo em abordagens de questionamentos de cunho científico. Ao contrário, como menciona Chomsky em seu artigo, a mente humana se consubstancia enquanto um sistema notoriamente eficiente e mesmo elegante, caracterizado por operações engendradas a partir de quantidades de informação notavelmente reduzidas. A mente humana não se restringe à inferência de correlações brutais entre pontos de dados; em contraposição, acentua-se a sua inclinação para forjar explicações que conferem sentido e coesão aos fenômenos objeto de sua apreensão. Chomsky conclui então que a discrepância entre a mente humana e os sistemas supracitados se sobressai na sua proficiência em extrair significados intrincados de contextos subjacentes, capacidade de discernir nuances nas interações comunicativas e discernimento que vai além da mera apreensão do conteúdo explícito.

[8] A arquitetura fundamentada no paradigma do transformador foi primordialmente introduzida no influente trabalho de VASWANI (2017). Esta concepção arquitetônica constitui um formato de rede neural especialmente apropriado para o tratamento de texto em linguagem natural. O modelo engloba uma composição estratificada de mecanismos de autoatenção e camadas de redes neurais de propagação direta. As camadas de autoatenção facultam ao modelo a capacidade de direcionar a sua atenção a distintas porções do texto inserido, ao passo que as camadas de propagação direta efetivam transformações não-lineares sobre as saídas originadas nas camadas de autoatenção.

De acordo com Rothman (2022), a primeira etapa do processo de treinamento consiste em utilizar uma vasta quantidade de texto da *web*, aproximadamente 45 *terabytes* de dados, o que equivale a cerca de 500 bilhões de palavras. Esse texto é fragmentado em *tokens*, que podem ser palavras ou componentes de palavras. O que distingue os *Large Language Models* é a abordagem utilizada para processar esses *tokens*, que se baseia nas informações de co-ocorrência de caracteres, ao invés de empregar uma análise morfológica tradicional. Para ilustrar esse método, consideremos a seguinte frase como exemplo: "ChatGPT é bom para revisão textual" é dividida em palavra ou componentes: "Chat", "G", "P", "T", "é", "bom", "para", "revisão", "textual". Palavras comuns e curtas, como "bom" e "para", são mantidas intactas, enquanto palavras que não estão presentes no *corpus* original, como "ChatGPT", são subdivididas em partes menores. Além disso, palavras como "textual" são segmentadas em unidades como "text-" e "-al", uma vez que o modelo reconhece que "text-" poderia ser combinado de diferentes formas, como "texto", e "-al", como "banal".

No contexto do relacionamento entre os *tokens*, é importante notar que o modelo GPT-3 tem como objetivo simples prever a próxima palavra com base em um número fixo de palavras anteriores, normalmente algumas centenas. Para isso, ele calcula a probabilidade de o *token* "textual" ocorrer após o token "revisão", por exemplo, considerando as informações contextuais disponíveis. Para avaliar a precisão da previsão, o modelo compara a parte da palavra prevista com a verdade básica, ou seja, a parte da palavra que realmente ocorreu naquela frase de treinamento. Esse processo é repetido iterativamente durante o treinamento, e o *feedback* resultante é utilizado para atualizar os muitos parâmetros do modelo, que totalizam mais de 100 bilhões.

Segunda etapa. Cada camada abarca um mecanismo de *gating*, designado como "atenção" (Vaswani et al., 2017), que permite que cada componente de palavra se direcione de maneira seletiva a qualquer um dos nodos correspondentes em camadas anteriores. O GPT-3, após o estágio de pré-treinamento, realiza um refinamento para desempenhar tarefas específicas de processamento de linguagem natural (PLN). Este processo de ajuste fino envolve treinar o modelo em um conjunto de dados rotulado, que é mais limitado e direcionado à tarefa em questão. Por exemplo, se uma parte do modelo é encarregada de gerar um pronome, ele pode discernir seletivamente os possíveis substantivos antecedentes a fim de gerar o pronome correto. Outro exemplo se dá com a tradução entre idiomas, quando o GPT-3 pode ser sintonizado com um conjunto de dados composto por frases correspondentes em diferentes línguas. Tal procedimento permite que o GPT-3 se adapte às particularidades inerentes à tarefa, aprimorando o seu desempenho.

Terceira etapa. Depois de ter passado pelo pré-treinamento e pelo ajuste fino, o GPT-3 é capaz de gerar texto em linguagem natural, prevendo a próxima palavra em uma sequência, dadas as palavras prévias. O modelo realiza a geração de texto por meio de amostragem a partir de uma distribuição de probabilidade associada ao vocabulário, fundamentando-se tanto no texto de entrada quanto nos parâmetros que foram aprendidos. Este procedimento confere ao GPT-3 a capacidade de gerar texto que é coeso e fluente, mesmo em sentenças longas e complexas[9].

Dessa forma, há a "compreensão contextual" (Imamguluyev, 2023). Uma das principais características distintivas do GPT-3 reside na sua habilidade de "compreender" (ou simular compreensão, conforme será tratado a seguir) e gerar texto contextualmente relevante. O modelo emprega uma técnica denominada "atenção", que viabiliza a focalização em diferentes partes do texto de entrada, permitindo a

[9] Mesmo assim, o GPT-3 pode produzir as denominadas "alucinações", conforme Lee (2023) e Alkaissi & Mcfarlane (2023).

geração de texto coerente com o contexto fornecido. Essa característica possibilita que o GPT-3 produza texto altamente pertinente em relação ao texto de entrada, sendo eficaz em uma ampla gama de tarefas de natural *language processing*. Na arquitetura do *transformer* (Vaswani *et al*., 2017), as palavras na entrada são processadas simultaneamente (não linearmente, palavra a palavra, como em modelos antecedentes), culminando na geração de uma previsão integral de uma só vez. Tal abordagem torna o processo de treinamento mais passível de paralelização e, consequentemente, mais eficiente, viabilizando seu treinamento em face da abrangente quantidade de dados requeridos.

No decorrer de seu treinamento para previsão vocabular, os modelos baseados em transformadores assimilam consideráveis informações acerca da estrutura da linguagem, incluindo atributos linguísticos que, até recentemente, eram percebidos como fora do âmbito das abordagens estatísticas. Não apenas obtiveram sucesso em avaliações de compreensão linguística gerais desenvolvidas pela comunidade de Processamento de Linguagem Natural (PLN), mas também, o que é crítico para nossos propósitos, em avaliações de competência linguística.

Considerando o exposto em *What Your Computer Can't Know*, é possível sustentar que Searle (2014), apesar de todo avanço tecnológico dos LLMs, ainda consideraria a IA generativa incapaz de compreensão. Neste texto, Searle, que dirige críticas diretas a Floridi (2014) e Bostrom (2014), procura sustentar a confusão filosófica ainda existente sobre as relações entre consciência, computação, informação, cognição, compreensão, entre outros fenômenos. No prosseguimento de sua exposição, Searle expande seu raciocínio original, baseado no clássico experimento mental da "Sala Chinesa". Nesse contexto, ele introduz a perspectiva de que a computação, conforme conceituada por Alan Turing e efetivamente implementada em máquinas tangíveis, está intrinsecamente ligada à posição do observador. As abruptas transições de estados físicos ocorridas em uma máquina eletrônica, em sua essência, são meros cálculos que dependem da existência de uma consciência real ou, pelo menos, uma consciência potencial capaz de interpretar esses processos computacionais.

Da mesma forma, uma inteligência, para ser real, deve ser intrínseca, isto é, independente do observador[10], o que não ocorre na realidade das máquinas computacionais e, consequentemente, da denominada IA. Para Searle (2014), a inteligência da IA "*is entirely observer relative. And what goes for intelligence goes for thinking, remembering, deciding, desiring, reasoning, motivation, learning, and information processing* [...]". Por isso, não é possível afirmar que máquinas computadoras possuam uma inteligência autêntica (original), pois "*while some of them do their jobs superbly, do not for a moment think that there is any psychological reality to them*".

Nesta seção do artigo, John Searle direciona seus esforços para refutar a tese proposta por Bostrom (2014), que sugere a possibilidade de uma superinteligência com intenções potencialmente prejudiciais à existência humana. Searle fundamenta seu argumento na premissa de que as máquinas computacionais não detêm uma inteligência ou consciência genuinamente psicológica. Assim, carece de sentido atribuir ao computador ou à inteligência artificial qualquer forma de motivação independente do observador. Ademais, Searle estende seu raciocínio para desafiar a perspectiva metafísica apresentada por Floridi (2014), que postula que a informação constitui a estrutura fundamental do universo. Esta visão implica que a realidade,

[10] Searle (2014) esclarece esta distinção da seguinte forma: "A related distinction is between those features of reality that exist regardless of what we think and those whose very existence depends on our attitudes. The first class I call observer independent or original, intrinsic, or absolute. This class includes mountains, molecules, and tectonic plates. They have an existence that is wholly independent of anybody's attitude, whereas money, property, government, and marriage exist only insofar as people have certain attitudes toward them. Their existence I call observer dependent or observer relative".

quando adequadamente interpretada, é composta inteiramente por informação. Searle (2014) observa que existe um equívoco no emprego do termo "informação". Portanto, torna-se imperativo estabelecer uma distinção entre o conceito de informação independente do observador, que possui uma realidade psicológica genuína, e o sentido de informação relativa ao observador, que não possui qualquer fundamentação psicológica concreta. De acordo com Searle (2014), "*conscious humans and animals have intrinsic information but there is no intrinsic information in maps, computers, books, or DNA […]. The sense in which they contain information is all relative to our conscious minds*".

Nesta crítica elaborada por Searle aos conceitos de inteligência computacional e informação empregados respectivamente por Bostrom e Floridi, é plausível empreender uma analogia entre o funcionamento surpreendente dos LLMs, exemplificado pelo GPT-3, e o que poderia ser chamado de uma versão sofisticada do *Searle-in-the-room*. Searle (2014) argumenta que persiste um resquício de behaviorismo nas ciências cognitivas, em que os profissionais muitas vezes incorrem na concepção de que, se for possível construir uma máquina capaz de exibir comportamentos inteligentes, essa máquina automaticamente se torna verdadeiramente inteligente. Adicionalmente, Searle alerta sobre a persistência de vestígios de dualismo na abordagem dos pesquisadores quando estes hesitam em abordar a consciência, o pensamento e o processamento de informações como fenômenos psicologicamente reais, e, ao invés disso, os tratam como entidades separadas e distintas dos processos biológicos comuns, tais como a fotossíntese ou a digestão.

A utilidade inegável da inteligência artificial e suas variantes mais avançadas, como os *Large Language Models*[11], muitas vezes leva a uma confusão entre esses avanços e a conquista de uma IA geral. Na perspectiva de Searle, esse constante equívoco decorre da presença de uma combinação estranha de princípios behavioristas e dualistas. Consequentemente, a "inteligência" que alguém pode erroneamente atribuir ao computador de Searle (1980; 2014) e ao programa que ele executa para identificar padrões correspondentes às palavras chinesas se assemelha de forma notável à "inteligência" que alguém pode equivocadamente perceber nas respostas de um *chatbot* em interações com comandos em linguagem natural. Em ambos os cenários, o que muitas vezes é interpretado como inteligência é, na realidade, uma forma de simulação. Isso ocorre porque, independentemente das notáveis diferenças na sofisticação do desenvolvimento dos programas, decorrentes do avanço do tempo, ambos se resumem, em sua essência, a buscas de força bruta de grandes quantidades de informações, a fim de prever com precisão quais devem ser as palavras subsequentes. Embora frequentemente obtenham êxito, ocasionalmente cometem equívocos, cujas consequências podem ser tanto positivas quanto negativas, ou até mesmo insignificantes[12].

[11] "That is the situation we are currently in with Artificial Intelligence. Computer engineering is useful for flying airplanes, diagnosing diseases, and writing articles like this one. But the results are for the most part irrelevant to understanding human thinking, reasoning, processing information, deciding, perceiving, etc., because the results are all observer relative and not the real thing" (SEARLE, 2014).

[12] Para exemplos de "alucinações", ver Alkaissi & Mcfarlane (2023), que tratam de equívocos no uso do ChatGPT na área médica. Já o trabalho de Lee (2023) apresenta uma análise abrangente do fenômeno da alucinação em generative pretrained transformer (GPT) models com base em uma abordagem matemática. O estudo em questão define e quantifica rigorosamente tanto a alucinação quanto a criatividade, fazendo uso de conceitos fundamentais da teoria da probabilidade e da teoria da informação. Por meio da introdução de uma família paramétrica de modelos GPT, busca-se caracterizar o ponto de equilíbrio que governa a relação entre alucinação e criatividade, identificando um equilíbrio capaz de otimizar o desempenho do modelo em diversas tarefas. Lee procura oferecer uma estrutura matemática inovadora para a compreensão das origens e implicações da alucinação em modelos GPT. "Despite the impressive performance of GPT models, they are known to exhibit a phenomenon called hallucination, wherein they generate outputs that are contextually implausible or inconsistent with the real world. The hallucination phenomenon has been attributed to the model's inherent limitations, particularly its inability to discern when there is no well-defined correct answer for a given input. Consequently, GPT models can generate low-likelihood outputs that deviate from the expected output based on the input context and the true data distribution" (LEE, 2023, p.2).

Considerações Finais

Pasquinelli e Joler (2020) oferecem uma importante reflexão para os propósitos finais deste artigo sobre os problemas relacionados à mistificação da inteligência artificial. Ao proporem a ideia da IA como um "nooscópio", os autores buscam desvincular a IA de sua carga ideológica de "máquina inteligente" e situá-la ao *status* de instrumento de conhecimento, fazendo uma analogia com as próteses ópticas, como o telescópio desenvolvido por Galileu. Eles sugerem, assim, que é mais sensato considerar a aprendizagem automática como um instrumento de expansão do conhecimento, auxiliando na identificação de características, padrões e correlações em vastos conjuntos de dados que estão além do alcance humano.

A revisitação ao argumento da Sala Chinesa de Searle, proposta neste artigo, demonstra a viabilidade de preservar sua crítica central — a inexistência de uma inteligência artificial no sentido forte. Para isso, é necessário retirar o "*Searle-in-the-room*" e substituí-lo pela metáfora do nooscópio, como argumentam Pasquinelli e Joler no Manifesto Nooscópio. Essa metáfora transforma a IA de um sistema alegadamente inteligente em um instrumento de ampliação do conhecimento, focado em padrões e correlações dentro de dados massivos. O nooscópio exemplifica como a IA deve ser compreendida: não como um substituto do pensamento humano, mas como uma ferramenta epistêmica que expõe os limites e os vieses embutidos em sua estrutura algorítmica. Essa abordagem resgata a crítica de Searle ao destacar que a IA, mesmo em modelos n-dimensionais avançados, permanece distante da semântica e da intencionalidade humanas.

Assim, ao retirar Searle da sala e adotar o nooscópio, o argumento revisitado reforça que a IA opera como um regime de extração de conhecimento, desmistificando sua autonomia. Essa transição, além de manter a validade da crítica original, também amplia o debate sobre os impactos sociotécnicos da IA, enfatizando a importância de compreender seus limites éticos e epistemológicos.

Declaração de disponibilidade de dados

O presente artigo tem como foco principal contribuições de natureza teórica ou metodológica, sem a utilização de conjuntos de dados empíricos. Dessa forma, conforme as diretrizes editoriais da revista, o artigo está isento de depósito no SciELO Data.

Referências

ALKAISSI, H.; MCFARLANE, S.I. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*, v. 15, n. 2, 2023. DOI: https://doi.org/10.7759/cureus.35179.

BODEN, Margaret A. Escaping from the Chinese Room. *In*: BODEN, Margaret A. *Computer Models On Mind*: Computational Approaches In Theoretical Psychology. Cambridge: Cambridge University Press, 1988. Disponível em: https://philpapers.org/rec/BODCMO. Acesso em: 29 abr. 2025.

BOSTROM, Nick (ed.). *Superintelligence:* paths, dangers, strategies. Oxford University Press, 2014.

CHALMERS, David. Could a Large Language Model be Conscious? *In*: NEURIPS, 36., 2022, virtual. *Anais* [...]. San Diego: NeurIPS, 2022. Disponível em: https://nips.cc/virtual/2022/invitedtalk/55867. Acesso em: 29 abr. 2025.

CHOMSKY, Noam. The false promise of ChatGPT. *New York Times*, New York. 08 mar. 2023. Disponível em: https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html. Acesso em: 29 abr. 2025.

DALE, Robert. Gpt-3: What's it good for? *Natural Language Engineering*, Cambridge University Press, v. 27, n. 1, p. 113–118, 2021. DOI: https://doi.org/10.1017/S1351324920000601. Disponível em: https://www.cambridge.org/core/journals/natural-language-engineering/article/gpt3-whats-it-good-for/0E05CFE68A7AC8BF794C8ECBE28AA990. Acesso em: 29 abr. 2025.

FLORIDI, Luciano. *The fourth revolution*: how the infosphere is reshaping human reality. Oxford University Press UK, 2014.

HARNAD, S. Minds, Machines and Searle. *Journal of Theoretical and Experimental Artificial Intelligence*, v. 1, p. 5-25, 1989.

IMAMGULUYEV, Rahib. The Rise of GPT-3: Implications for Natural Language Processing and Beyond. *International Journal of Research Publication and Reviews*, v. 4, n. 3, pp 4893-4903, March 2023. DOI: https://doi.org/10.55248/gengpi.2023.4.33987. Disponível em: https://ijrpr.com/uploads/V4ISSUE3/IJRPR10923.pdf. Acesso em: 29 abr. 2025.

LEE, M. A Mathematical Investigation of Hallucination and Creativity in GPT Models. *Mathematics*, v. 11, n. 10, 2023. DOI: https://doi.org/10.3390/math11102320.

MORAVEC, Hans. *Mind Children*: The Future of Robot and Human Intelligence. Cambridge, MA: Harvard University Press, 1988.

PASQUINELLI, Matteo; JOLER, Vladan. The Nooscope Manifested: Artificial Intelligence as Instrument of Knowledge Extractivism. *AI and Society*, v. 36, 21 November 2020. DOI: https://doi.org/10.1007/s00146-020-01097-6.

PRESTON, J.; BISHOP, M. (eds.). *Views into the Chinese Room*: New Essays on Searle and Artificial Intelligence. Oxford: Oxford University Press, 2002.

ROTHMAN, Denis. *Transformers for Natural Language Processing*: Build, train, and fine-tune deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, and GPT-3. Birmingham/Mumbai: Packt Publishing Ltd, 2022.

SEARLE, J. R. Chinese Room Argument. *In*: WILSON, Robert A.; KEIL, Frank. *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press, 1999

SEARLE, J. R. Minds, Brains and Programs. *Behavioral and Brain Sciences*, v. 3, 1980.

SEARLE, J. R. *Minds, Brains and Science*. London: BBC Publications and Cambridge, Mass Harvard University Press, 1984.

SEARLE, J. R. *The Mystery of Consciousness.* New York: A New York Review Book, 1997.

SEARLE, John. What Your Computer Can't Know. *New York Review of Books*, 2014.

SEJNOWSKI, Terrence J. Large language models and the reverse turing test. *Neural computation*, v. 35, n. 3, p. 309-342, 2023.

TURING, A. M. Computing Machinery and Intelligence. *Mind*, v. LIX, n. 236, Pages 433–460, October, 1950. DOI : https://doi.org/10.1093/mind/LIX.236.433

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N.; KAISER, Lukasz; POLOSUKHIN, Illia. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017.

Editor Responsável: Léo Peruzzo Júnior.

RECEBIDO: 11/11/2024 *RECEIVED: 11/11/2024*
APROVADO: 29/03/2025 *APPROVED: 03/29/2025*
PUBLICADO: 16/10/2025 *PUBLISHED: 10/16/2025*