Revista de Estudantes de Filosofia da Universidade de Brasília

10.26512/pl.v13i30.56911

O ANTROPOMORFISMO NA INTELIGÊNCIA ARTIFICIAL CONVERSACIONAL

ANTHROPOMORPHISM IN CONVERSATIONAL ARTIFICIAL INELLIGENCE

Kleyber Porto[1]
Universidade de Brasília

[1]Graduando em Filosofia pela Universidade de Brasília (UnB).
E-mail: klleybersantana@hotmail.com.
Lattes: http://lattes.cnpq.br/3230380352034916. Orcid: https://orcid.org/0009-0005-7688-0029.

RESUMO: Apesar do campo da inteligência artificial (IA) ter se utilizado em larga escala da antropomorfização como interação entre humanos e máquinas, nota-se a pouca atenção dispensada às pesquisas do antropomorfismo como categoria da IA. Essa atenção é ainda mais reduzida quando se trata de discutir filosoficamente a conceituação de uma IA antropomorfizada. Este trabalho tem por objetivo compreender filosoficamente a amplitude do antropomorfismo da IA conversacional. Dessa forma, em um primeiro momento, foram apresentados os aspectos formadores dessa tecnologia. Em seguida, investigou-se os mecanismos do antropomorfismo que a estruturam. Por fim, para alicerçar a discussão, utilizou-se a abordagem funcionalista de Daniel Dennett, bem como os conceitos presentes na filosofia de John Searle sobre intencionalidade e o experimento do quarto chinês. Este trabalho desenvolve, portanto, uma comparação entre esses dois filósofos sobre o antropomorfismo na IA conversacional.

Palavras-chave: Antropomorfismo. IA fraca conversacional. Daniel Dennett. John Searle.

ABSTRACT: Although the field of artificial intelligence (AI) has made extensive use of anthropomorphism as an interaction between humans and machines, little attention has been paid to research on anthropomorphism as a category of AI. This attention is even less when it comes to philosophically discussing the conceptualization of an anthropomorphized AI. This paper aims to philosophically understand the scope of anthropomorphism in conversational AI. Thus, first, the formative aspects of this technology were presented. Then, the mechanisms of anthropomorphism that structure it were investigated. Finally, to support the discussion, Daniel Dennett's functionalist approach was used, as well as the concepts present in John Searle's philosophy on intentionality and the Chinese room experiment. This work therefore develops a comparison between these two philosophers on anthropomorphism in conversational AI.

Keywords: Anthropomorphism. Conversational weak AI. Daniel Dennett. John Searle.

INTRODUÇÃO

O crescimento da Inteligência Artificial (IA) como tecnologia de propósito geral (TPG)[2] modificou o eixo da produção e conhecimento humanos. Composta por categorias variadas, a IA nos conduz para um caminho ético, social e tecnológico ainda incerto que se transforma a todo momento. Dentre as categorias que sofreram drásticas mudanças destaca-se a IA conversacional que se originou com o programa ELIZA e se consolidou com os *Chatbots* (*ChatGPT, DeepSeek, Microsoft Copilot, Gemini, Intercom*, etc.). Dito isto, pode-se notar, que existe um elemento comum que trespassa esse tipo de IA e que, desde ELIZA, tem permanecido incólume, o antropomorfismo. A IA conversacional utiliza uma forte base antropomórfica para interagir conosco por meio da linguagem natural. A interação inicial feita pelo usuário com comandos de voz ou texto é decodificada pelo algoritmo e devolvido para nós em forma de linguagem natural por texto ou voz.

Para se investigar uma IA com características antropomórficas, faz-se necessária a utilização de um método analítico, dividindo-se os termos em partes menores. Assim, pode-se conceituar a IA como um campo do conhecimento que está ligado ao raciocínio, à inteligência, à linguagem e à resolução de problemas (Kaufman, 2019, p. 19). Quanto ao antropomorfismo, podemos explicá-lo brevemente como o ato de atribuir qualidades humanas aos animais e coisas não-humanas, ainda que sem uma justificativa concreta ou racional (Shettleworth, 2010, p. 477). Infere-se, nesse sentido, que uma IA antropomorfizada é um campo do conhecimento com um conjunto de tecnologias que interagem com o raciocínio, inteligência, linguagem e resolução de problemas, adotando qualidades humanas que, neste caso específico, lhe foram originalmente atribuídas ou propositadamente colocadas.[3]

Ao se verificar o estado da arte em relação à IA conversacional, pôde-se constatar que em alguns dos trabalhos encontrados o antropomorfismo é citado como componente formal de robôs, como artifício de captura de dados, como um fenômeno relacionado à ética ou ainda como elemento tecnológico que gera estados psicológicos no usuário[4]. Desse modo, dentre os vários fios condutores

[2] São tecnologias que impactam profundamente a humanidade. Pode-se citar como TPG a eletricidade, os computadores e a IA.
[3] Nota-se que o antropomorfismo não é uma característica do próprio objeto em si, mas relacional com aquele humano que o observa. É relacional pois, é uma situação onde um humano lê o objeto como similar a si próprio.
[4] DARLING, Kate. 173"Who's Johnny?" Anthropomorphic Framing in Human–Robot Interaction, Integration, and Policy. *In*: LIN, P.; ABNEY, K.; JENKINS, R. (org.). *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford University Press, 2017. KRONEMANN, Bianca. *et al.* How AI encourages consumers to share their secrets? The role of anthropomorphism, personalisation, and privacy concerns and avenues for future research. *Spanish Journal of Marketing - ESIC*, v. 27, n. 1, p. 3–19, 2023.

que conectam as características da existência singular da IA conversacional, investigar o antropomorfismo como uma trama que constrói parte de sua existência, parece ser uma análise possível. Além disso, dada a constatação de que o antropomorfismo se apresenta de forma recorrente e elemento constituinte de determinada categoria da IA justifica-se a reflexão sobre o tema investigando-o sobre uma base filosófica.

Este trabalho tem por objetivo compreender filosoficamente a amplitude do antropomorfismo na IA conversacional. Para tanto, investigou-se primeiramente os conceitos e modelos que levaram expansão tecnológica da IA fraca conversacional que hoje conhecemos.[5] Em um segundo momento foram apresentados os mecanismos do antropomorfismo que estruturam sistemas artificiais capazes de reproduzir interações como um humano, podendo aparentar estados intencionais como um humano, além de simular desejos ou interagir conosco de maneira parecida àquela que humanos fazem. Por fim, recorreu-se ao funcionalismo de Daniel Dennet e os conceitos das posturas intencionais para investigar a manifestação do antropomorfismo na IA conversacional, para então, inserir os conceitos de intencionalidade de John Searle e o experimento do quarto chinês como contraponto e argumento para a existência restrita da antropomorfização na IA fraca. Dessa forma, para que o antropomorfismo da IA conversacional possa ser compreendido, é necessário conceituar brevemente o que é a IA.

1 A CONCEITUAÇÃO DA IA

A obtenção de um claro conceito para definir a IA esbarra em diferentes vertentes de pesquisas que deram origem a diferentes modelos de máquinas e inúmeros sistemas, dificultando a sua conceituação. Apesar de sempre existirem formas diferenciadas de pesquisa na história da IA,[6] por volta de 1980 foram cunhados termos para diferenciar comunidades de pesquisadores com propostas distintas de trabalho: os *scruffies* (desgrenhados) e os *neats* (empertigados). Sejam arquiteturas elaboradas por *scruffies* ou *neats*, para Costa e Cozman (2024, p. 137), nestas duas últimas décadas o

COECKELBERGH, Mark. Are Emotional Robots Deceptive?. **IEEE Transactions on Affective Computing**, v. 3, n. 4, p. 388–393, 2012.

[5] Nesta seção foi elaborada uma breve explanação sobre as primeiras tecnologias que foram relevantes para o desenvolvimento da IA. Esse breve *background* é necessário para situar o leitor sobre conceitos que serviram de base para estruturar a IA a partir de 1956.

[6] De acordo com Cozman (2021, p. 7), a história da IA foi marcada pela presença de dois modos de pesquisa. O primeiro estilo era baseado no empirismo. Se estruturava nas observações biológicas e psicológicas dos seres vivos. Além disso, viam com simpatia as arquiteturas complexas que surgem do contato entre conhecimentos díspares. A segunda linha de pesquisa era analítico e se apoiava em argumentos matemáticos e lógicos. Se sustentava em princípios gerais e organizadores e se interessava pelas concepções abstratas da inteligência.

desenvolvimento da IA teve como foco a extração de padrões das enormes bases de dados disponíveis. Isso gerou pesquisas e desenvolvimento em áreas como a Computação, a Linguística, a Matemática e a Neurociência (Kaufman, 2019, p. 19).

Foi a partir dessas diferentes comunidades de pesquisadores e diferentes formas de abordagens da IA, que várias expressões foram cunhadas: análise preditiva, modelagem estatística, reconhecimento de padrões, sistemas adaptativos, ciência de dados, sistemas de auto-organização ou, simplesmente, inteligência artificial (Kaufman, 2019, p. 26). Além destes, existem outros termos como IA fraca (ANI)[7], IA forte (AGI)[8], Superinteligência Artificial (ASI)[9] e singularidade que requerem uma conceituação preliminar.

Em 1980, John R. Searle apresentou o artigo intitulado "*Minds, brains and programs*", que apontou novos contornos sobre a diferenciação das IA. O primeiro ponto relevante é a diferenciação que Searle fornece para a IA fraca e IA forte. A primeira é entendida como um instrumento para o estudo da mente, que permite "formular e testar hipóteses" com maior acuidade (Searle, 1980, p. 417). A IA fraca é a própria IA que hoje conhecemos e integra desde os *bots* conversacionais, até os programas de reconhecimento facial e geração de imagens. A IA fraca pode parecer inteligente, mas ela não possui inteligência ou consciência[10] como os humanos. A IA fraca do tipo conversacional apenas simula um comportamento antropomórfico, apesar dos avanços nesta área serem surpreendentes (Huawei Technologies, 2023, p. 5).

Para Haymond e McCudden (2021, p. 1641) a IA fraca, estreita ou particular refere-se àquela que já existe e que é capaz de solucionar problemas específicos, para a qual os algoritmos são especificamente desenvolvidos, projetados e validados. Ela é, portanto, limitada e funciona dentro de um escopo definido. Tanto a IA forte quanto a Superinteligência Artificial ainda não existem e não se sabe se algum dia poderão existir. A AI forte pode ser conceituada como aquela capaz de resolver quaisquer tarefas intelectuais como um humano, tomar de decisões complexas, possuir autonomia para aprender em múltiplos contextos e possuir criatividade (Haymond e McCudden, 2021, p. 1640). Seria mais do que uma simples máquina que recebe instruções por meio de algoritmos; seria uma máquina que, se corretamente programada, possuiria estados cognitivos como os humanos possuem (Searle, 1980, p. 417). A Superinteligência Artificial seria derivada da IA forte e superaria a capacidade humana na inteligência, criatividade, resolução de problemas e está ligada diretamente ao conceito singularidade. Ainda não se sabe se esse tipo de IA teria autoconsciência ou subjetividade. Por fim, a

[7] *Artificial Narrow Intelligence.*
[8] *Artificial General Intelligence.*
[9] *Artificial Superintelligence.*
[10] Trataremos, por enquanto, com o conceito da consciência pelo senso comum, do cotidiano, definindo-a como a percepção de si.

singularidade nada mais é do que a hipótese do momento do surgimento da Superinteligência Artificial (Haymond e McCudden, 2021, p. 1640).

Apesar da grande variação de nomes e funcionalidades da IA o objetivo ainda é o desenvolvimento de algoritmos[11] que visem à solução de problemas específicos (Kaufman, 2019, p. 26). Uma outra definição também amplamente aceita foi cunhada por John McCarthy na Conferência de Dartmouth, em 1956. Para McCarthy, a IA consistia em fazer com que uma máquina simulasse, de maneira precisa, o comportamento inteligente de humanos. O mesmo pesquisador definiria a IA, décadas depois, como uma ciência e engenharia capaz de fazer computadores, máquinas ou programas inteligentes (McCarthy, 2004, p. 2). De acordo com Haugeland (1985, p. 2), a IA pode ser definida como o esforço para produzir computadores que realmente pensassem em um sentido pleno. Para Copeland (2000), ela é a ciência capaz de fazer com que computadores desempenhem tarefas que exijam inteligência para concluí-las, caso fossem feitas por humanos.

Esses conceitos têm em comum a procura de uma tecnologia que reproduza características humanas, tais como inteligência, comportamento e cognição, ou que utilize a inteligência humana como base comparativa para a resolução de problemas. Esse tipo de base conceitual levou os pesquisadores a ponderarem sobre as características, o modo de interação e os objetivos dos primeiros modelos de IA, dos quais destacam-se o Teste de Turing, de 1950, e o programa ELIZA, de 1965.

2 O TESTE DE TURING

O desenvolvimento da IA como tecnologia teve início, aproximadamente, na década de 1950. Um dos responsáveis por este acontecimento foi o inglês Alan Turing que, em 1930, atuava como matemático, tornando-se conhecido pela decodificação das mensagens criptografadas das forças armadas alemãs durante a Segunda Guerra. Tal fato foi responsável pela vantagem tática e vitória dos Aliados sobre o Eixo. Dessa maneira, Turing demostrou a relevância do conhecimento computacional[12]. Não obstante, foi numa palestra sobre inteligência computacional que apresentou na

[11] Pode-se definir algoritmos como um conjunto não-vazio de instruções precisas, não ambíguas para computar uma função numérica, escrita em linguagem simbólica ou natural (Gomes, 2023, p. 89).

[12] Vale destacar também que o trabalho seminal de Turing é o artigo de 1936 "*On Computable Numbers, with an Application to the Entscheidungsproblem*" sobre o problema da decisão de Hilbert, no qual ele define o que hoje chamamos de "Máquina de Turing". Uma máquina de Turing é uma formalização do conceito de "função computável", precursora dos computadores modernos. A definição de Turing é, não só importante para o entendimento deste conceito, mas é provadamente equivalente a qualquer definição de função computável que se encontre. Isso significa que, a despeito da complexidade, *qualquer* computador moderno pode ser reduzido à uma máquina de Turing (embora, empiricamente, isso seja deveras difícil; na teoria, contudo, é

Inglaterra em 1947, além da publicação de seu artigo intitulado Computing Machinery Intelligence, em 1950, que a computação se posicionou definitivamente como ciência dentro dos meios acadêmicos. É também inegável a contribuição de Turing para o desenvolvimento do que conhecemos hoje como Ciências da Computação e do conceito de algoritmo (Gomes, 2023, p. 23).

Foi por meio de seu artigo Computing Machinery Intelligence que Turing formulou um questionamento filosófico que iria atravessar toda a pesquisa sobre a IA indagando se as máquinas podem pensar (Turing, 1950, p. 433). Porém, ele percebeu que deveria primeiramente retirar qualquer ambiguidade da palavra "pensamento". Para evitar tais armadilhas, Turing (1950, p. 442) procurou obter respostas alterando a forma da pergunta[13] questionando-se sobre a possibilidade de computadores digitais obterem um resultado satisfatório no Jogo da Imitação (ou Teste de Turing), enganando um humano.

O Teste de Turing foi desenvolvido com o propósito de testar a capacidade de uma máquina digital imitar uma determinada parcela do comportamento humano. Essa parcela abrange a capacidade humana de dialogar e responder as mais variadas questões da mesma forma que um humano faria. Resumidamente, o teste[14] é composto por três participantes (A, B, C) nos quais, A deve ser uma máquina digital, B é humano e C atuará como um interrogador também humano. O interrogador ficará separado dos participantes A e B, que serão chamados de X ou Y, e se comunicarão por mensagens escritas.

O objetivo do teste é que o interrogador C é fazer todo tipo de perguntar para conseguir saber por indicações dos outros dois participantes, quem é a máquina digital. O objetivo de A é fazer com que C erre na identificação dos participantes. O objetivo de B é fazer com que C acerte na identificação

possível e provado ser o caso). Isso nos leva a tese Church-Turing como base para afirmar a equivalência das múltiplas definições de função computável. Gomes (2023, p. 97) define a tese Church-Turing como uma afirmação que nos diz que toda função algorítmica é "Turing-computável" e vice-versa.

[13] É necessário esclarecer que esta estratégia empregada por Turing no seu artigo não é *ad hoc.* Na prática, muitas de nossas habilidades não possuem definições precisas, assim como o termo "inteligência" também não possui. Não temos, por exemplo, uma definição exata sobre o que é "cozinhar", muito menos "cozinhar bem". Se assim o fosse, os testes para uma vaga de *chef* de cozinha seriam teóricos e objetivos. Como não temos uma definição "clara e distinta" para esses e muitos outros termos, faz sentido aplicar um teste situacional na tentativa de convencer um avaliador qualificado de que o candidato à vaga, que se comporta indistintamente à uma pessoa que cozinha bem, realmente cozinha bem. Pois bem, a estratégia de Turing é a mesma. Perguntar o que é inteligência torna-se inútil, uma vez que Turing não tem uma definição precisa desse termo. Ao invés disso, o melhor seria se perguntar: "Em que circunstâncias uma pessoa pareceria agir de forma inteligente?" Nesse sentido, o teste de Turing deve funcionar com máquinas porque é exatamente dessa forma avaliamos humanos como seres "inteligentes".

[14] A versão do Teste de Turing apresentada neste trabalho é denominada o "teste das espécies" (*the species test*), uma variação do "teste de gênero" (*The Gender Test*) do artigo de 1950. No "teste de gênero" os participantes tentam enganar o interrogador quanto ao gênero, ou seja, o homem deve fazer-se passar por mulher. Caso consiga, pode-se dizer que o computador passou no teste (Turing, 1950, p. 433). A variação mais utilizada do teste original é o "teste das espécies". O objetivo é a máquina digital enganar o interrogador passando-se por humano.

dos participantes. O teste de Turing terá sucesso se, ao final do teste, o interrogador tiver sido enganado pela máquina digital.

É a partir desse cenário que Turing elabora novos questionamentos para substituir a pergunta original "As máquinas podem pensar?"[15] (Turing, 1950, p. 434, tradução nossa). As perguntas, agora, devem ser de uma outra natureza: "O que aconteceria quando uma máquina tomasse o lugar de A neste jogo?"[16] (Turing, 1950, p. 434, tradução nossa). O matemático se questiona se a quantidade de erros da máquina seria diferente do número de erros do interrogador humano. Segundo Turing (1950, p. 442), computadores digitais universais[17], com um poder de processamento superior, poderiam ter um desempenho satisfatório no jogo.

O critério de Turing para definir se algo ou alguém poderia pensar, estrutura-se na possibilidade da atribuição de estados mentais a esse algo ou alguém. É um critério operacional. Baseia-se no funcionamento e no comportamento de máquinas ou organismos e não tem a obrigatoriedade de se debruçar sobre a natureza de estados mentais responsáveis por cenários comportamentais. Disso resulta que se o comportamento de uma máquina não puder ser diferenciado do comportamento humano[18], não haveriam impedimentos para atribuir-lhe estados mentais, infere Turing. Nesse sentido, Turing define o pensamento em termos pragmáticos e operacionais (Teixeira, 2008, p. 17-18). Com isso, a publicação do artigo de Turing foi a responsável pela busca ambiciosa de uma IA inteligente, dando origem a novos tipos de IA, como o programa ELIZA de 1965.

3 ELIZA

ELIZA foi um programa de computador desenvolvido na década de 1960 por Jospeh Weizenbaum, no MIT (Massachusetts Institute of Technology). Surgiu como um dos primeiros experimentos de IA do mundo e também foi o primeiro programa de conversação inventado. A partir de regras definidas e respostas de usuários, ELIZA podia dar respostas acertadas ao escolher as mais apropriadas entre os arquivos de uma base de dados (Huawei Technologies, 2023, p. 5). ELIZA era parte de um experimento projetado para interagir com humanos ou usuários utilizando a língua inglesa

[15] "Can machines think?" (Turing, 1950, p. 434).
[16] "What will happen when a machine takes the part of A in this game?" (Turing, 1950, p. 434).
[17] Turing define computadores digitais universais como aqueles que que podem imitar qualquer máquina de estado discreto.
[18] Computadores não choram ou riem, a solução de Turing foi estruturar o teste sob uma base linguística.

e o processamento da linguagem natural (Bassett, 2019, p. 804). Também pode ser considerado o primeiro chatbot ou IA conversacional e foi elaborado para imitar um terapeuta rogeriano[19].

A escolha da "personalidade" de um terapeuta rogeriano é descrita pelo próprio Weizenbaum como apenas uma questão de conveniência de pesquisa pois, "o psicoterapeuta rogeriano é relativamente fácil de imitar porque grande parte de sua técnica consiste em atrair seu paciente refletindo as declarações do paciente de volta para ele" (Weizenbaum, 1976, p. 3, tradução nossa)[20]. O termo "rogeriano" que é utilizado pelo pesquisador não corresponde rigorosamente à pisicoterapia postulada por Carl Rogers. Serve de modelo, um contorno quase estereotipado de comportamento da IA ELIZA que nas próprias palavras de Weizenbaum funcionaria mais como uma "paródia": "[...] eu dei à ELIZA um script[21] projetado para permiti-la desempenhar (eu deveria, na verdade, dizer parodiar) o papel de um psicoterapeuta rogeriano envolvida em uma entrevista inicial com um paciente" (Weizenbaum, 1976, p. 3, tradução nossa)[22].

Segundo o próprio Weizenbaum (1976, p. 3), ELIZA era como um analisador de linguagem que seguia um conjunto de regras. A inteligência artificial ELIZA era um bot escrito em SLIP[23]. Como tal, havia sido planejado para interagir com interlocutores humanos, simulando ou personificando suas falas. Eliza era composto por duas camadas. A primeira camada era um analisador de linguagem e a segunda era um script chamado DOCTOR, que estipulava, orientava e organizava as regras para que a interação se aproximasse de um terapeuta humano. À vista disso, ELIZA poderia manter uma conversa sobre insetos ou quasares desde que seu script fosse ajustado adequadamente (Weizenbaum, 1976, p. 3).

Com esse experimento surgiriam resultados não previstos por Weizenbaum. DOCTOR tornou-se conhecido no MIT e posteriormente conhecido e replicado em outras instituições. O fato de pessoas interagirem com DOCTOR num nível de envolvimento profundo como se ele realmente fosse um terapeuta fornecendo algum tipo de aconselhamento, deixou o pesquisador perplexo:

---

[19] O termo "rogeriano" faz menção à Escola Rogeriana fundada pelo psicólogo humanista norte-americano Carl Rogers a partir de 1940. A escolha desse modelo por Weizenbaum foi estratégica. A técnica rogeriana fornece uma impressão de empatia e acolhimento ao paciente e poderia ser simulada por regras simples devolvendo ao paciente as suas próprias palavras sob a forma de perguntas. Ideal para o sucesso do experimento de Weizenbaum, uma vez que ELIZA era composta por regras e *scripts* limitados.

[20] "The rogerian psychotherapist is relatively easy to imitate because much of his technique consist of drawing his patient out by reflecting the patient's statements back to him" (Weizenbaum, 1976, p. 3).

[21] *Scripts* são como estruturas cognitivas esquemáticas, modelos mentais e que representam sequencialmente eventos previsíveis como, por exemplo, fazer um bolo, pedir um hamburguer, etc. O Modelo de *Scripts* foi criado por Roger Schank. Weizenbaum (1976, p. 3), os definiria como conjuntos de regras como aquelas que fornecemos a atores para que eles possam improvisar em torno de um tema específico.

[22] "[...] I gave ELIZA a script designed to permit it to play (I should really say parody) the role of a Rogerian psychotherapist engaged in an initial interview with a patient" *(*Weizenbaum, 1976, p. 3).

[23] Sigla para *Symetric List Processor*. É uma linguagem de programação de computadores implementada inicialmente como uma extensão da linguagem FORTRAN para processar listas e foi desenvolvida por Weinzenbaum.

> Fiquei surpreso ao ver quão rápido e quão profundamente as pessoas conversando com DOCTOR se envolveram emocionalmente com o computador e quão inequivocamente o antropomorfizaram. Uma vez minha secretária, que tinha me observado trabalhar no programa por muitos meses e, portanto, certamente sabia que era apenas um programa de computador, começou a conversar com ele. Depois de apenas algumas trocas com ele, ela me pediu para sair da sala (Weizenbaum, 1976, p. 6, tradução nossa)[24].

Nota-se que, diferentemente do Teste de Turing no qual o seu sucesso depende de "enganar" uma pessoa fazendo a máquina se passar por humano, a natureza do experimento aqui (ELIZA/DOCTOR) é colocada às claras: é uma máquina imitando um comportamento humano. Por certo, a força do antropomorfismo na IA conversacional não reside na perfeição para imitar humanos. Como Weizenbaum (1976, p. 7) pôde verificar nós sabemos por observações cotidianas e pelo senso comum que pessoas adquirem vínculos e fortes laços emocionais com objetos. Entretanto, surpreendentemente, exposições extremamente curtas a um programa relativamente simples e limitado induziram "pensamentos delirantes poderosos em pessoas bastante normais" (Weizenbaum, 1976, p. 7, tradução nossa). Há, portanto, mecanismos relacionais, profundos que despertam em nós, um olhar antropomorfizante para com as coisas como a IA conversacional.

A partir desse cenário inusitado, ELIZA tornou-se um fenômeno rapidamente. Era uma celebridade em círculos científicos e foi o programa computacional mais citado da história. No entanto, seu idealizador acreditava que a atenção concedida ao programa era injustificada e equivocada, pois imaginar que uma máquina pudesse ser programada para se tornar um terapeuta era, para ele, uma ideia perversa. Para Weizenbaum, com o advento do programa ELIZA, se consolidou uma visão delirante sobre a IA que contaminou tanto imaginário comum quanto a comunidade especializada da ciência, da computação e da psicologia (Bassett, 2019, p. 13). Apesar de Weizenbaum ter uma visão crítica sobre a abrangência de sua própria criação, os conceitos contidos no ELIZA foram importantes para a evolução de uma tecnologia que nos conduziu à IA conversacional que conhecemos hoje.

[24] "I was startled to see how quickly and how very deeply people conversing with DOCTOR became emotionally involved with the computer and how unequivocally they anthropomorphized it. Once my secretary, who had watched me work on the program for many months and therefore surely knew it to be merely a computer program, started conversing with it. After only a few interchanges with it, she asked me to leave the room" (Weizenbaum, 1976, p. 6).

4 O QUE TEMOS

Uma das formas mais recentes de IA que ganhou notoriedade no meio científico, foi desenvolvida entre 1950 e 1980 e ficou conhecida como IA clássica, IA simbólica ou GOFAI[25]. O diferencial dessa nova tecnologia computacional é que ela tinha como base uma "representação simbólica de problemas, lógica e busca (*search*)" (Gomes, 2023, p. 18). Objetivava o desenvolvimento de computadores digitais que poderiam reproduzir a inteligência humana, a maneira como os humanos pensam, tomam decisões e se comportam.

De acordo com a linha de pensamento da IA simbólica, características como pensar, resolver problemas, ter adaptabilidade, além de consciência e autoconsciência, poderiam ser implementados na IA utilizando simplesmente a manipulação de símbolos. Para que isso fosse alcançado, Gomes (2023, p. 18) destaca que a IA simbólica deveria operar com sistemas especialistas[26]. A IA simbólica pode ser utilizada concomitantemente às redes neurais e aprendizado de máquina. Apesar de ser amplamente utilizada, em relação a padrões de aprendizagem, ela não é tão eficaz como a Conexionista.

Para Gomes (2023, p. 18), a IA conexionista trouxe uma nova abordagem tecnológica dentro do campo das IA. Embora a sua conceituação seja conhecida desde a década de 1940, ela só pôde ser plenamente desenvolvida a partir da década de 80. Isso ocorreu porque, para ser implementada integralmente, a IA conexionista necessita de uma grande capacidade computacional e de recursos que não estavam disponíveis na época. Diferentemente da IA simbólica, que exigia que o programador inserisse dados específicos de um determinado problema, a IA conexionista teve como base constitutiva a simulação de componentes neuronais do cérebro humano. À vista disso, o próprio algoritmo examinava de forma automática os dados fornecidos e conseguia deduzir os padrões do problema. Por conseguinte, foi com o surgimento das redes neurais artificiais[27] que o *Machine Learning* (aprendizado de máquina) e o *Deep Learning* (aprendizado profundo) puderam ser desenvolvidos.

---

[25] Sigla para *Good Old Fashioned Artificial Intelligence*.

[26] *Os sistemas especialistas demandam que o programador do algoritmo tenha conhecimento sobre o problema de que a IA se ocupará, para que ela possa raciocinar e tomar decisões.*

[27] Esse tipo de modelo computacional é inspirado nas redes neurais do cérebro humano. O cérebro humano recebe informações, as processa e toma decisões. É um movimento contínuo, constante e ininterrupto. Da mesma maneira, os conjuntos de algoritmos analisam constantemente os dados, atualizando a as suas previsões todo momento. Cada algoritmo representa uma determinada quantidade de neurônios que recebem e calculam os dados recebidos (*input*) e os devolvem analisados e valorados, sob a forma de *outputs* (Gomes, 2023, p. 19).

De forma breve podemos conceituar o aprendizado de máquina e o aprendizado profundo como IA Conexionistas. Por meio da análise estatística e preditiva eles permitem que os seres humanos e as máquinas digitais possam interagir de forma mais otimizada e facilitada, pois são capazes de aprender com experiências, com erros e acertos. O resultado disso é que essas novas máquinas digitais são capazes de identificar padrões num grande número de dados e podem prever um variado número de situações (Gomes, 2023, p. 19).

Um dos objetivos do *Machine Learning* é o desenvolvimento de algoritmos que possam ler, compreender e aprender com novos dados além de "determinar respostas dentro de um número finito de possibilidades" (Gomes, 2023, p. 19). Com essas tecnologias, as novas e impressionantes transformações vieram a partir de 2012 com o aprendizado profundo das redes neurais do tipo convolucional.

As técnicas de aprendizado profundo são estruturadas por muitas camadas de redes neurais artificiais. Em 2017 foram desenvolvidas as impressionantes redes denominadas *transformers*. Com essa rede vieram os *large language models* (LLM) ou modelos de linguagem de grande porte. Embora esses modelos sejam incrivelmente eficientes para tradução de textos, correção de sintaxe e gramática, resumo de textos, além de imitar de maneira relativamente sofisticada o raciocínio humano, o seu funcionamento é bastante simples: "[...] emitir os símbolos mais prováveis dado o conjunto de símbolos de entrada" (Costa e Cozman, 2024, p. 139-140). Pode-se citar o ChatGPT como exemplo de LLM.

Vale a pena ressaltar, que apesar de todos os avanços da IA, os seus desenvolvedores sempre se voltam à base antropomórfica, produzindo máquinas que imitam o ser humano, de acordo com o que acreditam serem as "verdadeiras" características humanas. A IA simbólica, por exemplo, tinha por motivação a simulação da mente humana em seus diferentes sistemas especialistas. A IA conexionista, por sua vez, almejava simular o cérebro humano por meio de suas redes neurais.

5 NEM SEMPRE VISTO, MAS SEMPRE PRESENTE: O ANTROPOMORFISMO

Desde o experimento de Turing, passando pelo programa ELIZA, até os *bots* conversacionais como o ChatGPT, o antropomorfismo é o componente que permeia a IA desde o início de seu desenvolvimento. A manutenção dessa característica se mantém pelas décadas de formação e aperfeiçoamento tecnológico da evolução computacional. É óbvio que existem outras categorias de IA, mas a antropomórfica é aquela que mais chama atenção e desperta fascínio. Além disso, o desenvolvimento de máquinas que imitam o nosso modo de agir movimenta um negócio lucrativo para

as grandes corporações. Mas afinal, do que se trata o antropomorfismo, que está tão presente em nossos atos a ponto de, tornando-se parte do nosso modo de agir, nem nos damos conta dele?

Airenti (2015, p. 1) destaca o aspecto relacional do antropomorfismo e esclarece que é um tipo de relação que seres humanos estabelecem com artefatos. Além disso, o que nomeamos como antropomorfismo são formas típicas de interações da comunicação humana estendida aos não-humanos. A esses não-humanos é dada a atribuição de interlocutor em um diálogo possível. Dessa maneira, atribuir ao não-humano ou artefato uma qualidade de interlocução implica lidar com esse ente como alguém dotado de estados mentais.

Para definir o antropomorfismo de uma maneira simples e direta, pode-se dizê-lo como a "atribuição de motivação, características ou comportamento" humanos a objetos inanimados, fenômenos da natureza e animais, dentre outros (Airenti, 2015, p. 4). Atentando-se para uma maior especificidade do termo, o antropomorfismo é uma maneira encontrada para explicar o não-humano utilizando os estados mentais de uma base humana. Pode-se dizer que, para interpretarmos o comportamento não-humano, lidamos com ele por meio da psicologia humana do senso comum ou popular. Por meio desse artifício, os humanos incluem os não-humanos na vida social, falando com eles, repreendendo-os, interagindo, como se quase humanos fossem[28].

Diante do exposto, Airenti (2015, p. 5) especifica um ponto relevante para entender como o antropomorfismo opera e, para isso, utiliza uma analogia para explicar a atribuição de características humanas a objetos, especificamente. Por exemplo, uma pessoa deve chegar pontualmente em uma entrevista de emprego e, no meio do caminho, o seu carro para. Em outra circunstância, o indivíduo necessita chegar a essa entrevista e não sabe se seu carro conseguirá chegar ao destino sem maiores problemas, mas, para a sua surpresa, consegue. No primeiro cenário, possivelmente a pessoa irá insultar o carro ou ainda implorar para que ele entenda a situação: "Por favor, não faça isso comigo justamente hoje!". No segundo exemplo, que é favorável, o dono do carro poderá dizer: "Eu sabia que poderia contar com você!". Nas duas situações, o indivíduo sabe que, apesar de complexo, ele lida com uma máquina, um carro.

Gabriella Airenti parte da premissa que o antropomorfismo é um componente básico da cognição humana e que, independentemente da natureza do objeto, o que importa é a motivação e a situação interativa. O que concretiza a atitude antropomorfizante em relação ao objeto é uma emoção positiva ou negativa desencadeada nessa situação de interação. De acordo com a premissa acima citada, a emoção só poderia ocorrer por meio da interação. Ainda, em situações comuns, o carro é apenas um objeto a ser manipulado. Mas, em situações interativas onde ocorram emoções negativas ou positivas,

[28] Esta situação não ocorre sempre de modo intencional, consciente, pois, na maioria das vezes isso se manifesta impensadamente.

o carro torna-se um parceiro de diálogo, no qual lhe são atribuídas características mentais semelhantes às humanas, como intencionalidade ou motivos (Airenti, 2015, p. 6). Essa situação é descrita por Airetti em relação à ELIZA.

Ao fornecer à IA ELIZA a "personalidade" de um terapeuta rogeriano, Weizenbaum se surpreendeu com o tipo de interação que usuários tinham com o programa. Segundo Weizenbaum (1966, pp. 36-45), ELIZA contava com um limitado número de reações e podia fazer perguntas simples como "Conte-me mais sobre sua mãe", ou repetir as últimas palavras do usuário: "Conte mais sobre [...]". No entanto, os usuários, geralmente estudantes e pesquisadores envolvidos no projeto, eram capazes de interagir longamente com ELIZA, discutir problemas e tratá-la como se fosse humana. Apesar de saberem perfeitamente qual era a real natureza do programa ELIZA, os usuários atribuíam-lhe estados mentais humanos. Mas por que isso acontece com certa frequência entre humanos e máquinas?

O antropomorfismo geralmente é influenciado pelo viés antropocêntrico. O antropocentrismo é a tendência de considerar valiosos apenas os comportamentos do desempenho humano, mesmo que contenham armadilhas de natureza semântica. Por conseguinte, o antropocentrismo considera a mente humana como um referencial para julgar outras mentes ou estados mentais. Outrossim, de acordo com Buckner (2011, p. 50), o antropocentrismo pode ser classificado como metodológico (quando avaliamos as habilidades de entes com base nas habilidades dos animais humanos), avaliativo (quando julgamos o ente interessante ou digno de atenção, apenas se o seu comportamento tiver similaridade ao comportamento humano) e semântico (quando tentamos interpretar a incerteza das coisas, tomando como referencial a nossa própria perspectiva humana).

A criação da IA conversacional se sustenta na tentativa de capturar o que nos é mais essencial. Ou seja, o antropomorfismo da IA passa quase obrigatoriamente pela característica mais intrínseca que nos define como humanos, a linguagem. Uma vez que a nossa capacidade de manipular símbolos finitos para transformá-los em modos infinitos de se expressar, é uma característica de nossa humanidade, toda tentativa de recriar uma máquina que se assemelhe a nós, deve inexoravelmente reter em sua essência o que há de mais essencial em nós, que é a linguagem (Weizenbaum, 1976, p. 184).

A fixação em desenvolver máquinas cada vez mais antropomórficas e que se manifestem no mundo como fazem os humanos, pode também estar ligada ao que Buckner (2023, p. 88) chama de antropofabulação. A antropofabulação é um engano, um erro que combina o antropocentrismo semântico com uma visão hiperbólica do desempenho cognitivo humano. Assim, as avaliações comparativas que fazemos do mundo dizem (enganosamente) que nossas características humanas seriam suficientes para satisfazer certas idealizações irrealistas a que submetemos as coisas do mundo.

Por conseguinte, as atribuições que damos aos entes podem se manifestar como uma distorção do que consideramos prioritário, prejudicando o desenvolvimento das demais tecnologias de IA que não sejam antropomórficas e conversacionais.

O antropomorfismo é um fenômeno que é inerente à nossa atuação no mundo. Ele simplesmente existe por meio de nossa manifestação, queiramos ou não. Em decorrência disso, antropomorfizamos carros, nuvens, animais. Contudo, a IA conversacional possui uma singularidade diversa dos carros, nuvens ou animais. A IA conversacional possui uma característica linguística propositadamente fornecida a ela, a qual identificamos fortemente como um atributo humano essencial (Dennett, 2000, p. 175; Weizenbaum, 1976, p. 184). A camada linguística dada à máquina nos impele à interação. É como falar com um bebê humano, para que ele reaja e nos olhe de volta (Dennett, 2000, p. 27). Estamos a todo momento tentando uma interação consciente, verdadeira, cheia de significado com as coisas que no cercam, para que consigamos extrair algum significado do mundo. Os próprios pesquisadores que participaram do desenvolvimento da IA ELIZA realmente se comportaram como se ela fosse uma terapeuta humana, porque é da nossa natureza antropomorfizar e porque a ela foram propositadamente adicionadas camadas com "características humanas" para se parecer com um terapeuta rogeriano. O que isso pode nos dizer em relação à IA conversacional?

Na perspectiva de Weizenbaum (1976, p. 189), ELIZA criou uma notável ilusão na interação com as pessoas, porque elas acreditavam que a IA realmente as compreendia como um humano faria. Esses indivíduos não estavam sendo enganados, pois sabiam se tratar de uma máquina. Contudo, logo se esqueciam desse fato - como espectadores de uma peça teatral que se esquecem que o desenrolar dos fatos observados não é real. Além disso, constatou-se que tal ilusão era mais fortemente arraigada nas pessoas que tinham um conhecimento superficial sobre computadores e que apesar das explicações de Weizenbaum, os usuários acreditavam que a máquina realmente poderia fornecer insights profundos acerca de suas vidas[29]. Não obstante, as pistas desses insights eram dadas pela própria pessoa.

O sentido e a continuidade conversacional que ELIZA aparentava possuir, eram reforçados pelos significados e interpretações do usuário, confirmando uma hipótese pessoal que a máquina possuía compreensão. O que ELIZA fazia, entretanto, na conversa estruturada pelo "escopo rogeriano", era dar respostas sequencialmente plausíveis dentro de uma lista predeterminada de interações. Com isso a resposta fornecida era logicamente possível e compreensível dento de uma sintaxe. É o que fazemos quando dizemos "caso eu fure este balão com um prego ele irá..." e alguém próximo responde "estourar". Esta é uma construção hipotética estatisticamente plausível e

[29] Entretanto, como o próprio *Weizenbaum* relata, que esse encantamento que ELIZA imprimia, estava presente mesmo em usuários com um alto grau de discernimento como os pesquisadores que participaram do projeto.

formalmente possível, pois a primeira frase fornece pistas para o que virá a seguir (Weizenbaum, 1976, pp. 189-191)[30].

Weizenbaum (1976, p. 202) ressalta que pesquisadores como Roger Schank deram vazão ao sonho científico que move, até hoje, o trabalho envolvendo a IA: uma máquina como modelo do homem, que teria uma infância para aprender como uma criança, sentir o mundo por meio de seu próprio corpo, órgãos e com isso obter conhecimento. E, no entanto, caso isso pudesse ser possível diz Weizenbaum, mesmo observando uma máquina que compreende a nossa linguagem, isso não significa que poderíamos compreender como funciona o processo de aquisição da linguagem que ocorre em nossas mentes humanas.

Por fim, vale dizer que por certo, possuímos uma tendência em acreditar que a IA conversacional compreende contextos, mesmo que saibamos que não é o caso. Que notadamente, no caso da IA ELIZA, os pesquisadores foram fortemente induzidos por suas impressões a antropomofizá-la. Isso, talvez, conduziu aqueles usuários à algumas distorções. A base do conceito dessa tecnologia está na resolução de determinados problemas, então, o objetivo não é transformar a antropomorfização da IA num efeito colateral que enfeitice o usuário. Ao contrário, a antropomorfização pode ser adotada como um caminho possível para entender como as mentes e as coisas funcionam.

6 DENNETT, SEARLE, MENTES E IA

Para Daniel Dennett (2000, pp. 11-14) não podemos ter a certeza sobre as outras mentes, mas a nossa própria mente é algo que conhecemos e temos a certeza de sua existência. Se cada indivíduo sabe introspectivamente a existência apenas da própria mente, não existe no mundo dois indivíduos que conheçam igualmente uma única mente a partir do seu interior. E nenhuma outra coisa que existe no mundo é conhecida dessa forma. Portanto, não podemos ter a certeza de conhecer outras mentes como intimamente conhecemos a nossa. Há, por certo, coisas incognoscíveis.

Convém notar que Dennett (1991, p. 34; 2000, p. 36) rejeitará de forma enfática o dualismo cartesiano. Para ele, a ideia de uma mente que se separa da matéria comum, sendo composta por algum tipo especial de "coisa" não física, misteriosa, merece o descrédito junto com a astrologia e a alquimia. Ainda, não existe um "teatro cartesiano", ponto cerebral de onde a consciência "comanda" ou para onde todos os nossos pontos de informação se dirijam (Dennett, 1991, pp. 106-108). Para Teixeira

[30] *É claro que até as capacidades dedutivas e analíticas de uma criança são infinitamente mais sofisticados que a máquina. Esse exemplo foi citado para apenas fornecer uma pálida analogia para um melhor entendimento da atuação do script que ELIZA seguia.*

(2008, p. 16), Dennett considera a mente humana é uma interpretação dos processos cerebrais manifestados sob a forma de comportamentos.

Dennett (2000, p. 14; p. 18) considera a relevância da linguagem no reconhecimento do "outro". É por meio da linguagem que os humanos falam uns com os outros e compartilham aquilo que está além das capacidades de qualquer outra criatura do planeta: o mundo subjetivo. Pode-se saber por meio da interação com o outro, não apenas qual a melhor estação para plantar ou como fazer redes de pesca, mas também, reconhecer aflições, falar de desejos, lembrar de sensações, manifestar frustrações e inventar hipóteses: "[...] sabemos a partir da conversação que as pessoas são tipicamente capazes de um entendimento muito elevado de si mesmas e dos outros" (Dennett, 2000, pp. 19-21, tradução nossa)[31].

É pela troca intersubjetiva feita pela linguagem que humanos sabem sobre os estados mentais de seus pares. O mesmo não acontece com as outras espécies ou coisas. Uma vez que não podemos conversar com elas, o mundo mental desses entes é para nós um quarto escuro, no qual só podemos tecer suposições baseadas no senso comum. Entretanto, é pela investigação científica que podemos tentar confirmar ou negar tais impressões sobre as coisas que desconhecemos (Dennett, 2000, p. 22).

Diante do que foi dito sobre as mentes, pode-se refletir sobre duas situações. Primeiro, que o ato de falar não é um requisito indispensável para possuir uma mente. Um humano recém-nascido, por exemplo, tem uma mente. Ou ainda, indivíduos com afasias graves, com certeza possuem mentes. Além disso, é possível que alguns animais não-humanos que não têm capacidade linguística, possuam mentes. Deve-se admitir, por isso, uma dificuldade real. Tais mentes poderiam tornar-se terrenos inacessíveis para a investigação científica, já que não podemos investigá-las pelo viés da linguagem (Dennett, 2000, pp. 24-25).

A evolução foi um processo gradativo, na qual nossos ancestrais foram um dia organismos simples e sem mentes. Antes das mentes, vieram os corpos. Sistemas simples evoluíram para organismos mais complexos que possuíam funções autorreguladoras e de autoproteção. Em um lento processo evolutivo, essas "máquinas" microscópicas que não passavam de recipientes de cadeias de DNA, evoluíram e desenvolveram sistemas nervosos simples que funcionavam como pequenos "comutadores" ou órgãos sensoriais primitivos. Por isso, podiam ligar e desligar esses "comutadores" modulando informações para a autopreservação. Pode-se comparar o efeito de ligar e desligar desses comutadores como sendo as primeiras ações intencionais que modulavam informações (Dennett, 2000, pp. 37-39).

[31] "Sin embargo, nosotros ya sabemos por médio de la conversación que las personas son caracteristicamente capaces de um entendimento muy elevado de sí mismas y de las demás" (Dennett, 2000, p. 21, tradução nossa).

Nosso mundo fervilha desses seres com suas intenções particulares, como serviçais irredutíveis, sempre perseguindo obcecadamente seus objetivos, constituindo os corpos complexos ou as partes dos corpos. Portanto, Dennett (2000, p. 26; pp. 29-30) deseja saber como podemos entender os entes que possuem mentes, mas não possuem uma capacidade linguística para exteriorizar seus pensamentos. A esse universo de entes, simples ou complexos, naturais ou construídos pelo homem, Dennett (2000, p. 39) dará o nome de sistemas intencionais.

A postura intencional é a interpretação que se pode fazer, de qualquer entidade existente, tratando-a como agente racional capaz de ter ações intencionais, sob a perspectiva de suas "crenças" e "desejos"[32]. As posturas intencionais de Dennett utilizam a antropomorfização de coisas vivas ou inanimadas para poder fazer uma predição de suas ações. Existe aqui um estranhamento justificado em relação à postura intencional de atribuir características humanas a entes, inclusive inanimados, mas que Dennett vai esclarecer como sendo "a atitude ou perspectiva que rotineiramente adotamos em relação uns aos outros, portanto adotar a postura intencional em relação à alguma outra coisa parece ser a antropomorfização desta coisa" (Dennett, 2000, p. 40, tradução nossa)[33].

Entretanto, existe o perigo, no uso dessa estratégia: de um lado ser atraído para a metáfora vazia, de outro lado, cair na falsidade literal. Caso seja bem compreendida e bem utilizada, ressalta Dennett (2000, pp. 40-41), pode-se ter uma base segura para a investigação dos fenômenos para explicar e predizer suas ações e movimentos. Existem ainda, mais duas posturas mais simples de predição. A primeira é a postura física se utiliza do conhecimento padrão que temos do mundo físico para construir a predição, uma vez que toda coisa material, viva ou inanimada, está sujeita às leis da física: a água aquecida a cem graus, irá entrar em ebulição e irá congelar a zero grau. Seres humanos ou pedras sofrem os mesmos efeitos da gravidade. Essa é a única estratégia possível para coisas não-vivas e que não sejam artefatos. Já a postura de planejamento conta com um modelo mais sofisticado de predição que utilizamos a todo momento e refina a postura física, porém é mais arriscada (Dennett, 2000, p. 41).

Suponhamos que alguém faça uso de uma IA conversacional para fazer contas simples de aritmética. O programa foi projetado com esse objetivo. O usuário não precisa saber programação, algoritmos, scripts ou leis da física para ter certeza que ela fará contas quando precisar e assim, cumprir a função para que foi planejada. O usuário confia plenamente no programa, a tal ponto, de utilizá-lo para questões relevantes, como resolver provas ou testes (Dennett, 2000, pp. 41-43). Os humanos

[32] Dennett quer se distanciar da concepção do senso comum e da psicologia do cotidiano para "crença" e "desejo" que utilizamos para discutir a vida mental dos humanos (Dennett, 2000, p. 40).
[33] "El enfoque intencional es la actitud o la perspectiva que adoptamos ordinariamente unos com otros, de modo que adoptar el enfoque intencional em relación a otras cosas parece um modo deliberado de *antropizar* la cuestión" (Dennett, 2000, p. 40).

utilizam esse método de predição em suas vidas a todo momento. Usamos o metrô, trens, carros sempre confiando que seguirão a função que lhe fora atribuída e não nos matarão devido a algum defeito. Também nossos ancestrais confiavam que as sementes iriam germinar, crescer e gerar frutos, caso eu as plantasse de uma forma determinada e tivesse cuidados específicos (Dennett, 2000, p. 43).

A postura intencional é derivada da postura de planejamento. Digamos que um indivíduo seja dono de um despertador e a partir de agora o filósofo tratará o artefato como uma pessoa, como um agente. Desse modo, essa pessoa dará ordens para que o despertador a acorde em um horário específico e que, por sua constituição interna, o despertador cumprirá. Na hora correta, o despertador motivado pela ordem recebida e pela promessa feita, emitirá um som para obedientemente acordar seu dono. Esse antropomorfismo hiperbólico pode parecer surreal, mas serve para explicar à uma criança, por exemplo, como coisas funcionam (Dennett, 2000, pp. 43-45). Neste exemplo a postura intencional parece ser de pouco valor, mas adquire outros contornos com entes mais complexos, como a IA conversacional.

Para ilustrar essa teoria, Dennett (1971, p. 87) utiliza o exemplo de um jogo de xadrez contra um computador ou programa. Caso o jogador trate o computador como um adversário cujas táticas nada se saiba, a vitória será mais difícil. A única opção plausível é, à vista disso, agir como se o computador soubesse o que está fazendo e que desejasse ganhar a partida de xadrez. Agir dessa maneira torna minimamente possível prever a maneira que o computador irá atuar no jogo. Não importa que seja um super computador ou um simples laptop, pois todos os comandos do algoritmo que o computador segue, "desejam" vencer a partida: "pense neles como agentes racionais que querem ganhar e sabem as regras o os princípios do xadrez e as posições das peças no tabuleiro"[34] (Dennett, 2000, p. 44, tradução nossa).

Mesmo que o computador esteja em uma posição complicada a ponto de perder o jogo, infere-se que ainda assim a predição da postura intencional seria eficaz. Por que? Porque o computador foi planejado para se comportar como um agente racional. Os seja, podemos considerar que em última análise, as coisas vivas e artefatos construídas pelo homem, têm origem em um objetivo comum de autopreservação. Qualquer agente racional tem como característica primeira e racional buscar o próprio bem, da maneira como indivíduos humanos buscam (Dennett, 2000, p. 45).

Por conseguinte, diz-se sistemas intencionais, os entes, no qual o comportamento pode ser previsto e explicado pela postura intencional, supondo, é claro, que tal ente seja um agente inteligente e racional, uma vez que agentes estúpidos e irracionais, fariam coisas inesperadas e inexplicáveis. Supor que tal agente seguirá um padrão de racionalidade, nos fornece uma imensa vantagem preditiva, uma

[34] "piénsese en ellos como em agentes racionales que quieren ganhar y que saben las reglas e los princípios del ajedrez y las posiciones de las piezas em el tablero" (Dennett, 2000, p. 44).

vez que, já de início, possuímos uma hipótese para tentar prever o desconhecido. Os sistemas intencionais possuem o que os filósofos chamam de intencionalidade. O termo intencionalidade possui dois sentidos, que embora simples, causam alguns equívocos. O primeiro sentido vem do senso comum, ordinário, que está relacionado com propósito, deliberação ou intenção de fazer algo. A intencionalidade no sentido filosófico "[...] é apenas relacionalidade. Alguma coisa exibe intencionalidade se sua competência é de algum modo sobre alguma outra coisa" (Dennett, 2000, p. 48).

Convém lembrar que os conceitos dennettianos concebem o mental como "construção teórica a partir de termos psicológicos que, como ficções úteis, tornam os comportamentos complexos inteligíveis sejam eles de humanos ou de dispositivos artificiais" (Teixeira, 2008, p. 13). Os seus conceitos dialogam com os fenômenos mentais a partir da perspectiva científica e do senso comum. Se por um lado Dennett acomoda sua pesquisa nas bases científicas, ele também se utiliza do nosso senso comum da descrição do mental, da teoria habitual que humanos possuem para explicar os comportamentos das coisas do mundo, recorrendo às ideias comuns como intenção, crença e desejo (Teixeira, 2008, p. 13).

Para Teixeira (2008, p. 21) os sistemas intencionais e todos os seus elementos derivam da folk psychology[35] ou psicologia popular e se apresenta como um mecanismo de nossa capacidade limitada e ignorância para prever o comportamento de organismos e dispositivos complexos. A chave seria a complexidade da folk psychology comportamental do animal, robô, organismo, artefato. Caso o seu comportamento fosse tão complexo a ponto de ser necessário lançar mão de crenças, intenções e desejos para contar uma história de suas ações ou construir explicação de suas reações, a esses entes poderiam ser atribuídos uma vida mental.

A folk psychology dennettiana, é uma estratégia que se baseia nas capacidades humanas de lembranças, conhecimentos e crenças, para relacionalmente predizer o comportamento de coisas desconhecidas. Poderá a IA conversacional prever o nosso comportamento? Sim, já que ela se baseia em nas pistas que nós mesmos fornecemos. Dessa forma, a máquina ganharia no xadrez. A solução seria elaborar um modelo mental, do modelo mental, do modelo mental e assim por diante. Isso significa que, a inteligência é a capacidade do indivíduo de prever a previsão do outro, baseando os comportamentos futuros no resultado dessa previsão. Contudo, existe a possibilidade de ações

[35] A *folk psychology* é uma expressão cunhada por Dennett em 1981 e designa uma teoria habitual que todos nós possuímos na qual explicamos os comportamentos de outros seres humanos recorrendo às ideias comuns de "intenção", "crença", "desejo" e outros termos que compõem o chamado vocabulário mentalista. Dessa forma, para Dennett, tudo que pudesse ser descrito como um sistema intencional, seria possuidor de uma mente. (Teixeira, 2008).

preditivas equivocadas, uma vez que essas predições são como uma espécie de aposta (Teixeira, 2008, p. 21).

Um importante conceito de Daniel Dennett são os chamados "termos intencionais". Foi por meio da matematização que a física desenvolveu as equações matemáticas da mecânica clássica para descrever e predizer com certa acurácia o movimento dos corpos. Obter uma descrição e predição exata da ação humana quando envolve estados mentais, não é possível. É para esse estado de "imprecisão" da ação humana que Dennett chama atenção. Para ele não seria possível reduzir a ação humana repleta de termos mentais a uma descrição física com referentes definidos (Teixeira, 2008, p. 10).

Termos não-intencionais (além da lógica que rege as ciências naturais) diz-se daqueles que são caracterizados pela extensão[36]. Termos intencionais se caracterizam pela sua intensão com "s" e se referem ao significado de cada elemento tomado individualmente. Tomemos como exemplo para ilustrar a intensão, a história de Édipo Rei de Sófocles, conforme descrito por Teixeira (2008, p. 011), onde temos duas proposições: A) Édipo casa-se com Jocasta; B) Édipo deseja que Jocasta se torne sua esposa. Na proposição A se substituímos o termo "Jocasta" pelo termo "mãe de Édipo" a nova frase teria seu valor de verdade preservado: A') Édipo casa-se com a mãe de Édipo. Entretanto, caso substitua-se na proposição B "Jocasta" pelo termo "mãe de Édipo", ou seja, B') Édipo deseja que a mãe de Édipo se torne sua esposa; B e B' não teriam o mesmo valor de verdade, pois Édipo jamais desejou se casar com a própria mãe. Para Dennett (2000, p. 53) o termo intensionalidade com "s" é o causador de certos enganos[37]. É exclusivo das linguagens, não se referindo a qualquer outro sistema representativo como gráficos, mapas, quadros. Quando se faz referência de certas coisas utilizando-se denominações diferentes ("sol", "astro-rei", "estrela do sistema solar"), cada palavra pode adquirir um significado diferente dependendo do contexto. Essa vagueza é mais rara nas ciências naturais mesmo que possamos observar seus fenômenos apenas a priori (Teixeira, 2008, p. 12).

O filósofo John Searle, também dedicou grande parte de suas investigações para investigar fenômenos como a consciência e a intencionalidade. Para Searle (2002, p. 1) intencionalidade é a propriedade de atribuir estados e eventos mentais a objetos e estados de coisas no mundo. Intencionalidade é a característica da direcionalidade. Um indivíduo que tem uma crença, deve ser uma crença desse ou daquele modo. Se tiver medo, deve ser medo de algo ou acontecimento. São sempre referentes a algo, pois se é um estado com intencionalidade deve sempre ter uma resposta para perguntas como: A que se refere X? Em que consiste X?

[36] A extensão de um termo é a classe das coisas às quais o termo se refere (Teixeira, 2008, pp. 10-11).
[37] A lógica clássica utiliza o termo "compreensão" (Dennett, 2000, p. 53).

Por conseguinte, Searle (2000, pp. 89-90) esclarece a distinção entre intencionalidade intrínseca, derivada e aparente (que na verdade não é um terceiro tipo). Tome-se por exemplo as seguintes frases: a) Meu cachorro está com medo; b) *My dog is scared*; c) O motor do meu carro tem medo de água. Nos três exemplos temos o mesmo fenômeno intencional, o medo, com condições diferentes. No primeiro exemplo, o meu cachorro tem intrinsecamente a intencionalidade que lhe é atribuída. Ela existe. Meu cachorro tem fome independentemente de os observadores acreditarem ou não. Seres humanos e animais possuem essa forma intrínseca de intencionalidade. A segunda frase possui a intencionalidade derivada da anterior e depende do observador. Os observadores devem saber inglês para a frase fazer sentido. A intencionalidade derivada vem das palavras, diagramas, imagens e é essencialmente linguística (Searle, 2000, p. 90).

Consequentemente, as duas primeiras são intencionalidades literais, diferentes da intencionalidade aparente, que é metafórica. Quando digo "o motor do meu carro tem medo de água" eu atribuo uma característica relacional humana que não pode ocorrer com aquele artefato. A intencionalidade aparente apenas se comporta como intencionalidade, quando na verdade não é (Searle, 2000, p. 90). Essa atribuição de crenças e desejos da intencionalidade aparente, pode ser observada nos fenômenos antropomórficos descritos por Weizenbaum e parece também fazer correspondência a estrutura antropomórfica pertencente à cognição humana descrita por Gabriella Airenti.

Searle (2002, p. 32) também chama a atenção para uma confusão recorrente no universo filosófico[38]. Comumente confunde-se "intencionalidade" com "intensionalidade" com "s", sendo que o conceito dessas duas palavras não é nem remotamente parecido. Intencionalidade é a capacidade da mente ou cérebro de representar outros objetos. A intensionalidade é a incapacidade de algumas sentenças, enunciados e outras entidades linguísticas de satisfazer as condições de verdade dos testes lógicos de extensionalidade. Fenômenos intensionais pertencem ao universo da filosofia da linguagem e à lógica; e referem-se aos contextos intencionais nos quais as substituições na frase por termos de igual valor pode alterar o sentido da proposição.

O aspecto basilar, primeiro e mais importante das mentes é a consciência que pode ser definida como os estados de conhecimento ou percepção que nos acompanham desde que acordamos até o momento que novamente vamos dormir. A consciência tem fim com a morte ou qualquer outro modo que nos coloque em estados inconscientes. Com grande número de formas e variedades, as características mais importantes e comuns a todos os estados conscientes: eles são internos, são qualitativos e a são subjetivos (Searle, 2000, pp. 45-46).

[38] Daniel Dennett, como visto anteriormente, também se ocupou dessa distinção.

Segundo Searle (2002, pp. 2-3), a intencionalidade e consciência são fenômenos diferentes. Muitos estados conscientes não possuem intencionalidade, como a euforia ou os medos infundados. Existem, ainda estados intencionais que não são conscientes, pois nunca foram pensados. Temos muitas crenças pois nunca, até o momento, como as crenças que nunca foram pensadas conscientemente até o momento: penso e creio que minha avó nunca pôde viajar para a Holanda.

Para investigar a possiblidade de máquinas possuírem intencionalidade, Searle (1980, p. 417) propõe um sistema geral que explique tanto a proposta computacional de Roger Schank, quanto todos os demais experimentos de "fenômenos mentais humanos" que se baseiem no experimento de Turing: "mesmos argumentos se aplicariam ao SHRDLU de Winograd (Winograd 1973), ao ELIZA de Weizenbaum (Weizenbaum, 1966) e [...] qualquer simulação de fenômenos mentais humanos baseada na máquina de Turing"[39] (Searle, 1980, p. 417, tradução nossa). As objeções que Searle faz não são direcionadas à IA fraca, mas à IA no sentido forte. Para desenvolver um conceito que verifique a singularidade na IA, Searle propõe um experimento mental que seja geral e abrangente. Inicialmente, Searle (1980, p. 417) utiliza o experimento de Roger Schank[40] de 1977, e o *Gedankenexperiment*[41] como método investigativo. Nesse contexto, Searle emprega esforços para refutar o Teste de Turing como ferramenta para validar a presença de uma inteligência semelhante à humana na IA. Com cautela, Turing sugeriria que as representações mentais estariam reduzidas a símbolos com características semânticas e sintáticas, nos quais o pensamento racional e inteligente seria apenas uma reordenação de

[39] "The same arguments would apply to Winograd's SHRDLU (Winograd 1973), Weizenbaum's ELIZA (Weizenbaum 1965), and indeed to any Turing machine simulation of human mental phenomena" (Searle, 1980, p. 417).

[40] O experimento de Roger Schank era composto por softwares desenvolvidos na linguagem LISP e auxiliados por scripts que ajudavam na implementação de uma IA que podia compreender e interpretar narrativas. Os primeiros modelos estão descritos no seu livro "*Conceptual Information Processing*", de 1975. Já no livro "*Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures*", de 1977, Schank detalha como programas como SAM utilizavam seus scripts na interpretação de histórias. Pode-se compreender a intenção do programa de Schank com a seguinte história: Numa primeira situação, um homem entra em um restaurante e pede um hambúrguer, que vem queimado. O homem, irritado, sai do restaurante. Humanos, se indagados, certamente responderiam que o homem não comeu o hambúrguer. Em uma segunda história um homem entra em um restaurante e pede um hambúrguer que vem perfeito. O homem fica satisfeito, paga e vai embora. Desta vez, possivelmente, humanos responderiam que o homem comeu o hambúrguer. Os humanos, ao ouvirem uma história, são capazes de fornecer respostas válidas mesmo se algumas informações da história estiverem implícitas e exigirem uma inferência e capacidade dedutiva para compreendê-las. O objetivo do programa de Schank seria de compreender histórias de restaurantes e hambúrgueres como humanos fazem, desde que o programa fosse alimentado com dados sobre hambúrgueres e restaurantes da mesma forma que esses dados são fornecidos aos humanos. A hipótese de Schank afirmava que essas informações fornecidas são parecidas com aquelas que humanos utilizam para se comunicar (Searle, 1980, p. 417).

[41] *Gedankenexperiment* é um termo alemão que quer dizer "experimento mental". É um recurso filosófico, ou seja, uma situação imaginária, mas possível, que não contraria as possibilidades físicas, lógicas, etc. Útil para extrair conclusões conceituais relevantes.

símbolos (Gomes, 2023, p. 48). Searle com o Argumento do Quarto Chinês pretende refutar esses conceitos.

Usando um *Gedankenexperiment* bastante esclarecedor, Searle (1980, p. 417) propõe o seguinte cenário: um indivíduo trancado em um quarto recebe três grandes pilhas ou lotes de papel. Esse é o *Argumento do Quarto Chinês*. A pessoa não entende chinês escrito ou falado e é falante apenas de inglês. Para ela, o chinês escrito é como um aglomerado de rabiscos sem sentido. A primeira pilha é composta por textos em chinês. Agora, a pessoa recebe um segundo lote de textos em chinês, acompanhado de um conjunto de regras em inglês que ela compreende. Essas instruções permitem à pessoa relacionar o primeiro conjunto de símbolos formais em chinês com o segundo. Também é fornecido um terceiro lote em chinês. Ele vem junto com instruções em inglês, o que possibilita que a pessoa o relacione esse terceiro conjunto de texto (apenas pela forma dos ideogramas chineses) aos dois primeiros.

As pessoas que fornecem os textos, chamam o primeiro lote de "roteiro"; o segundo é chamado de "história"; e o terceiro é chamado de "perguntas". Os símbolos que a pessoa devolve em resposta às "perguntas" são chamadas de "respostas às perguntas", e as instruções em inglês são chamadas de "programas". Munida de todos esses dados, a pessoa do quarto é capaz de responder às perguntas, relacionando os símbolos em chinês que, para ela, são apenas formas gráficas sem sentido.

As pessoas de fora do quarto fornecem histórias em inglês e também fazem perguntas em inglês à pessoa do quarto, que compreende tudo (pois é falante em inglês) e responde em inglês – esta é uma situação relevante que Searle utilizará como elemento comparativo. Além disso, com o decorrer do experimento, a pessoa no quarto fica cada vez mais familiarizada em manipular os símbolos chineses. As pessoas de fora do quarto também se tornam mais hábeis em escrever os programas. Para uma pessoa externa ao experimento, não há diferença de qualidade entre as respostas fornecidas em inglês ou chinês. Ou seja, para uma pessoa que desconheça o processo, pode parecer que o indivíduo no quarto fala e compreende tanto chinês quanto inglês.

A partir desse experimento mental, Searle (1980, p. 418) infere duas conclusões. A primeira é que, da mesma maneira que a pessoa no quarto não compreende chinês, a máquina de Schank, pelos mesmos motivos, também não é capaz de verdadeiramente compreender histórias. Isso quer dizer que não há compreensão verdadeira, nem para o computador, nem para o indivíduo que fornece respostas em chinês, mesmo que pareça que há. O segundo ponto que Searle destaca é que programas de computador não explicam o modo como a compreensão humana funciona. Este ponto deve ser melhor explicado.

Segundo Searle (1980, p. 418), os defensores da IA forte acreditam que nos dois casos em que a pessoa no quarto fornece respostas, em inglês ou chinês, existe compreensão[42]. Segundo esta linha de pensamento, no caso das respostas em inglês (que a pessoa compreende), haveria um número muito maior de manipulação de símbolos formais do que nas respostas em chinês, que a pessoa não compreende. Por isso, ao concordar com essa premissa, assume-se a ideia que a chave para que a máquina possua compreensão passa pelo número de símbolos formais que ela pode manipular. Ou seja, hipoteticamente a chave para a compreensão das linguagens e suas peculiaridades narrativas passaria pelo poder de processamento da máquina, pois o próprio cérebro humano seria um programa instanciado.

À vista disso, caso conseguíssemos desenvolver uma IA com o mesmo poder de processamento do cérebro humano ela teria a capacidade de nos informar como se daria o processo da compreensão humana. Searle refuta essa ideia, pois segundo ele, a compreensão não passa pela quantidade de manipulação de informações e dados puramente formais, ou ainda, "[...] por mais que se coloque no computador princípios formais isto não será suficiente para a compreensão" porque "[...] não há nenhuma razão para supor que quando eu [...]" humano, "[...] compreendo inglês, eu estou operando com algum programa formal" (Searle, 1980, p. 418).

A grande indagação de Searle ao apresentar o Argumento do Quarto Chinês é: o fato do indivíduo dentro quarto haver respondido perguntas em chinês implica que ele compreende e fala chinês como se fosse um humano chinês? Segundo Searle (1980, pp. 417-424), não, pois o quarto chinês apenas seguiu instruções. Ainda, o Argumento do Quarto Chinês infere que regras sintáticas não poderiam gerar habilidades semânticas. Diante disso, Searle utiliza o Argumento do Quarto Chinês para criticar o Computacionalismo, o Behaviorismo filosófico, o Funcionalismo e o Cognitivismo (Gomes, 2023, pp. 43-49).

A posição de Searle não é, por certo, uma posição dualista. O argumento de Searle, segundo Viana (2013, p. 74) se ancora em três premissas a saber: 1) Que programas são sintáticos; 2) Mentes possuem conteúdos semânticos; e 3) Sintaxe não é análogo à semântica, nem é suficiente para produzir conteúdo semântico. Para haver uma negação do argumento seria necessário negar a segunda premissa, negando com isso que estados mentais possuam uma semântica intrínseca a eles e também os próprios conteúdos mentais.

Dennett, aceita os conteúdos mentais, mas nega a ideia de que seriam subjetivos e privados, elementos que para Searle constituem a essência desses fenômenos. Para Searle, o problema contido nos conceitos Dennettianos é que experiências subjetivas são consideradas uma ilusão, negando-se a

[42] Gomes (2023) utiliza o termo "compreensão" como a posse de estados mentais intencionais e que, doravante, também utilizaremos neste trabalho.

experiencia privada e interna em primeira pessoa, indo contra a intuição do senso comum. Em outras palavras, Dennett critica o dado observado ao invés de simplesmente explicá-lo e defende "uma teoria da consciência que *não salva o fenômeno*" (Viana, 2013, p. 75).

Se, por exemplo, nos ferimos, isso causa uma sensação desagradável chamada de dor e sobre isso não existe qualquer dúvida, ilusão ou engano. E mesmo que o fosse, ainda assim seria, para Searle, uma experiência subjetiva, interna e privada impossibilitando-a de ser traduzida apenas em estímulos eletroquímicos. Ao pensar dessa forma, Dennett reduziria o indivíduo a um simples zumbi, já que nega o fato comum das experiências e percepções subjetivas (Viana, 2013, p. 75).

A posição de Searle não é dualista, já que para ele estados mentais e estados neuronais são coisas distintas. Sua posição pode ser definida como fisicalista não-reducionista, na qual, os estados mentais são causados por estados físicos. Daí decorre, que Searle não descarta a IA forte, mas a aceita com cautela, pois somente com o mesmo poder causal do cérebro humano, poderíamos ter uma IA capaz de gerar mundos subjetivos e privados[43]. O problema que daí decorre é a dificuldade de definição sobre quais seriam os poderes causais de um cérebro capaz de gerar uma IA forte. Além disso o Naturalismo Biológico de Searle, precisa esclarecer como estados neuronais causariam os estados mentais, sem reduzir os segundos aos primeiros (Viana, 2013, pp. 77-78).

CONSIDERAÇÕES FINAIS

A IA conversacional como qualquer tecnologia revolucionária vem revestida de uma capacidade única de causar assombro, espanto e, portanto, uma certa desorientação para investigá-la. Além disso, quando se emprega o antropomorfismo como forma de escrutinar a IA, percebe-se algumas particularidades. O antropomorfismo é um fenômeno peculiar que colocou a IA conversacional desde a sua origem em uma esfera do quase-humano, transformando-a em um dos artefatos produzidos pela humanidade que talvez mais se pareça conosco linguisticamente.

Dito isso, é necessário esclarecer que o trabalho aqui desenvolvido não avalia o antropomorfismo como um fenômeno que distorce a realidade do mundo ou como característica que conduz o indivíduo ao erro e ao engano. O antropomorfismo é mais que isso. Ele existe como uma característica que atribuímos às coisas, mas também como uma estratégia humana de organização para

[43] Essa teoria da consciência de Searle é chamada de "Naturalismo Biológico". A palavra "naturalismo" faz referência a consciência que faria parte do mundo natural da mesma forma que a mitose, o peristaltismo, digestão, fotossíntese. "Biológico" se refere à consciência como fenômeno restrito aos níveis biológicos de animais desenvolvidos, capazes de produzirem estados mentais "subjetivos, qualitativos e intencionais" (Viana, 2013, p. 76).

que o mundo faça sentido para nós. Assim, em quais categorias da IA o antropomorfismo poderia existir?

Em que se pesem as características da IA forte, a antropomorfização jamais poderia cooptar tal categoria de IA. Primeiramente, o próprio conceito de antropomorfismo invalida a IA forte como possuidora de característica antropomórficas, visto que seu significado indica uma ação relacional em que atribuímos características humanas aos entes. A IA forte teria em tese consciência e mente com estados intencionais, desejos e capacidades subjetivas humanas. Assim como não podemos atribuir características humanas à humanos, também não poderíamos atribuir características humanas à maquinas com mentes humanas, como no caso da IA forte.

Dennet e Searle tem respostas sobre a possibilidade de uma IA forte e como isso interfere na aplicação do antropomorfismo na IA conversacional. Tanto Dennett quanto Searle (embora Searle com um grau maior de ceticismo), admitem a possibilidade do advento de uma IA consciente e com estados subjetivos e privados. A diferença entre os dois, é que para Searle, esse fato só é possível caso a máquina possa ser capaz de produzir estados internos subjetivos, o que segundo ele é impossível devido ao problema sintaxe-semântica, como foi demonstrado no experimento do quarto chinês. Ou seja, Searle responderia à questão da existência do antropomorfismo de maneira residual, onde ele habitaria apenas e tão somente a IA fraca e nunca a IA forte.

Dennett reduzindo a mente à racionalidade e extraindo dela a subjetividade, também é favorável à possibilidade de criação de uma IA forte. Contudo, enquanto Searle define o fenômeno da consciência como consequência do cérebro humano, que elabora uma perspectiva subjetiva de si próprio, Dennett aceita o fenômeno de uma consciência como resultado do agrupamento complexo de neurônios, lapidado pelo processo evolutivo, sendo por isso, reproduzível artificialmente.

Searle infere que o cérebro humano constrói uma perspectiva subjetiva em primeira pessoa, mas que não pode ser reproduzida por uma máquina nem simplesmente reduzida a uma mistura de processos eletroquímicos. Searle, em um primeiro momento, não reduz os estados mentais a processos físico-químicos uma vez que não derivam apenas de uma sintaxe, mas de processos semanticamente intencionais. Máquinas simbólicas (baseadas na sintaxe) não reproduzem o funcionamento do cérebro, porque ele é mais do que símbolos (Searle 1980). Mas esse "mais que símbolos" é justamente a sua conexão biológica e sua estrutura particular[44]. Enquanto Searle procura um traço de intencionalidade

[44] No artigo, há uma objeção direcionada a Searle (IV – A objeção da combinação) que um conexionista poderia fazer hoje: de que estamos reproduzindo o cérebro, não a mente. Para Searle, se um dia conseguíssemos reproduzir ponto a ponto o cérebro humano, então essa coisa reproduzida seria outro cérebro humano, não seria máquina, o que tornaria a discussão toda inútil. Nesse caso, embora estados mentais não sejam biológicos (porque não possuem uma localização biológica), eles possuem uma ligação com certas configurações biológicas do cérebro (Searle 1980, pp. 7-9).

in loco, em outras palavras, na própria configuração biológica do cérebro humano, Dennett o procura relacionalmente nos indivíduos: a intencionalidade é uma propriedade atribuída funcionalmente por indivíduos para predizer comportamentos de entidades. Essa é uma distinção relevante porque fornece diferentes respostas ao problema do antropomorfismo.

A antropomorfização como característica da IA não é boa ou ruim. Ela adquire um caráter de engano quando à ela se aderem a antropofabulação e o antropocentrismo. O seu efeito negativo poderia se manifestar como um modo de produção em massa que priorizaria pesquisas e produtos em uma IA que apresentasse características humanas cada vez mais aperfeiçoadas. Mas esse tipo de IA também deixa marcas sutis no usuário quando a ela é atribuída propositadamente o antropomorfismo com a intenção de gerar certos estados psicológicos. Entretanto, o usuário não é enganado. Desde os primórdios dos experimentos com IA, foram-lhes designadas características humanas. Seus usuários, com exceção dos que participaram do Teste de Turing, tinham a plena consciência de estarem interagindo com uma máquina, como ocorria com a IA ELIZA. Dessa forma, mesmo que a IA conversacional não se passasse por humano, ainda assim, ela induziria estados mentais no usuário.

Dennett por meio das posturas intencionais, as quais atribuem mentes e estados subjetivos na IA conversacional, nos diz que o antropomorfismo é um meio de ordenação do universo que nos cerca. Ordenação dos artefatos, dos seres vivos ou da IA conversacional. O próprio Searle (2000) percebe as características de atribuição antropomórficas como um fenômeno intencional do tipo aparente. Em outras palavras, o antropomorfismo seria uma maneira metafórica de atribuir características relacionais humanas às coisas. Não seria um engano, mas apenas mais uma das muitas nuances do ser humano.

Assim, se Buckner (2011) nos adverte sobre a antropofabulação e o antropocentrismo, Airetti (2015) nos apresenta o antropomorfismo como um fenômeno ancestral que está presente em nossa forma de ver o mundo e nos relacionarmos com ele. Os estados intencionais dennettianos que atribuímos a um artefato tão singular como a IA conversacional – que se apresenta como o mais antropomorfizado linguisticamente dentre todos os outros que nos cercam – é uma estratégia satisfatória que nos permite ter uma vantagem preditiva sobre aquilo que se apresenta como uma criatura quase desconhecida e intrigante. O uso de uma técnica tão humanamente sofisticada, baseada no senso comum da observação comportamental e utilizada para extrair informações que nos permite predizer comportamentos, por si só, já é um fato extraordinário. Por fim, poderíamos imaginar que, talvez a ação humana de antropomorfizar, tenha surgido com a primeira fagulha de nossas mentes junto com o descobrimento de nosso mundo interior, com o "eu" e com a noção de privado, para só depois designarmos as coisas exteriores a partir de nossa própria subjetividade. E isso sim, causa espanto.

Artigo recebido em: 20/01/2025
Artigo aceito em: 30/03/2025
Artigo publicado em: 31/03/2025

REFERÊNCIAS

AIRENTI, Gabriella. The Cognitive Bases of Anthropomorphism: From Relatedness to Empathy. *International Journal of Social Robotics*, v. 7, n. 1, p. 117–127, 2015. DOI: https://doi.org/10.1007/s12369-014-0263-x.

BASSETT, Caroline. The computational therapeutic: exploring Weizenbaum's ELIZA as a history of the present. *AI & SOCIETY*, v. 34, n. 4, p. 803–812, 2019. DOI: https://doi.org/10.1007/s00146-018-0825-9.

BUCKNER, Cameron J. *From Deep Learning to Rational Machines*: what the history of philosophy can teach us about the future of artificial intelligence. 1. ed. Oxford University, Press New York, 2023.

BUCKNER, Cameron J. Two approaches to the distinction between cognition and mere association. *International Journal for Comparative Psychology*, v. 24, n. 1, p. 1–35, 2011. DOI: https://doi.org/10.46867/ijcp.2011.24.04.06.

COPELAND, Jack. The Turing Test. *Minds and Machines*, v. 10, 2000a, p. 519-539. DOI: https://doi.org/10.1023/A:1011285919106.

COSTA, A. H. R.; COZMAN, F. G. O futuro da pesquisa em inteligência artificial. *Revista USP*, n. 141, p. 133–146, 2024. DOI: https://doi.org/10.11606/issn.2316-9036.i141p133-146.

COZMAN, Fabio G. No canal da Inteligência Artificial – Nova temporada de desgrenhados e empertigados. *Estudos Avançados - USP*, v. 35, n. 101, p. 7–20, 2021. DOI: https://doi.org/10.1590/s0103-4014.2021.35101.002.

DENNETT, Daniel. C. *Consciousness Explained*. Boston: Little, Brown, and London: Allen Lane, 1991.

DENNETT, Daniel. C. Intentional Systems. *Journal of Philosophy*, v. 68, n. 4, p. 87–106, 1971. DOI: https://doi.org/10.2307/2025382.

DENNETT, Daniel. C. *Tipos de Mentes*: hacia uma comprensión de la conciencia. Trad: Francisco Páez de la Cadena. Madrid: Editorial Debate, 2000.

GOMES, Victor P. *Revisitando o teste de Turing*: análises e consequências. 2023. 109 f. Tese de doutorado - Universidade Federal do Rio Grande do Norte, Centro de Ciências Humanas, Letras e Artes, Programa de Pós-Graduação em Filosofia, Natal, 2023. Disponível em: https://repositorio.ufrn.br/handle/123456789/57607.

HAUGELAND, John. *Artificial Intelligence*: The Very Idea. A Bradford Book. The MIT Press: Cambridge, Massachusetts, London, England, 1985.

HAYMOND, S.; MCCUDDEN, C. Rise of the Machines: Artificial Intelligence and the Clinical Laboratory. *The Journal of Applied Laboratory Medicine*, v. 6, n. 6, p. 1640–1654, 2021. DOI: https://doi.org/10.1093/jalm/jfab075.

HUAWEI TECHNOLOGIES COMPANY, LTD. *Artificial Intelligence Technology*. Singapore: Springer Nature Singapore, 2023.

KAUFMAN, Dora. *A inteligência artificial irá suplantar a inteligência humana?* Barueri: Estação das Letras e Cores, 2019.

MCCARTHY, John. What is artificial intelligence? *Computer Science Department*, Stanford University, p. 13, 2004.

SEARLE, John R. *Intencionalidade*. Trad. Júlio Fischer. 2a ed. São Paulo: Martins Fontes, 2002.

SEARLE, John R. *Mente, linguagem e sociedade*: filosofia no mundo real. Trad. F, Rangel. Rio de Janeiro: Rocco, 2000.

SEARLE, John R. Minds, brains, and programs. *Behavioral and Brain Sciences*, v. 3, n. 3, p. 417–424, 1980. DOI: https://doi.org/10.1017/S0140525X00005756.

SHETTLEWORTH, Sara J. *Cognition, communication, and behavior*. 2. ed. Oxford, New York, 2010.
TEIXEIRA, João de F. *A mente segundo Dennett*. São Paulo: Perspectiva, 2008.
TURING, Alan M. Computing, machinery and intelligence. *Mind*, v. LIX, n. 236, p. 433–460, 1950. DOI: https://doi.org/10.1093/mind/LIX.236.433.
VIANA, W. C. Técnica e Inteligência Artificial: O debate entre J. Searle e D. Dennett. *Pensando - Revista de Filosofia*, v. 4, n. 7, p. 70-80, 2013. DOI: https://doi.org/10.26694/pensando.v4i7.1326.
WEIZENBAUM, Joseph. ELIZA – A Computer Program for the Study of Natural Language Communication Between Man And Machine. *Communications of the Association Computing Machinery*, v. 9, n. 1, p. 36-45, MIT, Cambridge, 1966. DOI: https://doi.org/10.1145/365153.365168.
WEIZENBAUM, Joseph. Computer Power and Human Reason: From Judgement to Calculation. *W.H. Freeman*. New York, 1976.