LINGUAGEM, INTENCIONALIDADE E O PROBLEMA DA INTELIGÊNCIA ARTIFICIAL

LANGUAGE, INTENTIONALITY, AND THE PROBLEM OF ARTIFICIAL INTELLIGENCE

Libni E. Teles[1]

Resumo: Esse trabalho pretende entender quais mecanismos da linguagem revelam a intencionalidade, analisando o trabalho de John Searle. Após a exposição de parte da teoria de Searle sobre a intencionalidade, será feito um comentário a respeito dos recentes avanços da Inteligência Artificial, em especial àquelas ferramentas que lidam com a linguagem natural. Por fim, explicaremos por que, por mais sofisticados que sejam, essas ferramentas estão longe de apresentar algum tipo de intencionalidade aos moldes da teoria de Searle.

Palavras-Chave: Intencionalidade. John Searle. Inteligência Artificial. Atos de Fala.

Abstract: This work intends to understand which language mechanisms reveal intentionality, analyzing the work of John Searle. After exposing part of Searle's theory on intentionality, a comment will be made on recent advances in Artificial Intelligence, on tools that deal with natural language. Finally, we will explain why, however sophisticated they are, these tools are far from presenting some kind of intentionality along the lines of Searle's theory.

Keywords: Intentionality. John Searle. Artificial Intelligence. Speech Acts.

Introdução: Intencionalidade - um estudo a partir dos Atos de Fala

Searle é provavelmente um dos mais importantes filósofos contemporâneos. Detentor de uma popularidade inquestionável, seu trabalho é indispensável, especialmente aquele que concerne suas críticas à Inteligência Artificial. Seu popular experimento *The Chinese Room* argumenta, em resumo, que máquinas são ferramentas meramente formais e não são capazes de instanciar uma mente. Trata-se da crítica mais contundente à Inteligência Artificial Geral. Sua bibliografia, no entanto, está muito mais comprometida com a filosofia da linguagem do que com o computacionalismo. Em seu primeiro trabalho, *Speech Acts*, de 1969, Searle se compromete a construir uma taxonomia das teorias dos Atos de Fala, criada originalmente por seu orientador, John L. Austin. Searle se dedicou a teoria dos atos de fala de tal forma que, a partir dela, construiu um arcabouço teórico que explica, a partir da análise da linguagem, desde as intenções

[1] Licenciado em Filosofia pela Universidade Federal de Alagoas (UFAL). Mestrando em Filosofia, com ênfase em Linguagem e Cognição, pelo Programa de Pós-Graduação em Filosofia (PPGFil). Lattes: https://lattes.cnpq.br/9191541014000883. E-mail: libni_teles25@live.com. ORCID: https://orcid.org/0009-0005-7620-3474.

escondidas por trás dos enunciados, até uma ontologia da realidade social que busca entender como a sociedade em que vivemos, com suas instituições, compromissos, promessas, casamentos, leis e governos estão diretamente ligados à satisfação desses enunciados. Para entender como Searle liga os enunciados do nosso dia a dia, enquanto seres sociais, com a intencionalidade é preciso primeiro entender onde estão as raízes do seu pensamento.

Investiguemos resumidamente o trabalho do seu mentor, John Austin. Na primeira metade do século XX, os positivistas criaram uma condição em que o significado das palavras só poderia ser analisado por meio de condições de verificação. Isso ocorreu porque os positivistas se interessavam apenas pela função especulativa da linguagem. Sendo assim, o que é expresso por meio de uma frase como "Batizo este navio *Queen Mary*" ultrapassa o conteúdo empírico do que está sendo enunciado, e aos filósofos da linguagem caberia apenas analisar o sentido dos enunciados, e não aquilo que acrescentamos ao sentido, com vistas a nossa comunicação. Mais tarde, o conceito de "significado como uso" foi introduzido pelos trabalhos de Wittgenstein e, subsequentemente, Austin. A concepção de Austin em alguns aspectos se assemelha ao segundo Wittgenstein, pelo menos no que diz respeito à atenção que ele concede ao conceito de uso nas suas investigações sobre a linguagem. Porém, não há, até onde se sabe, qualquer paralelo direto entre o trabalho de Austin e o de Wittgenstein. Apesar de Austin ser contemporâneo ao Wittgenstein, sua teoria se desenvolve de maneira independente. O rompimento de Austin com a filosofia da linguagem inspirada na lógica fregeana é primeiramente representado por seu artigo de 1946, chamado *Other Minds*. Como o título sugere, nesse trabalho Austin investiga sobre como sabemos a respeito da existência de pensamento, sensações e desejos de outras pessoas, partindo de uma análise de enunciados como: "Estou certo de que vai chover". No fim das contas, se trata de uma análise sobre a autoridade que alguém concede ao dizer certos enunciados como "Eu sei que..." ou "Estou certo de que...". Esse artigo já serve de pista para o que Austin viria desenvolver a seguir. Em seu livro, *How To Do Things With Words*, editado a partir de uma série de aulas ministradas em Havard em 1955, Austin tenta entender a construção daquilo que chamou de Atos Performativos, uma teoria que seria extremamente influente dali em diante.

Lembremos da frase "Batizo este navio *Queen Mary*". Segundo o positivismo lógico, esse enunciado, por carecer de verdade, seria ausente de sentido. Austin, por sua, vez vai classificar este e outros enunciados desse mesmo tipo como sendo enunciados

performativos (*performative utterances*). Tais enunciados são definidos como sendo bem-sucedidos ou malsucedidos. Se, por exemplo, quebro a garrafa de champanhe no casco para batizar o navio (sendo eu também aquele que tem o poder para batizar o navio), então "Batizo este navio *Queen Mary*" é um enunciado performativo bem-sucedido, conforme Austin descreve. A isto Austin vai chamar de "condição de felicidade". Como o enunciado não pode ter um valor verdade, a ele é atribuído uma condição de felicidade, que pode ser satisfeita ou não. Tendo isso em mente, Austin define que o uso da linguagem é um ato, e, portanto, uma teoria da linguagem deve ser inserida numa teoria da ação.

Dessa ideia surgirá a sua teoria dos "atos de fala", em que a teoria da ação linguística deve compreender a fonética, a sintaxe e a semântica. Toda ação linguística ocorre em uma linguagem. Nesse sentido, o enunciado "Batizo este navio *Queen Mary*" possuí três atos, o ato locutório, o ato ilocutório, e o ato perlocutório. O ato locutório é simplesmente o ato de dizer algo, como por exemplo "Batizo este navio *Queen Mary*", onde se analisa apenas o aspecto fonético e a sintaxe do enunciado, suas regras. O ato ilocutório por sua vez traz muito mais do que a sintaxe, se trata da força ilocutória, onde ao dizer "Batizo este navio *Queen Mary*", o digo como se fosse uma ordem. Sempre que se diz algo, se diz com uma força particular, como se fosse uma asserção, uma pergunta, uma promessa, uma súplica, ordem etc. Por fim, o ato perlocutório é o que acontece a partir da enunciação original, neste caso, se a pessoa que enuncia "Batizo este navio *Queen Mary*" tem poder para dar nome a um navio, então a frase "Batizo este navio *Queen Mary*" deve batizar o navio com o nome *Queen Mary*.

Em *How To do Things With Words*, no entanto, a coisa não fica tão bem esquematizada como vimos acima. Grande parte da teoria-desenvolvida por Austin, especialmente no que concerne aos Atos Perlocutórios, não fica clara. O trabalho de sistematizar e desenvolver uma taxonomia dessa teoria posteriormente ficou a cargo de outros filósofos, dos quais podemos destacar John Searle. Além de construir uma taxionomia para os atos de fala e dar substância a teoria, a partir do seu livro *Speech Acts* (1969), Searle busca entender também como as múltiplas dimensões dos atos de fala, enunciado, significado e ação, podem ser unificadas. Isto é, o que há no ato de falar algo, ou escrever algo, ações meramente físicas de pronunciar sons ou rabiscar símbolos, que é capaz de possuir algum significado e é capaz de fazer com que executemos uma ação? O que explica o caso de ao ouvir "Cuidado, há tubarões no mar!" estes sons possam ser entendidos como um alerta para o perigo e sejam capazes de dissuadir a minha vontade

de nadar? Searle primeiro distingue o que ele chama de regras regulativas e constativas (*regulative rules* e *constative rules*). Enquanto a primeira apenas "regula" um comportamento que já existe, como, por exemplo, descrever o ato de comer, a segunda constrói novas formas de comportamento. Essas regras constativas não acontecem sozinhas, de forma independente. Se peço desculpas ao espirrar em público, é porque já existe um conjunto de outras regras que convencionam que ao espirrar em público deve-se pedir desculpas. Os atos de fala estão sempre dentro desse conjunto de regras constativas.

Quando alguém diz "Desculpe, estou resfriado.", esse alguém tem a intenção de informar algo. Nesta intenção está a chave para entender a intencionalidade por trás de um ato de fala. Se alguém enuncia uma sentença, que chamaremos de S, (como "Desculpe, estou resfriado") três condições necessitam ser satisfeitas, a ver:

1. Esse alguém tem intenção, que chamaremos de I, que seu enunciado produza no ouvinte ou nos ouvintes, a consciência do estado de relações que levou a S;
2. O falante intenciona produzir essa consciência por meio do reconhecimento da intenção I
3. O falante intenciona que essa intenção I será reconhecida em virtude das regras que regem os elementos da sentença S (Searle, 1969, p. 49).

É fácil ver como essas condições se aplicam a nossa sentença usada como exemplo ("Desculpe, estou resfriado"). O ponto interessante é que Searle passa a fazer uma sistematização lógica do funcionamento dessas regras. O uso desse cálculo lógico, o qual Searle usará para explicar como os enunciados se relacionam com intencionalidade, pode ser visto como uma espécie de distanciamento de Austin, que evitava justamente o logicismo excessivo dos positivistas, buscando uma alternativa mais próxima da informalidade do uso da linguagem no dia a dia.

Falamos do sistema de regras constativas ao qual um simples enunciado como "desculpe, estou resfriado" pode estar envolvido. Esse sistema de regras constativas Searle vai chamar de "fato institucional" (*Institutional fact*), que ele define como um fato no qual a existência pressupõe esse sistema de regras constativas. Quando alguém performa um ato de fala, essa pessoa também cria um certo fato institucional. Esses fatos existem porque nós estamos em um ambiente e uma sociedade na qual necessitamos lidar de uma certa maneira com o estado das coisas, em determinados contextos. Nesse sentido,

existem duas coisas nesse ambiente que estão em contraste, a primeira delas é a existência de fatos que não estão sob nosso domínio, como as leis da física. Outros, como dinheiro, propriedade, casamento etc., no entanto, são produtos da nossa criação, fatos dos fatos institucionais antes mencionados.

Além da taxionomia geral dos atos de fala, Searle promoveu também uma nova estrutura detalhada dos atos de fala. Ele distingue em duas as condições de felicidade que foram originalmente definidas por Austin: as condições de performance e as condições de satisfação (*Performance conditions* e *satisfaction conditions*). A primeira, como o nome sugere, se realiza pelo simples fato de se enunciar o ato. A segunda exige alguma ação daquele que enuncia. Se digo “Amanhã vou lavar o carro”, o enunciado já foi performado, e eu me propus a lavar o carro. Para de fato satisfazer esse enunciado eu preciso lavar o carro.

Pois bem, nossa intenção é entender como, para Searle, a linguagem se relaciona com a intencionalidade. Vale esclarecer antes certos conceitos que serão usados por ele para explicar a relação que existe entre os enunciados e a realidade social, já que, como vimos, Searle está inicialmente preocupado em entender como frases e rabiscos, meramente físicos, são capazes de causar alguma ação no mundo. Ele expõe esses conceitos no segundo capítulo do livro *Speech Acts*. Nesse capítulo Searle vai fornecer uma classificação das "direções de ajuste" (*direction of fit*), entre ambos, enunciado e realidade, ou, em resumo, como os enunciados se relacionam com a realidade e vice-e-versa. O conceito de direções de ajuste não é original de Searle, sua definição lógica diz que “direções de ajuste” é um conceito usado para descrever as distinções que são oferecidas por dois conjuntos de termos opostos relacionados. Consideremos um exemplo usado pela filosofa Elizabete Anscombe para total entendimento:

> Uma mulher manda o marido ao supermercado com uma lista de coisas a se comprar; sem que ele saiba, ele também está sendo perseguido por um detetive preocupado em fazer uma lista do que o homem compra. Quando o marido e o detetive estão na fila do caixa, suas duas listas contêm exatamente os mesmos itens. O conteúdo das duas listas difere, porém, ao longo de outra dimensão. Pois o conteúdo da lista do marido orienta o que ele coloca no carrinho de compras. Nesse sentido, sua lista exibe uma direção de ajuste "mundo-se-ajusta-para-mente": é, por assim dizer, o trabalho dos itens em seu carrinho de acordo com o que está em sua lista. Ao contrário, cabe à lista do detetive conformar-se com o mundo, em particular com o que está no carrinho do marido. Como tal, a lista do detetive tem uma direção de ajuste "palavra-se-ajusta-para-mundo": o ônus recai sobre essas palavras para se adequar a como as coisas são. (Anscombe, 1963, p. 57)

Searle usa o caso das "*directions of fit*" para classificar os atos de fala da seguinte maneira: Os atos de fala com força assertiva, isto é, atos que, por exemplo, afirmam alguma coisa, têm uma direção "palavra-se-ajusta-para-o-mundo" (*Word-to-World*), atos com força diretiva tem uma direção "mundo-se-ajusta-para-a-palavra" (*World-to-Word*), assim como os atos com força comissiva. Os Atos com força expressiva, como congratulações, por exemplo, não têm direção alguma, uma vez que eles já pressupõem alguma verdade. Atos declarativos possuem direção "palavra-se-ajusta-para-o-mundo".

Em seu livro Intencionalidade (1983), Searle desenvolve sua teoria da causação intencional. Ele considera o fato que uma intenção é satisfeita apenas se a intenção por si mesma causa a satisfação do resto da sua condição de satisfação. Por exemplo, se desejo correr, não basta apenas desejar correr, eu necessito de fato o fazê-lo para que essa intenção seja satisfeita. Ou seja, a ação causal é autorreferenciável. Não significa, Searle esclarece logo nos primeiros capítulos do livro, que a intencionalidade depende da linguagem para existir, mas é factual que ela reflete as nossas atividades cognitivas mais fundamentais, e, acima de todas elas, a capacidade da mente em representar os estados das coisas. Observemos a teoria da Intencionalidade de Searle mais detidamente, conforme ele a desenvolve em seu livro de 1983.

É preciso notar que, apesar de se tratar de um conceito que carrega tradição histórica, Searle não faz questão de resgatá-la ao desenvolver sua teoria. Há pontos de semelhança entre a Intencionalidade de Searle a intencionalidade de Franz Bretano, mas essas semelhanças são iniciais, em pouco tempo a Intencionalidade de Searle, graças a sua análise da linguagem, constrói-se como um conceito particular desenvolvido pela sua filosofia. Sendo assim, podemos distinguir suas formulações iniciais, presentes ainda no primeiro capítulo do livro *Intentionality*, em que a Intencionalidade é uma propriedade dos estados e eventos mentais no qual eles são dirigidos a um objeto ou estado de relações, ou então são sobre um objeto, ou estado de relações no mundo. Em resumo, se tenho uma intenção, essa intenção deve ser uma intenção sobre algo. Nem todos os estados e eventos mentais possuem intencionalidade, existem algumas sensações que não são intencionais. Por exemplo, ansiedade pode não está relacionado com algo em específico; desejos e crenças, no entanto, necessitam sempre ser sobre algo.

Lembremos das regras numeradas que descrevemos mais acima. Se tenho um estado intencional I, então esse estado deve responder certas questões, como, por exemplo: I é sobre o quê? O fato de nem todos os estados mentais serem intencionais

demonstra que, para Searle, Intencionalidade não é o mesmo que consciência. No sistema desenvolvido por Searle é possível dizer que a consciência está em um nível mais "geral" do que a intencionalidade (Searle, 1992, p. 83).

Com o termo "intenção", Searle quer dizer que algo se pretende com uma ação, nesse sentido, ele descreve apenas uma forma de Intencionalidade. Eu posso ter um estado mental *x* que é sobre a lua, tal estado possui, naturalmente, um objeto sobre o qual ele é dirigido, mas isso não significa que esse estado mental me leve a fazer algo, ou seja, que ele tenha alguma intenção. Tal conceito aparece da mesma forma em Paul Grice (1968). Nesse sentido, Crenças e desejos são intencionais no sentido de que eles são estados mentais que possuem intencionalidade, mas eles não necessariamente intendem a alguma coisa. Essa relação entre Desejos e Crenças cria a primeira peculiaridade na teoria da Intencionalidade de Searle. Poderíamos resumir todos os estados intencionais a crenças e desejos, mas o caso é que Searle acredita que o modelo Crença/Desejo não é capaz de promover uma análise completa das intenções (Zaibert, 2003, p. 53). O que se perde é justamente o papel causal especial das intenções na produção do nosso comportamento. Um exemplo simples, seguindo a lógica modal, é o seguinte:

Tenho a intenção (de ser rico), implica que eu acredito (que eu possa ser rico) e desejo (ser rico).

Em notação lógica, esta frase fica assim:

$$\text{Intendo (Eu faço } A\text{)} \rightarrow \text{Acredito } (\Diamond \text{Eu faço } A) \wedge \text{Desejo (Eu faço } A\text{)}$$

Para que tenha sucesso, uma intenção precisa possuir uma condição de satisfação, conforme expusemos mais acima. Se Maria deseja ficar rica, ou acredita que vai ficar rica, isso não satisfaz nenhuma condição de satisfação, porque o simples ato de acreditar ou desejar não gera, *per se*, ação alguma. O caso para Maria é que sua condição de satisfação só pode ser satisfeita se ela ficar rica a partir de sua intenção em ficar rica. Por isso que Crenças e Desejos não intendem a nada.

Tendo feito esse esclarecimento, partimos para a relação que existe entre os estados intencionais e os objetos ou estados de relações ao qual eles estão dirigidos. Searle entende a Intencionalidade como modelos representacionais. É aí que há a primeira ligação explicita entre linguagem (no caso, atos de fala) e Intencionalidade. Estados

intencionais representam objetos e estados de relações no mesmo sentido em que os atos de fala representam objetos e estados de relações. É importante destacar o que Searle quer dizer com o uso do termo "representação". Assim como faz com a intencionalidade, ele procura uma definição própria. Primeiro Searle se afasta da definição tradicional, bem como da definição usada na cognição e inteligência artificial. Com "representação" Searle pretende identificar a "representação" da condição de satisfação de uma crença. Ele declina definitivamente o conceito do cognitivismo, em que representação é lido como a estrutura formal das relações dos objetos com suas propriedades. Para Searle representação é o conteúdo e a forma. Em resumo, cada estado Intencional consiste em um conteúdo Intencional em um modo psicológico, onde este conteúdo é uma proposição por inteiro, e onde há uma direção de ajuste no qual o conteúdo intencional determina a condição de satisfação (Searle, 1983, p. 12). Esse esclarecimento será importante quando discutirmos a relação entre Intencionalidade e Inteligência Artificial.

Naturalmente, ao explicar a linguagem em termos de Intencionalidade, Searle não quer dizer que a Intencionalidade é linguística, tampouco que ela seja derivada da Intencionalidade. É justamente o contrário. O que acontece aqui é que a linguagem revela a intencionalidade, por assim dizer, nesse sentido, Searle lista ao menos 4 pontos de conexão entre estados intencionais e os atos de fala. A ver:

1. A distinção entre o conteúdo proposicional (está nevando) e a força ilocucionária (por exemplo, afirmar que, questionar se) dos atos de fala se estende a estados mentais como crença, desejo, intenção e assim por diante. Podemos ter diferentes "atitudes" (crença, desejo, esperança, medo etc.) com o mesmo conteúdo proposicional.
2. Alguns estados mentais têm uma direção de ajuste (mente-para-o-mundo vs. Mundo-para-à-mente) correspondente à direção do ajuste (palavra-para-o-mundo vs. mundo-para-a-palavra) dos atos de fala.
3. Quando executados, os atos de fala com conteúdo proposicional expressam estados mentais com o mesmo conteúdo proposicional: ao fazer a afirmação de que P eu expresso a crença de que P, ao dar uma ordem para fazer P eu expresso um desejo ou desejo de que você faça P, e assim por diante.
4. As condições de satisfação de um ato de fala (uma afirmação é satisfeita se e somente se for verdadeira, uma ordem é satisfeita se e somente se for obedecida

> etc.) são idênticas às condições de satisfação do estado mental expresso (Searle, 1983, p. 6).

Esses pontos de semelhança levam Searle a concluir que certos estados mentais, os que tem uma direção de ajuste, representam suas condições de satisfação da mesma forma que os atos de fala. Cada estado Intencional, portanto, consiste em um conteúdo representativo em um certo modo psicológico. Estado intencionais representam objetos e estados de relações no mesmo sentido que os atos de fala representam tais coisas.

Inteligência Artificial: um olhar quanto à intencionalidade

Alan Turing propôs seu teste para inteligência baseando-se em linguagem pela simples razão de que a linguagem consegue capturar a maior parte do comportamento inteligente. Ela não apenas tem poder de transmitir um conhecimento, mas é também capaz de revelar a intenção de seus falantes. Paralelo a isso, Searle, em seu artigo de 1980, onde explorou pela primeira vez o experimento *The Chinese Room*, criticou essencialmente a ausência de intencionalidade da Inteligência Artificial, especialmente as que lidavam com linguagem natural. Mais tarde ele viria a refinar sua crítica, lidando com a ausência de significado desses agentes artificiais. O programa que Searle tinha às vistas era o *Schank* de Roger Schank e Peter Alberson. Outros programas também considerados foram ELIZA (Weizenbaum, 1966) e SHRDLU (Winograd, 1973). Esses programas ficaram popularmente conhecidos como *Chatterbots*. Houve recentemente um avanço considerável nessa classe de programas. Em 2022 a empresa OpenAI lançou ao grande público a ferramenta ChatGPT-3, programa baseado no modelo de linguagem autorregressiva GPT-3 (*Generative Pre-trained Transformer*). O programa ChatGPT-3 é provavelmente a mais avançada ferramenta de processamento de linguagem que já foi lançada até então, sendo ele capaz de gerar textos dos mais diversos tipos, textos técnicos, ensaísticos e com estilo artístico. Tendo isso em vista, naturalmente retornou ao debate público afirmações que atribuíam a essas ferramentas senciência, criatividade, pensamento, intencionalidade ou consciência (Tiku, 2022). Não temos objetivo de debater especificamente o *ChatGPT* ou a maneira como ele ou qualquer programa em específico funciona. A ideia é entendermos o panorama geral. Fato é que devido ao *hype* da inteligência artificial nos últimos dois anos, acompanhar os desenvolvimentos nesta área tornou-se um desafio, pois as empresas de tecnologia correm para competir no

desenvolvimento de diversas versões mais poderosas desses programas. No dia em 10 de março de 2023 o Google anunciou o *PaLM-E*, uma versão multimodal do modelo de linguagem *PaLM*. A Baidu apresentou seu *Chatbot ERNIE* baseado em LLM no mesmo mês e a *OpenAI* revelou sua próxima versão do GPT, o GPT-4, no dia 14 de março.

Investigamos acima sobre como a linguagem, segundo Searle, pode revelar a intencionalidade. As novas ferramentas de geração de texto possuem uma alta sofisticação ao lidar com linguagem natural. Seus textos são coesos, e muitas vezes são capazes enganar olhares incautos. Por exemplo, uma pessoa pode realmente se convencer que algo escrito pelo *ChatGPT* foi feito por um ser humano. Nesse sentido, vale dizer que a maneira como essas ferramentas lidam com a linguagem revela alguma intencionalidade? Para responder a essa pergunta vamos, em resumo, descrever como funcionam essas ferramentas e depois avaliar os pontos levantados por Searle em 1980.

O que um programa como o *ChatGPT* pode fazer? O *ChatGPT* se trata de uma *LLM* (*large language model*), programas desse tipo são versáteis. Eles podem escrever e depurar programas de computador, escrever texto técnicos e corporativos, compor músicas, escrever redações de estudantes, responder a perguntas de testes (às vezes, dependendo do teste, em um nível acima da média dos seres humanos), escrever poesias e letras de canções, recuperar e refinar informações (Tung, 2023). Ainda assim, os programas estão sujeitos a falhas e às vezes escrevem respostas que soam plausíveis, mas que estão incorretas ou não tem sentido. Esse comportamento é comum a modelos de linguagem grandes e é chamado de "alucinação" (*hallucination*). Internamente, os modelos do tipo GPT são redes neurais artificiais baseadas na arquitetura de transformador, pré-treinados em grandes conjuntos de dados de texto não rotulado e capazes de gerar novos textos semelhantes aos humanos. O modelo GPT-3 (2020) possui 175 bilhões de parâmetros e foi treinado em 400 bilhões de tokens de texto. O termo "GPT" também é usado nos nomes de alguns *LLMs* generativos desenvolvidos por outras empresas, como uma série de modelos inspirados no GPT-3 criados pela empresa EleutherAI.

Um *LLM*, ou sua forma mais simples, chamada de *language model* (modelo de linguagem), é um mecanismo probabilístico para gerar texto. Tal definição é geral o suficiente para incluir uma grande variedade de programas. No entanto, deve ser feita uma distinção entre modelos generativos, que podem, em princípio, ser usados para sintetizar texto artificial, como é caso do GPT-3, e os modelos de captura de informação, que não geram texto artificial.

Uma *language model* é, portanto, uma distribuição de probabilidade que descreve a probabilidade de qualquer *string*[2]. Quando alimentado com um enunciado como, por exemplo, "Eu ouso perturbar o universo?", tal modelo pode responder que a tradução tem uma probabilidade de 99% de ser "*Do I dare disturb the universe?*". Com um modelo de linguagem, podemos prever quais palavras provavelmente virão a seguir em um texto. Nesse sentido, o programa pode sugerir uma conclusão para um e-mail ou mensagem de texto, por exemplo (Russell; Norvig, 2020, p. 824).

Claude Shannon[3] foi o primeiro a propor um modelo de linguagem estatística, a saber, Shannon apresentou as palavras como dados estatísticos, números que podem ser calculados. Shannon considerou a linguagem como uma fonte estatística e mediu como conjuntos de palavras, chamados de modelos *n*-grama[4], podem prever ou, de forma equivalente, comprimir texto natural. Shannon fez experimentos com seres humanos para estimar o quanto o inglês é previsível e avaliou o desempenho desses modelos *n-grama* ao tentar prever o que vem depois nas frases reais. A capacidade dos modelos de linguagem de serem avaliados quantitativamente dessa maneira é uma de suas virtudes essenciais (Lafferty; Zhai; Croft; 2020).

As linguagens naturais são complexas, portanto, qualquer modelo de linguagem será, na melhor das hipóteses, uma aproximação. Não existe um modelo de linguagem definitivo para o inglês da mesma forma que existe para Python, um conjunto de sequências; "print (2 + 2)" é um programa válido na linguagem Python, enquanto "2) + (2 print" não é (Russell; Norvig, 2020, p. 824). Donald Davidson disse: "Se uma linguagem é uma estrutura compartilhada claramente definida, então não existe linguagem" (Davidson, 1986, p. 174).

Nesse sentido, podemos resumir que os *LLMs* são modelos formais que lidam com padrões em linguagem natural, sem realmente possuírem algum tipo de "fotografia" das *strings* que eles lidam. Parte da crítica atual aos *LLM* argumenta justamente que em uma possível experiência com objetos reais, uma *LLM* não se sairia bem, uma vez que a linguagem com a qual eles lidam não possui conexões com o mundo físico real (Bender; Koller, 2020). Isso pode ser comprovado pela ausência de intenção nos enunciados

[2] Uma sequência de caracteres, geralmente utilizada para representar palavras, frases ou textos de um programa.
[3] Claude Elwood Shannon. Matemático, engenheiro eletrônico e criptógrafo estadunidense, conhecido como "o pai da teoria da informação". Seu trabalho com modelos de linguagem está Shannon (1951).
[4] "N-grama" é um termo utilizado na área de processamento de linguagem natural. Trata-se de uma sequência contígua de *n* itens em um enunciado. Esse "n" em "n-grama" representa o número de itens na sequência (Jurafsky; Martin, 2009).

gerados por uma *LLM*. Searle, em 1980, listou em síntese alguns argumentos baseados na teoria da intencionalidade que invalidam a tese de que esses programas podem possuir alguma intencionalidade.

Primeiro, Intencionalidade é um produto das funções causais do cérebro. Tal relação causal entre processos mentais e o cérebro é um fato empírico. Sabe-se que para Searle a mente é algo natural, sendo o cérebro uma máquina biológica. O que Searle busca esclarecer é toda a cadeia de relações entre as regras formalmente extraíveis da linguagem e a Intencionalidade, que não é formalmente extraível. Dentro disto, está a capacidade dos Atos de fala em representar objetos e relações no mundo da mesma forma que ocorre com a Intencionalidade, onde Searle esclarece as ligações entre os Atos de fala e a Intencionalidade a partir dos quatro pontos mencionados em §2. Bender e Koller (2020), baseando-se na teoria do significado de Grice (1968) e usando as mesmas premissas de Searle (1980), esclarecem que a forma é qualquer conteúdo observacional da linguagem, enquanto o significado é a conexão entre a forma e algo externo. Significado não é o mesmo que Intencionalidade, mas compreende-se que cada enunciado possuí uma intenção comunicativa. Os *LLM*s, por sua vez, capturam padrões semânticos formais, e entre esses padrões estão também os usos que damos a certas sentenças, suas intenções não são capturadas. Ainda que consideremos o argumento representacional conexionista, em que as representações das redes neurais artificiais conseguem representar formalmente as propriedades dos objetos (Clark, 2015, p. 492-493), este é apenas um reflexo fraco do significado real. As representações neurais não se qualificam como significados permanentes, nem como interpretações, nem como intenções comunicativas, sendo insuficientes para, por exemplo apontar relações no mundo real (Bender; Koller, 2020). E o que seria essa casualidade, em suma? Basicamente aquilo que mencionamos em §2 sobre a ação causal ser autorreferenciável. Nesse sentido, uma intenção só é satisfeita se o agente intende a algo com seu enunciado. As *LLM*s em nada intendem com seus enunciados, uma vez que elas carecem de conteúdo representacional.

Segundo, uma vez que operam de maneira formal, os programas são meramente instanciados. Isso é perceptível no caso dos *LLM*s quando lidamos com os problemas chamados "alucinações". Apesar do nome, as "alucinações" dos *LLM* pouco tem a ver com as alucinações humanas, onde o que ocorre são percepções falsas que, entre outras coisas, levam a falhas de julgamento. As *LLM*s por seu turno, acabam por sofrer ou com a quantidade insuficiente de dados, ou com a ausência de treinamento quanto às informações que ela está lidando. Por isso, uma *LLM* pode gerar respostas confidentes

que tenham resultados não factíveis ou absurdos. Não existe uma interpretação de "bom senso" por parte de uma *LLM*, apenas o uso de *datasets* ou modelos de treinamento ruins (Dziri et al, 2022, p.1).

Terceiro e quarto, não se pode explicar que o cérebro produz intencionalidade da mesma maneira que um computador instância um programa. Isto é, a intencionalidade não é um "software" do cérebro. E para se criar intencionalidade artificialmente é necessário que a máquina duplique as funções causais do cérebro humano. Essas duas últimas conclusões são derivadas das duas primeiras, que podemos resumir dizendo que o cérebro não "simula" um programa que interpreta os enunciados que recebe. Não existe uma rede neural artificial rodando em segunda instância, o cérebro é a própria rede neural.

Conclusão

Ainda que ferramentas do tipo LLM demonstrem uma impressionante capacidade de geração de conteúdo, a ponto de se equiparar com seres humanos, não é possível que exista alguma intenção comunicativa por detrás dos prompts gerados por essas ferramentas, conforme demonstra a teoria da intencionalidade de Searle, em que um enunciado intende a uma ação auto referenciável que pode ser satisfeita ou não. Em adição a isso, somamos o argumento de Bender e Koller, no qual a ausência de experiência no mundo real e a possível incapacidade das LLM em lidar com objetos pode comprovar que, por baixo de cada resultado gerado para cada um os prompts que alimentam um LLM, não há outra coisa senão um resultado probabilístico para tal e tal *string*.

Referências

AUSTIN, J. L. **Philosophical Papers**. Oxford University Press, 1979.
ANSCOMBE, Elizabeth. **Intention**. Harvard University Press, 2000.
BENDER, Emily M.; KOLLER, Alexander. On Meaning, Form, and Understanding in the Age of Data. **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, p. 5185–5198, 2020.
CLARK, Stephen. Vector space models of lexical meaning. In: LAPIN, Shalom; FOX, Chris (Eds.). **Handbook of Contemporary Semantic Theory, 2nd edition**. Wiley-Blackwell, 2015.
DAVIDSON, Donald. A Nice Derangement of Epitaphs. In: GRANDY, Richard E.; WARNER, Richard (Eds.). **Philosophical Grounds of Rationality**. Clarendon Press, 1986.
DRETSKE, Fred. The Intentionality of Perception. In: SMITH, Barry. **John Searle**. Cambridge University Press, 2003.
DZIRI, Nouha; MILTON, Sivan; YU, Mo; ZAINE, Osmar; REDDY, Siva. On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models? In:

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p. 5271 – 5285, 2022.
GRICE, Paul H. Utterer's Meaning, Sentence Meaning, and Word-Meaning. In: KULAS, Jack; FETZER, James H.; RANKIN, Terry L.; **Philosophy, Language and Artificial Intelligence**. Kluwer Academic Publishers, 1988.
JURAFSKY, Dan; MARTIN, James H.; **Speech and Language Processing**: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson Prentice Hall, 2009.
LAFFERTY, John; ZHAI, Cheng Xiang; CROFT, W. Bruce. **Language Modeling for Information Retrieval**. Springer, 2020.
LEVINSON, Stephen C. **Pragmatics**. Cambridge University Press, 1983.
RUSSELL, Stuart; NORVIG, Peter. **Artificial Intelligence, A Modern Approach**. 4th Edition. Pearson, 2020.
SEARLE, John R. **Speech Acts**. Cambridge University Press, 1969.
SEARLE, John R. **Minds, Brains, and Programs**. Behavioral Brain Sciences, 1980a.
SEARLE, John R. **Intentionality**. Cambridge University Press, 1983.
SEARLE, John R. **Consciousness and Language**. Cambridge University Press, 1995.
SEARLE, John R. **The Mystery of Consciousness**. Cambridge University Press, 1992.
SEARLE, John R. **The Construction of Social Reality**. Free Press, 1995.
Shannon, Claude Elwood. Prediction and entropy of printed English**. Bell System Technical Journal**, Vol. 30, p. 51-64, 1951.
SMITH, Barry. **John Searle**. Cambridge University Press, 2003.
ZAIBERT, Leo. Intentions, Promises, and Obligations. In: SMITH, Barry. **John Searle**. Cambridge University Press, 2003.
PENCO, Carlo. **Introdução a Filosofia da Linguagem**. Editora Vozes, 2006.
TIKU, Nitasha. "The Google engineer who thinks the company's AI has come to life". **The Washington Post**, June 11, 2022. ISSN 0190-8286. Disponível em: [https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/]. Acesso em: 12 de junho de 2022.
TUNG, Liam. "ChatGPT can write code. Now researchers say it's good at fixing bugs, too". **ZDNET**, January 26, 2023. Disponível em: [https://www.zdnet.com/article/chatgpt-can-write-code-now-researchers-say-its-good-at-fixing-bugs-too/]. Acesso em: 30 de janeiro de 2023.
WEIZENBAUM, Joseph. "ELIZA—a computer program for the study of natural language communication between man and machine". **Communications of the ACM**, 9, 1966, p. 36–45. doi:10.1145/365153.365168. S2CID 1896290.
WINOGRAD, Terry. "Procedures as a Representation for Data in a Computer Program for Understanding Natural Language". [https://dspace.mit.edu/handle/1721.1/7095]. Acesso em: [Junho de 2023]. What's the next word in large language models? **Nat Mach Intell** 5, 331–332 (2023). https://doi.org/10.1038/s42256-023-00655-z
What's the next word in large language models?. **Nat Mach Intell** 5, 331–332 (2023). https://doi.org/10.1038/s42256-023-00655-z.

Recebido em: 13/10/2023
Aprovado em: 03/05/2024