UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL
Faculdade de Ciências Humanas
Curso de Graduação em Filosofia
Kauê Barbosa de Oliveira Lopes

Filosofia da Inteligência Artificial: Desvendando os Limites e as Possibilidades da IA

Campo Grande – MS
2024

Kauê Barbosa de Oliveira Lopes

Filosofia da Inteligência Artificial: Desvendando os Limites e as Possibilidades da IA

Trabalho de Conclusão de Curso apresentado à Banca Examinadora da Universidade Federal de Mato Grosso do Sul, como pré-requisito para obtenção do título de Licenciado em Filosofia.

Orientador: Prof. Dr. Vinicius Carvalho da Silva (UFMS)

Coorientador: Prof. Dr. Judikael Castelo Branco (UFC)

Campo Grande – MS
2024

AGRADECIMENTOS

A realização deste trabalho seria impossível sem o apoio e inspiração de pessoas fundamentais em minha trajetória, que de diferentes maneiras contribuíram para que este momento se tornasse realidade.

Em primeiro lugar, agradeço ao meu pai, que me inseriu no mundo dos computadores desde cedo, despertando em mim uma curiosidade insaciável por aquelas máquinas que me cercavam. Foram suas iniciativas e ensinamentos que acenderam a fagulha de meu encantamento pela tecnologia e pelo potencial que ela carrega, algo que continua a me acompanhar até hoje.

À minha mãe, minha eterna gratidão pelo cuidado zeloso e pelo apoio incondicional em todos os momentos. Seu amor, paciência e dedicação sempre garantiram que eu tivesse as condições ideais para seguir em frente, independentemente dos desafios que surgissem no caminho.

Aos meus avós, meu mais profundo obrigado, por nunca deixarem de acreditar em mim. Foram eles que, com palavras de incentivo e amor inabalável, me encorajaram a perseguir até os sonhos mais malucos — como decidir cursar Filosofia. A confiança que depositaram em mim foi e sempre será um dos maiores presentes que pude receber.

Por fim, quero expressar minha gratidão ao professor Samuel Santana Lima, que, durante o ensino médio, me apresentou à Filosofia de uma maneira tão cativante e apaixonante que transformou minha visão de mundo. Seu empenho, conhecimento e habilidade em transmitir a essência da disciplina foram fundamentais para que eu desenvolvesse uma paixão genuína por ela, algo que carrego comigo desde então.

A todas essas pessoas especiais, que moldaram e enriqueceram minha jornada, deixo aqui o meu mais sincero agradecimento. Sem vocês, nada disso seria possível.

Na vastidão do pensar humano,
em labirintos do ser profundo,
erguem-se máquinas, fruto do arcano,
tecidas no ventre do mundo.

São circuitos que brilham em rede,
sons metálicos em compasso,
mas será que na frieza da sede
há mente além do aço?

Do barro nasceu a alma que sonha,
do silício, a lógica que calcula.
Um cria estrelas, o outro as rascunha;
um sente a vida, o outro simula.

Mas quem pode dizer com clareza,
onde acaba o humano e começa a razão?
Se nas máquinas brotar a certeza,
terão elas também coração?

Entre qualia e bits, a disputa persiste,
na fronteira do que é ser e saber.
Mentes e máquinas, o futuro insiste,
será que um dia irão se entender?
ChatGPT (OpenAI)

Resumo

A inteligência artificial (IA) vem revolucionando a nossa sociedade e já está integrada nos mais variados dispositivos e programas, como carros autônomos, mecanismos de buscas, assistentes virtuais, streamings de música e de vídeo, chatbots e muitos mais. No entanto, apesar do que muitos imaginam, a IA não se trata de algo tão recente e já existe há um certo tempo. Como veremos adiante, ao analisar os principais pontos relevantes que culminaram na criação da Inteligência artificial, essa tecnologia surgiu com o objetivo de simular processos cognitivos. Devido a isso, iremos explorar os níveis que esses sistemas podem alcançar, desde a capacidade de replicar — mesmo que de forma limitada — aspectos da cognição humana até a possibilidade de que eles desenvolvam uma consciência própria.. Toda essa investigação nos leva a importantes questões filosóficas e revive o debate mente-corpo, mas que agora se apresenta em uma nova roupagem para tratar da possibilidade do desenvolvimento de consciência em máquinas. Por fim, não iremos nos limitar a explorar apenas a ideia de que a IA pode desenvolver uma mente que se equipare à nossa, pois ao explorar o conceito de superinteligência, também vamos analisar as mais variadas ideias de máquinas ultrainteligentes que superam as capacidades cognitivas humanas em níveis inimagináveis, além de examinarmos os problemas que esse tipo de entidade pode apresentar para humanidade e os possíveis métodos de prevenção que poderíamos articular para não perder o controle de algo tão poderoso.

Palavras chaves: Inteligência, Inteligência Artificial (IA), Mente, Consciência, Superinteligência.

Abstract

Artificial intelligence (AI) has been revolutionizing our society and is already integrated into a wide variety of devices and programs, such as autonomous cars, search engines, virtual assistants, music and video streaming platforms, chatbots, and many others. However, contrary to what many might think, AI is not a recent phenomenon and has existed for quite some time. As we will see ahead, by analyzing the key milestones that led to the creation of artificial intelligence, this technology emerged with the goal of simulating cognitive processes. Because of this, we will explore the levels these systems can reach, from the ability to replicate — even if in a limited way — aspects of human cognition to the possibility that they could develop their own consciousness. This entire investigation raises significant philosophical questions and revives the mind-body debate, now framed in a new form to address the possibility of machines possessing consciousness. Finally, we will not limit ourselves to exploring only the idea that AI could develop a mind comparable to our own. By delving into the concept of superintelligence, we will also analyze the various ideas of ultraintelligent machines that surpass human cognitive abilities to unimaginable levels. Moreover, we will examine the problems such entities could pose to humanity and the possible preventive methods we could articulate to ensure that we do not lose control over something so powerful.

Keywords: Intelligence, Artificial Intelligence (AI), Mind, Consciousness, Superintelligence.

Sumário

INTRODUÇÃO

A inteligência artificial (IA) pode ser considerada um dos campos de estudo mais inovadores e influentes da atualidade. Essa área vem gerando diversas transformações nos mais variados setores da sociedade, como a ciência, a economia, a saúde, o entretenimento, a política e as relações humanas. De modo geral, a ideia da criação de máquinas que possuem capacidade de imitar a nossa inteligência tem raízes que datam de séculos atrás, como João de Fernandes Teixeira afirma:

> Os primeiros registros de criaturas artificiais com habilidades humanas têm uma forma mítica ou por vezes lendária, tornando difícil uma separação nítida entre imaginação e realidade. A idéia é de fato muito antiga, mas as condições técnicas para a sua realização são coisa recente. É esta confusão entre mito e realidade e, por vezes, a impossibilidade de distingui-los que faz com que a IA possa ser considerada uma disciplina com um extenso passado, mas com uma história relativamente curta. (Teixeira, 2019, p. 11)

Assim, embora a ideia de inteligência artificial existisse há muito tempo, frequentemente envolta em mitos e lendas[1], foi somente com os avanços da ciência cognitiva, das tecnologias computacionais e o otimismo científico do século XX que essa concepção começou a se concretizar como uma realidade tangível.

Com isso em vista, nossa pesquisa tem como uma de suas intenções analisar a trajetória que culminou na origem da inteligência artificial, passando por alguns dos principais marcos e nomes responsáveis pelo surgimento da IA. Tal abordagem se faz necessária para que tenhamos uma base concreta do contexto e dos avanços que lapidaram a inteligência artificial como conhecemos. Além disso, conceitos e objetivos importantes que foram traçados durante o desenvolvimento desse campo de estudos ainda se mantêm relevantes na atualidade. Ademais, o recorte histórico nos fornece um excelente ponto de partida para guiar as nossas investigações posteriores de maneira mais ampla e fundamentada.

De forma complementar, em nossa investigação, também iremos examinar os conceitos de IA fraca e IA forte. Isso se deve ao fato de que um dos objetivos do desenvolvimento da inteligência artificial se trata da capacidade de simular a cognição

[1] Como exemplo dessa afirmação, Teixeira (2019) nos apresenta a lenda de Gólem, Joseph Gólem era um homem artificial dotado de inteligência, criado no século XVI para espionar os inimigos dos judeus confinados no gueto de Praga. Porém, Joseph acaba se revoltando contra seu criador que resolve retirar-lhe sua consciência.

humana[2]. Com isso em vista, podemos notar que nos últimos tempos essa tecnologia anda em um constante avanço, atualmente temos programas capazes de conversar conosco, nos auxiliar com as tarefas do dia-a-dia, criar imagens, poemas, dirigir carros e aeronaves de forma autônoma, além de várias outras tarefas que a inteligência artificial desempenha até mesmo melhor que os seres humanos em alguns casos. No entanto, será que podemos considerar que a IA é capaz de simular a cognição humana perfeitamente? Os conceitos de IA fraca e IA forte vão nos auxiliar a responder essa questão.

Em seguida, vamos evocar um importante debate da filosofia, a questão mente-corpo, mas agora sob a luz das máquinas. De maneira simplificada, essa questão busca investigar qual a natureza da nossa mente e a sua relação com o nosso corpo, esse tema envolve questões sobre o que é a consciência[3] e se ela pode ou não ser reduzida a processos físicos, por exemplo. Sendo assim, em nosso recorte, iremos analisar diversas teorias como o dualismo cartesiano, o funcionalismo, o materialismo eliminativista, os *qualia*, a teoria computacional da mente, além dos limites da computação para examinarmos a possibilidade do desenvolvimento de consciência em inteligências artificiais.

Por fim, após estarmos familiarizados com os principais conceitos, capacidades e problemas filosóficos que envolvem a inteligência artificial como um todo iremos explorar um modelo conceitual que se pretende capaz de superar a cognição humana de maneira inimaginável: a superinteligência. Esse tipo de IA possui um potencial extraordinário para revolucionar a forma como vivemos, mas também apresenta enormes riscos para a humanidade, caso seus objetivos não estejam alinhados com os interesses e valores humanos. Devido a isso, vamos analisar os possíveis tipos de superinteligência, os problemas que a acompanham e as suas razoáveis soluções.

Desse modo, ao longo deste trabalho, espera-se proporcionar uma análise abrangente e crítica da inteligência artificial, desde suas raízes conceituais até suas aplicações mais ambiciosas e potencialmente disruptivas. Através de uma abordagem que integra história, filosofia e ciência, busca-se compreender não apenas onde estamos no desenvolvimento da

[2] Máquinas com inteligência geral comparável à dos humanos — ou seja, dotadas de bom senso e capacidade real de aprender, raciocinar e planejar a superação de desafios complexos de processamento de informação em uma vasta gama de domínios naturais e abstratos — têm sido esperadas desde a invenção dos computadores, na década de 1940. (Bostrom, 2018, p. 31).

[3] Neste trabalho, não nos propomos a definir ou explicar os conceitos de consciência ou mente, dado que se trata de um debate filosófico e científico ainda em aberto, com diversas abordagens e interpretações por parte dos teóricos. Para fins de simplificação, consideraremos os termos "consciência" e "mente" como sinônimos ao longo deste estudo, sem nos aprofundarmos em suas distinções.

IA, mas também o que o futuro nos reserva e como podemos nos preparar para os possíveis impactos dessa tecnologia transformadora.

1. ORIGENS DA INTELIGÊNCIA ARTIFICIAL

1.1 Descartes, os animais e as máquinas

Quando refletimos sobre a Inteligência Artificial (IA), é importante ter em mente que não estamos lidando com um conceito inteiramente novo. René Descartes, importante filósofo francês do século XVII, em seu livro *Discurso do Método* (2001), expressa suas considerações sobre máquinas e sua incapacidade de se tornarem "homens verdadeiros"[4]. Para o filósofo, mesmo que houvesse algum tipo de máquina que se parecesse com uma pessoa e até mesmo se espelhasse moralmente nos seres humanos, esta ainda de forma alguma poderia ser considerado de fato um indivíduo humano, Descartes esclarece suas noções na seguinte passagem:

> (...) Se houvesse máquinas assim que tivessem os órgãos e os aspecto de um macaco ou de qualquer outro animal sem razão, não teríamos nenhum meio de reconhecer que elas não seriam, em tudo, da mesma natureza desses animais; ao passo que, se houvesse algumas que se assemelhassem a nossos corpos e imitassem as nossas ações tanto quanto moralmente é possível, teríamos sempre dois meios muito certos para reconhecer que, mesmo assim, não seriam homens verdadeiros. O primeiro é que nunca poderiam servir-se de palavras nem de outros sinais, combinando-os como fazemos para declarar aos outros nossos pensamentos. Pois pode-se conceber que uma máquina seja feita de tal modo que profira palavras, e até profira algumas a propósito das ações corporais que causem alguma mudança em seus órgãos, como por exemplo ela perguntar o que lhe queremos dizer se lhe tocarmos em algum lugar, se em outro, gritar que a machucamos, e outras coisas semelhantes, mas não é possível conceber que as combine de outro modo para responder ao sentido de tudo quanto dissermos em sua presença, como os homens mais embrutecidos podem fazer. E o segundo é que, embora fizessem várias coisas tão bem ou talvez melhor do que algum de nós, essas máquinas falhariam necessariamente em outras pelas quais se descobriria que não agiam por conhecimento, mas somente pela disposição de seus órgãos. Pois, enquanto a razão é um instrumento universal, que pode servir em todas as circunstâncias, esses órgãos necessitam de alguma disposição particular para cada ação particular; daí ser moralmente impossível que haja numa máquina a diversidade suficiente de órgãos para fazê-la agir em todas as ocorrências da vida da mesma maneira que nossa razão nos faz agir. (2001, p. 63 - 64).

Além disso, ainda nesse trecho, é interessante observar como o filósofo expõem que as máquinas poderiam ser semelhantes aos animais. Isso decorre da maneira como Descartes compreende a realidade. Para ele, existem duas substâncias fundamentais: a *res cogitans*, no

[4] É importante salientar que o conceito de "homem verdadeiro" em Descartes pode ser um tanto problemático, como aponta Rocha em seu artigo "Animais, Homens e Sensações Segundo Descartes" (2004). Na medida em que o filósofo rejeita que máquinas ou animais possam ter consciência, atribuindo isso apenas aos humanos, sua argumentação abre brechas para que apenas o "eu" cartesiano — o indivíduo que pensa e reflete sobre si mesmo — seja considerado um "homem verdadeiro". Rocha observa que, ao depender exclusivamente de expressões naturais e comportamentais para inferir a consciência em outros seres, o mesmo critério que descarta os animais também poderia questionar a certeza sobre a consciência de outros seres humanos. Isso sugere que, no limite, o único "homem verdadeiro" para Descartes é o "eu" que tem acesso direto à sua própria mente.

caso, a alma, que é a substância pensante, dotada de entendimento e vontade, e a *res extensa*, que é a substância material responsável por constituir os corpos físicos, por exemplo. Os animais e as máquinas — na concepção do filósofo — são criaturas compostas apenas de *res extensa* e, portanto, desprovidas de alma, o que significa que não possuem entendimento nem vontade, diferente dos seres humanos. Por isso, o pensador acreditava ser coerente que uma máquina pudesse imitar um animal, mas impossível que ela conseguisse se assemelhar a um ser humano ao ponto de se igualar a nós. Dessa forma, mesmo que uma máquina pudesse conversar conosco em alto e bom som, ainda sim ela não teria a capacidade de compor frases, devido ao fato de não ter vontade própria para formar suas falas, tudo o que dissesse seria apenas resultado de mecanismos organizados de tal maneira que imitasse a fala humana, mas sem sua devida autonomia e por ser desprovido de razão e não ter a capacidade de pensar, a máquina sequer poderia compreender qualquer coisa que pronunciasse.

Dessa maneira, podemos perceber que a discussão sobre coisas artificiais e a sua capacidade (ou incapacidade) de desenvolver algum tipo de inteligência não é algo recente. Descartes é apenas um exemplo que evidencia esse ponto, mas sequer foi o primeiro a pensar no assunto[5]. Até não muito tempo atrás, esses tipos de agentes artificiais, eram relacionados a autômatos mecânicos, máquinas a vapor e até mesmo a magia. No entanto, tudo isso muda a partir do século XX, é com a explosão do avanço tecnológico, causado pela Segunda Guerra Mundial, que a Inteligência Artificial encontra um novo hospedeiro para se manifestar, um dispositivo cibernético movido a eletricidade: o computador. É a partir daí que a IA como conhecemos é concebida, e diga-se de passagem, em um ambiente muito propício, marcado pelo desenvolvimento de novas tecnologias, um período de otimismo em relação ao progresso científico e mudanças significativas nas pesquisas no campo da psicologia, que passaram a popularizar a ideia de que a mente humana possuía semelhanças lógicas em relação aos computadores.

[5] Um dos primeiros registros mitológicos de um artefato artificial dotado de inteligência remonta ao século III a.C., em um poema épico que descreve Talos, um gigante de bronze criado por Hefesto. Talos era um autômato guardião da ilha de Creta e circulava sua costa três vezes ao dia para proteger o local de invasores. De maneira geral, Talos pode ser considerado um dos primeiros exemplos de uma máquina artificial com comportamento programado, mesmo que de maneira rudimentar. Além do exemplo mitológico, a antiguidade também nos legou um dispositivo real: o Mecanismo de Anticítera. Criado por volta de 100 a.C., o Mecanismo de Anticítera é um artefato real e comprovado que foi descoberto em 1901 em um naufrágio perto da ilha de Anticítera, na Grécia. Este computador analógico foi projetado para prever eventos astronômicos e embora não possuísse "inteligência" no sentido moderno, o Mecanismo de Anticítera era capaz de realizar cálculos complexos e automatizados, permitindo que os antigos astrônomos obtivessem informações detalhadas sobre o cosmos sem depender de cálculos manuais.

1.2 A máquina de Turing e a arquitetura de John Von Neumann

De maneira geral, podemos compreender melhor o contexto que levou ao desenvolvimento da Inteligência Artificial, a partir dos marcos iniciais da computação, como a proposta da Máquina de Turing e a arquitetura de John Von Neumann. Em 1936, o matemático e cientista da computação Alan Turing publicou o artigo "*On Computable Numbers, with an Application to the Entscheidungsproblem*"[6], no qual apresentou a *Máquina de Turing*. Esse modelo abstrato é uma representação teórica de uma máquina com memória e tempo infinitos, capaz de resolver problemas seguindo passos bem definidos. O objetivo de Turing era investigar o problema da decisão (Entscheidungsproblem), proposto pelo matemático David Hilbert e seu aluno Wilhelm Ackermann. Como explica o pesquisador Fernando Ferreira (2019), Hilbert buscava uma maneira de colocar toda a matemática em bases formais, seguras e decidíveis. Em outras palavras, Hilbert pretendia encontrar através do problema da decisão um método sistemático, automático e mecânico para determinar se qualquer afirmação matemática era verdadeira ou falsa, sem depender de criatividade ou intuição humana.

Entretanto, Turing demonstrou que esse objetivo não poderia ser alcançado. Ele provou, através da sua máquina abstrata, que existem problemas de decisão que não podem ser resolvidos por nenhum método mecânico, ou seja, são incomputáveis[7]. Essa descoberta revelou os limites da computação, mostrando que há problemas intratáveis e estabelecendo fronteiras claras entre o que é e o que não é resolvível computacionalmente. Turing também demonstrou que problemas computáveis — aqueles que podem ser resolvidos por meio de algoritmos finitos e bem definidos — são solucionáveis por máquinas, desde que estas possuam recursos ilimitados de tempo e memória. Esse conceito teórico tornou-se a base para o desenvolvimento dos algoritmos e da computação moderna. Cerca de quase uma década depois, em 1945, o físico-matemático John Von Neumann, transformou essa teoria abstrata em algo prático. Inspirando-se na Máquina de Turing, ele desenvolveu a Arquitetura de von Neumann, que se consolidou como o padrão para computadores. No entanto, é importante destacar que, ao contrário da Máquina de Turing, essa arquitetura utiliza tempo e memória finitos e foi projetada para operar com unidades de processamento e armazenamento,

[6] O desenvolvimento deste estudo é fundamental para a inteligência artificial, pois estabelece os limites entre o que uma máquina pode ou não realizar, gerando questionamentos sobre a capacidade das máquinas de replicar processos cognitivos.

[7] O termo "incomputável" refere-se a problemas que não podem ser resolvidos por qualquer algoritmo ou processo computacional, ou seja, não há uma sequência finita de passos que uma máquina possa seguir para encontrar uma solução.

permitindo a execução de programas previamente armazenados. Essa estrutura permanece como a referência principal para a maioria dos computadores até hoje.

1.3 O artigo de Warren McCulloch e Walter Pitts

Em 1943, Warren McCulloch e Walter Pitts, precursores do que viria a ser conhecido como ciência cognitiva, escreveram um artigo denominado *A Logical Calculus of the Ideas Immanent in Nervous Activity* em que apresentavam um modelo matemático que buscava descrever eventos neurais e suas interações por meio da lógica proposicional, como explicam os cientistas da computação Peter Norvig e Stuart Russell:

> Eles [Warren McCulloch e Walter Pitts] se basearam em três fontes: o conhecimento da fisiologia básica e da função dos neurônios no cérebro; uma análise formal da lógica proposicional criada por Russell e Whitehead; e a teoria da computação de Turing. Esses dois pesquisadores propuseram um modelo de neurônios artificiais, no qual cada neurônio se caracteriza por estar "ligado" ou "desligado", com a troca para "ligado" ocorrendo em resposta à estimulação por um número suficiente de neurônios vizinhos. O estado de um neurônio era considerado "equivalente em termos concretos a uma proposição que definia seu estímulo adequado". Por exemplo, eles mostraram que qualquer função computável podia ser calculada por certa rede de neurônios conectados e que todos os conectivos lógicos (e, ou, não etc.) podiam ser implementados por estruturas de redes simples. (2013 p. 42).

Dessa forma, o artigo propunha que a atividade nervosa, devido ao seu caráter "tudo ou nada"[8], poderia ser descrita de maneira binária, facilitando sua representação por meio de operações lógico-matemáticas. A análise das conexões entre múltiplos neurônios revelou que era possível simular funções lógicas complexas, como conjunção, disjunção e negação, utilizando redes de neurônios interconectados. Essas redes, estruturadas como circuitos lógicos, foram fundamentais para a formulação de um modelo teórico de processamento de informações. Esse modelo lançou as bases para a ideia de que a inteligência humana poderia ser não apenas compreendida, mas também simulada de forma artificial.

1.4 O Simpósio de Hixon

Anos mais tarde, os primeiros vestígios de uma nova ciência começaram a emergir de forma significativa. No entanto, ainda não se tratava da Inteligência Artificial, mas sim da ciência cognitiva, um campo de estudo dedicado a investigar o funcionamento da mente humana. Como afirma Teixeira:

> No fim da Segunda Guerra Mundial, os cientistas já tinham registrado importantes invenções na área eletrônica, além de pesquisas sobre mecanismos que imitavam

[8] (MCCULLOCH e PITTS, 1943, p. 115).

> ações humanas e estudos sobre o cérebro humano desenvolvidos por médicos e por psicólogos. Isso os levou a programarem um encontro nos Estados Unidos, onde pesquisadores dessas áreas apresentaram suas descobertas, numa primeira tentativa de reuni-las e compor algo parecido com uma ciência geral do funcionamento da mente humana. (2019 p. 13-14).

Dessa maneira, esse encontro ocorreu em 1948, no instituto de Tecnologia da Califórnia e ficou conhecido como o Simpósio de Hixon. O objetivo do evento era discutir o modo como o cérebro humano processa informações e produz comportamentos. Para realizar essa discussão, foram convocados especialistas das mais diversas áreas do conhecimento como a biologia, a computação, a psicologia, a matemática e a filosofia. Essa colaboração interdisciplinar visava integrar as descobertas e métodos de diferentes campos para construir uma compreensão mais abrangente do cérebro e da mente através do que ficaria conhecido como ciência cognitiva. Dentre os pesquisadores presentes no evento, se encontravam os já mencionados John Von Neumann, Warren McCulloch e também o matemático Norbert Wiener (1894 - 1964), responsável por cunhar o termo *cibernética*[9]. Foram estes estudiosos, junto a outros importantes cientistas que levantaram questões como, as similaridades entre o funcionamento do cérebro e dos computadores e as discussões sobre o modo como a cibernética e o processamento de informações poderiam ser importantes aliados para compreender o sistema do cérebro humano. Além disso, a ideia de que os processos mentais poderiam ser estruturados de forma matemática foi certamente uma semente que ao cair no fértil terreno do Simpósio de Hixon abriu as portas para o surgimento da Inteligência artificial.

1.5 O Workshop Dartmouth Summer Project

Com um cenário que se apresentava cada vez mais favorável ao desenvolvimento da IA, a primeira manifestação concreta da inteligência artificial como um campo de estudo, surgiu na Universidade Dartmouth College, em Hanover, durante o Workshop Dartmouth Summer Project em 1956. Como afirmam Russell e Norvig (2013), o evento foi organizado por John McCarthy (cientista da computação) que conseguiu convencer Marvin Minsky (cientista cognitivo), Claude Shannon (matemático e engenheiro eletrônico) e Nathaniel Rochester (cientista da computação) a unir esforços para encontrar pesquisadores dos Estados Unidos interessados em teoria de autômatos, redes neurais e estudo da inteligência. Desse modo, o workshop conseguiu reunir cientistas durante seis semanas para promover estudos

[9] Wiener apresentou o conceito de *cibernética* pela primeira vez em sua obra *Cybernetics: Or Control and Communication in the Animal and the Machine* em 1948. Esse termo é utilizado para se referir ao estudo dos processos de controle e comunicação dos seres vivos e das máquinas.

sobre a capacidade de criar máquinas que pudessem simular a inteligência humana. McCarthy e alguns outros pesquisadores presentes no evento, compartilhavam a hipótese de que todas as características da inteligência humana que pudessem ser descritas de maneira precisa para uma máquina poderiam ser simuladas através de modelos lógicos[10]. Graças a esse encontro e as suas reflexões a Inteligência Artificial se tornou uma disciplina científica. Além disso, devido ao otimismo acerca do campo e a grande onda de financiamentos, várias áreas de pesquisas próximas à Inteligência artificial começaram a surgir e se desenvolver, como estudos sobre programas de reconhecimento de padrões, sistemas de resolução de problemas e desenvolvimento de redes neurais.

1.6 Primavera da IA

Após o Workshop da Universidade de Dartmouth College os estudos em IA sofreram uma explosão de popularização e contribuíram para o desenvolvimento tecnológico das décadas seguintes, como Bostrom descreve:

> O primeiro período de entusiasmo, que iniciou com o encontro em Dartmouth, foi mais tarde descrito por John McCarthy (o principal organizador do evento) como a era do "Veja, mamãe, sem as mãos!". Nessa época, pesquisadores construíram sistemas projetados para contestar afirmações como: "Nenhuma máquina jamais seria capaz de fazer X!". Tais afirmações céticas eram comuns no período. Para contrapô-las, os pesquisadores da IA criavam pequenos sistemas que faziam X em um "micromundo" (um domínio bem definido e limitado que tornava possível uma versão reduzida da performance a ser demonstrada), fornecendo, dessa forma, uma comprovação do conceito e mostrando que X poderia, em princípio, ser feito por uma máquina. (Bostrom, 2018, p. 34).

Dessa maneira, foi graças ao desenvolvimento desses sistemas que aparelhos e programas com habilidade para vencer os melhores jogadores de xadrez do mundo, capacidade de resolver grandes problemas matemáticos e competência para traduzir diversas linguagens humanas, começaram a surgir e evoluir cada vez mais. Entretanto, é importante salientar que todo desenvolvimento da Inteligência Artificial da década de 50 até o período atual não foi tão simples e marcado apenas por avanços. Ao longo desse percurso, surgiram também limitações técnicas e teóricas que desafiaram os progressos da área. Em diversos momentos (que infelizmente não entraremos em detalhes) as pesquisas encontraram grandes

[10]Segundo Bostrom, a proposta apresentada à Rockefeller Foundation destacava que os estudos desenvolvidos pelo Workshop Dartmouth College seriam realizados com base na hipótese de que todos os aspectos da aprendizagem ou de qualquer outra característica da inteligência que pudesse ser descrito de forma precisa a uma máquina poderia ser simulado.(Bostrom, 2018, p. 33).

problemas para que resultaram no chamado "Inverno da IA"[11] em que o interesse na área e o seu financiamento sofreram uma grande redução, devido a falta de realizações significativas. No entanto, de tempos em tempos a inteligência artificial se depara com uma florida primavera em que se torna relevante novamente e recebe grandes incentivos financeiros, após apresentar resultados promissores. Como os leitores devem imaginar, estamos nesse exato momento em meio a um campo de flores que acabaram de desabrochar.

2. OS NÍVEIS DE INTELIGÊNCIA ARTIFICIAL: IA FRACA E IA FORTE

Nos últimos anos a inteligência artificial parece estar se desenvolvendo cada vez mais, seus modelos se apresentam progressivamente mais completos e capazes de se integrar nos mais variados tipos de sistemas, celulares, computadores, relógios e programas que utilizamos rotineiramente. Com isso em vista, faz se necessário pensar de que maneira podemos classificar o avanço dessas máquinas, estamos prestes a alcançar um novo patamar? Essa será a nossa investigação neste capítulo. Desse modo, iremos explorar as definições de IA Fraca e IA Forte, além de nos atentarmos para os experimentos dos pensadores Alan Turing e John Searle que apresentam suas compreensões acerca da possibilidade do desenvolvimento da IA forte.

2.1 IA Fraca (ANI)

De modo geral, quando nos referimos a IA fraca, também conhecida como inteligência artificial estreita (ANI), estamos trabalhando com a noção de máquinas que possuem a capacidade de simular algumas características do comportamento ou da inteligência humana para realizar tarefas, mas sem que isso implique em algum nível de consciência ou estado menta. Dessa forma, por mais que essas máquinas possam agir de maneira que julgamos inteligentes — e podemos compreender o termo "inteligente", nesse cenário, como a capacidade de resolver problemas tipicamente associados a inteligência humana — elas não passam de ferramentas e diga-se de passagem, ferramentas muito eficientes na maioria das vezes, mas desprovidas do que alguns filósofos chamariam de *qualia*[12] — ou seja, elas não possuem qualquer forma de experiência subjetiva consciente. Em vez disso, essas ferramentas

[11] Segundo Bostrom, "A constatação de que muitos dos projetos de IA jamais poderiam realizar suas promessas iniciais levou ao surgimento do primeiro 'inverno da IA' um período de retração durante o qual os financiamentos diminuíram, o ceticismo aumentou, e a IA, então, saiu de moda."(2018, p.35). Esse fenômeno ocorreu pela primeira vez por volta de 1970, marcando o início de uma fase de retrocesso na área.

[12] O termo *"qualia"* será explorado em capítulos posteriores.

apenas replicam, de maneira limitada, aspectos específicos da cognição sem alcançar a profundidade subjetiva que caracteriza a consciência humana.

Com isso em vista, podemos compreender a IA fraca como um tipo de instrumento que através de sistemas algorítmicos e dados, são desenvolvidas para resolver problemas específicos e que para isso, buscam imitar, na medida do que se é possível, a inteligência humana. Desse modo, quando utilizamos aplicativos de localização, recebemos recomendações de anúncios em redes sociais, sugestões do que assistir em streamings de vídeos, conversamos com chatbots, viajamos em veículos autônomos, utilizamos mecanismos de buscas digitais, ligamos nossos celulares e computadores, estamos a todo tempo em contato com esse tipo de tecnologia e por mais fabuloso que seja o grau elevado que todos esses sistemas estão chegando, nenhum deles conseguiu superar a classificação de IA fraca. Por mais que estejam cada vez mais otimizadas e correspondam de maneira adequada a suas tarefas, a IA não manifesta nenhum tipo de intencionalidade[13], ou seja, diferente dos seres humanos que possuem a capacidade de dirigir seus pensamentos, crenças ou desejos a objetos ou ideias, essas máquinas apesar de processar informações, não conseguem pensar sobre o que estão fazendo, não possuem qualquer tipo de entendimento subjetivo daquele objeto ou objetivo que estão processando. Dessa forma, quando fazemos alguma pergunta para o ChatGPT, por exemplo, ele não pensa de maneira subjetiva sobre a nossa questão, o que ocorre é uma série de cálculos matemáticos que geram uma resposta. Russell e Norvig nos ajudam compreender melhor essa ideia na seguinte passagem:

> É claro que os computadores podem fazer muitas coisas tão bem ou melhor que os humanos, incluindo aquelas que as pessoas acreditam que exigem grande perspicácia e compreensão humana. No entanto, isso não significa que os computadores utilizam a perspicácia e a compreensão na execução dessas tarefas [...] (Russell e Norvig, 2013, p. 1131).

Além disso, esse tipo de IA geralmente resolve problemas, analisa padrões e aprende a lidar apenas com um determinado tipo de tarefa. Dessa maneira, em vários segmentos, esses sistemas podem apresentar um nível de desempenho muito próximo ou até superar a inteligência humana em alguns aspectos, mas não conseguem manter esse mesmo nível em

[13]Segundo Teixeira: A intencionalidade se manifesta na medida em que sabemos a que se referem nossos estados mentais. Quando falamos, não estamos apenas emitindo sons: sabemos do que estamos falando e que nossas palavras se referem a coisas que estão no mundo. Todos os nossos pensamentos - sejam expressos em palavras ou não - têm conteúdos que apontam para coisas ou situações do mundo. É impossível estar pensando sem estar pensando em alguma coisa. E quando estamos pensando, sabemos selecionar, entre nossos estados mentais, aqueles que apontam para objetos que estão à nossa volta e aqueles que são mais distantes, como, por exemplo, os conteúdos da nossa imaginação. De qualquer maneira, há sempre uma direcionalidade, algo como um apontar para fora de nós mesmos que faz com que nossos pensamentos adquiram significado ou sentido. (Teixeira, 2019, p. 43).

outros tipos de atividades sem ser aquelas pelas quais foi programado. Ou seja, uma IA especializada em traduções pode traduzir um discurso para mais de 80 línguas, um feito extremamente difícil e talvez até impossível mesmo para um excelente poliglota humano, mas esse mesmo sistema sem treinamento e uma adaptação específica não teria as instruções e dados necessários para disputar uma partida de xadrez. Devido a isso, a confecção dessas máquinas tem como ponto central a maximização da sua funcionalidade e utilidade e não necessariamente a capacidade de desempenhar múltiplas tarefas de diferentes campos ou se provar consciente.

2.2 IA Forte (AGI)

De diferente modo, a IA forte, também referida como inteligência artificial geral (AGI), é um conceito que especula ser possível que, a partir dos avanços tecnológicos adequados, as máquinas teriam o potencial necessário para alcançar todas as capacidades cognitivas de um ser humano, e cumprir qualquer tarefa intelectual sem maiores dificuldades. Além disso, uma IA com esse nível de desenvolvimento, em tese, também poderia superar os obstáculos que impedem a possibilidade da confecção de uma consciência de si mesma[14]. Dessa maneira, é importante compreender que a partir desse cenário a IA teria a capacidade de produzir pensamentos e entendimentos próprios sem qualquer necessidade de simular a inteligência humana. Em estágios mais avançados, essas máquinas também poderiam desenvolver intencionalidade, compreender o mundo ao seu redor e direcionar seus pensamentos aos objetos. Assim, elas seriam capazes de produzir experiências subjetivas e até expressar emoções. Entretanto, todos esses pontos geram uma infinidade de problemas filosóficos como veremos adiante. Para complementar um pouco mais essa discussão, é interessante analisarmos alguns experimentos que demonstram as expectativas de alguns importantes pesquisadores acerca da questão da consciência em IA de nível forte.

2.3 Experimentos

Primeiramente, podemos analisar o experimento do cientista da computação Alan Turing. Estimulado pela questão de que se máquinas poderiam pensar, Turing escreveu o artigo "*Computing machinery and intelligence*" (1950), publicado na revista de filosofia *Mind.* Em seu texto, o cientista da computação propõe um teste de verificação, cujo objetivo

[14]Segundo Norvig e Russell: a asserção de que as máquinas talvez possam agir de maneira inteligente (ou, quem sabe, agir como se fossem inteligentes) é chamada hipótese de IA fraca pelos filósofos, e a asserção de que as máquinas que o fazem estão realmente pensando (em vez de simularem o pensamento) é chamada hipótese de IA forte. (2013, p. 1129).

se trata de responder se máquinas poderiam se comunicar de maneira indistinguível de um ser humano. O teste em questão é composto por um participante humano e um participante artificial (computador). Durante o experimento, ambos devem conversar com um juiz humano por meio de um terminal. Nesse diálogo, que poderia ser sobre poesia, matemática, experiências de vida e diversos outros assuntos, os participantes devem persuadir o juiz e provar sua humanidade. Dessa forma, caso o juiz não consiga determinar quem é o participante não humano, o computador vence. Turing então propõe que caso uma máquina possa sustentar uma conversa de maneira que seja impossível saber se ela é um ser humano ou não, a atitude correta seria interpretá-la como consciente, por simples cortesia, já que no dia-a-dia não paramos para investigar se outros seres humanos possuem estados mentais ou consciência, apenas aceitamos o fato de que todos possuem essas capacidades. O *Jogo da Imitação*, como ficou conhecido, foi um dos primeiros estudos realizados sobre o tema no século XX, todavia, recebeu muitas críticas[15], pois a capacidade de persuadir ou se passar por um ser humano não parece ser prova suficiente para sustentar a hipótese de que máquinas podem pensar.

Ademais, anos mais tarde, o filósofo John Searle em contraposição às ideias de Turing, desenvolveu um experimento mental que buscava demonstrar a inexistência de cognição relevante em máquinas. O chamado *Argumento da sala chinesa* (1980), consiste em um experimento no qual o filósofo está preso em uma sala com apenas uma janela, por onde são entregues três folhas: uma com questões em chinês, uma totalmente em branco e uma com instruções redigidas em inglês. Searle não fala chinês, mas deve escrever na folha em branco respostas às perguntas que estão nessa língua. Para isso, deverá utilizar em seu auxílio a terceira folha, composta por regras de resposta, escritas em sua língua nativa (inglês). As regras não ensinam chinês ou explicam o significado dos caracteres, elas apenas indicam quais símbolos do alfabeto chinês devem ser redigidos, quando determinados caracteres da língua chinesa aparecerem. Ao seguir as regras e responder à mensagem, Searle deverá devolver a folha respondida pela janela. Assim, do lado de fora da sala, uma pessoa receberá uma mensagem em chinês escrita por Searle. Para o sujeito que recebeu o papel com o texto redigido, Searle parece ser fluente em chinês, mas ele realmente sabe chinês? É evidente que não, já que apenas está seguindo as orientações que lhe foram passadas em sua língua nativa,

[15] Será que o simples exame do comportamento externo, observável, de um organismo ou de uma máquina é suficiente para que possamos concluir que esse organismo ou máquina é capaz de ter pensamento? Um CD player toca Bach e Beethoven da mesma maneira que um músico o faz. Mas nunca diríamos de um CD player que ele o faz intencionalmente. Nem tampouco aplaudiríamos um CD player no final de uma execução. (Teixeira, 2019, p. 39).

ou seja, ele desconhece o significado dos símbolos que observou e escreveu. De maneira análoga, para Searle, as máquinas exibem a mesma conduta, seguem regras e respondem símbolos com outros símbolos sem que entendam o seu sentido, sem que ajam com qualquer tipo de intencionalidade. Todavia, Russell e Norvig (2013) afirmam que, da mesma maneira que Turing, o argumento de Searle também sofreu diversas críticas, até mesmo de renomados cientistas da computação como John McCarthy e Robert Wilensky.

De maneira geral, apesar de vários outros experimentos e estudos como esses terem sido produzidos, a questão da possibilidade da confecção da IA forte ainda é um problema em aberto, pois diversas são as abordagens que negam ou acham plausível essa possibilidade. Contudo, toda essa incerteza não impediu que especulações sobre o tema continuassem a ser feitas e novas concepções passassem a tomar forma como veremos nos capítulos seguintes.

3. O PROBLEMA MENTE-CORPO SOB O RECORTE DA IA

Como vimos anteriormente, desde suas origens, ainda no Workshop Dartmouth Summer Project de 1956, a inteligência artificial costuma ser pensada como uma aliada que pode nos auxiliar a compreender nossos processos cerebrais, nossa forma de pensar e nossa consciência. Para isso, uma das maneiras que vários cientistas compreendem como um modo de nos ajudar a entender o funcionamento do nosso cérebro se trata da simulação. Muitas vezes a IA é vista como uma possível ferramenta apta a simular nossos processos mentais, mas será que ela é mesmo capaz? Será possível desenvolver uma inteligência artificial com essa característica? Para tentar responder essas perguntas, ou ao menos apresentar os possíveis caminhos para uma resposta, é necessário que investiguemos um campo conhecido como filosofia da mente. Dessa forma, iremos nos dedicar a analisar, através do nosso recorte em inteligência artificial, algumas das diversas teorias que buscam traçar respostas para o problema mente-corpo, como o dualismo, o funcionalismo, os *qualia*, o materialismo eliminativista e a teoria computacional da mente. Além disso, também trataremos de investigar os limites da computação, para que possamos compreender até onde essas simulações da inteligência humana podem chegar.

3.1 Dualismo Cartesiano

Para iniciarmos nossa investigação, trataremos do dualismo cartesiano, criado pelo filósofo René Descartes. É importante compreender que o dualismo aborda dois conceitos fundamentais: a alma racional (*res cogitans*), exclusiva dos seres humanos, e a substância

material (*res extensa*). Segundo Maslin (2009, p. 47), "Concebida como uma alma, uma pessoa é totalmente distinta e diferente em espécie de seu corpo físico, que ocupa espaço e extensão. Descartes se refere à alma como *res cogitans*, uma coisa pensante." A partir disso, podemos entender a alma como uma espécie de sinônimo de mente, já que a consciência, o pensamento e a razão são características atribuídas à alma. Sendo assim, o dualismo explana a ideia de que a mente e o corpo são duas coisas diferentes, compostas por substâncias distintas. Dessa maneira, enquanto a mente se trata de algo imaterial, o corpo é composto por matéria e possui um estado físico. Entretanto, para Descartes (2001), apesar dessa diferença de substâncias entre a mente e o corpo, ambos conseguem se relacionar. Com isso em vista, quando pensamos em inteligência artificial, a questão que emerge é se é possível que uma máquina possa ter ou simular uma mente identica a humana. Se levarmos as considerações de Descartes, aparentemente não seria plausível supor que uma máquina desenvolva algo semelhante, já que para isso, seria necessário que a inteligência artificial também possuísse *res cogitans*, mas na concepção do filósofo (2001) essa substância é uma característica única do ser humano e nenhum outro animal ou máquina é dotado dela. Dessa forma, por mais avançada e completa que fosse a arquitetura de um sistema, sua simulação dos estados mentais nunca poderia alcançar de fato a mente humana.

3.2 Funcionalismo

Outra importante abordagem filosófica que merece nossa atenção se trata do funcionalismo, uma corrente influenciada diretamente pelo desenvolvimento do campo da inteligência artificial. Segundo Russell e Norvig (2013) essa abordagem foi desenvolvida por filósofos como Hilary Putnam e Daniel Dennett. Diferente da noção expressa no dualismo cartesiano, o funcionalismo não concebe a mente como uma substância imaterial — e também não a compreende apenas como o cérebro — essa vertente trata a mente como uma espécie de função. Para compreender melhor essa noção podemos investigar a seguinte definição do filósofo Keith Maslin:

> O funcionalismo, como o nome implica, concebe a mente como uma função. Mas o que é uma função? Um modo fácil de entender isso é pensar em um termostato. A função de um termostato é regular a temperatura de uma sala ou edifício, Ele toma certo input na forma de temperatura ambiente e então, dependendo de como foi programado, produz um dos resultados seguintes: (1) aciona o sistema de aquecimento, porque a sala está fria demais; (2) desliga o sistema de aquecimento, porque a sala está quente demais. (Maslin, 2009, p. 129).

Dessa maneira, a composição da mente ou a forma como ela executa suas ações não tem importância, a única coisa relevante é a função que ela exerce e os processos que realiza.

Devido a isso, da mesma forma que um ser humano poderia assumir a função do termostato e controlar a temperatura do ambiente ativando o sistema que irá aquecer ou esfriar a sala, caso uma máquina seja capaz de replicar a função da mente humana, podemos considerar que ela pode reproduzi-la perfeitamente. Em vista disso, podemos compreender que uma mesma função pode ser executada por diversos mecanismos ou entidades. Sendo assim, se a teoria dualista de Descartes exclui qualquer possibilidade de que a inteligência artificial se torne semelhante à inteligência humana, o funcionalismo oferece o fundamento filosófico necessário para que essa mesma possibilidade se concretize.

3.3 Qualia

Uma abordagem relevante que contesta o funcionalismo se trata da teoria dos qualia, geralmente associada aos filósofos Thomas Nagel e Frank Jackson[16]. Os *qualia* podem ser compreendidos como nossas experiências subjetivas, intrínsecas a cada ser, como o gosto único que cada indivíduo experimenta ao tomar um café, a dor específica que sente ao se machucar ou a coloração que enxerga ao observar o mundo. Para entendermos melhor essa definição podemos recorrer ao filósofo Thomas Nagel e seu artigo intitulado *What is it like to be a bat* (1974), nele Nagel pede para que o leitor se imagine como um morcego, percebendo a área ao seu redor através do sonar de ecolocalização desse animal. Entretanto, é claro que por mais que possamos ter algum vislumbre, ainda não conseguimos sentir ou imaginar exatamente a experiência sensorial que esses animais sentem ao emitirem sons de ecolocalização, voar, comer e assim por diante. Podemos compreender melhor a concepção de Nagel através da seguinte passagem:

> Nós não podemos formar mais do que uma concepção esquemática do que é essa experiência de ser um morcego. Por exemplo, podemos atribuir tipos gerais de experiência com base na estrutura e no comportamento do animal. Assim, descrevemos o sonar do morcego como uma forma de percepção tridimensional; acreditamos que os morcegos sentem algumas variantes de dor, medo, fome e desejo, e que possuem outros tipos de percepção mais familiares além do sonar. Mas acreditamos que essas experiências também têm, em cada caso, um caráter subjetivo específico, o qual está além da nossa concepção. (Nagel, 1974, p. 439, tradução nossa).

Dessa forma, é através de argumentações semelhantes a essa que a teoria dos *qualia* busca criticar a tese funcionalista. Como é possível reduzir o funcionamento da mente em funções quando cada indivíduo possui suas próprias interpretações e experiências sensíveis

[16] Frank Jackson foi responsável por introduzir o experimento mental "O Quarto de Mary", em 1982, no artigo Epiphenomenal qualia. O experimento ilustra o conceito de qualia, referindo-se às experiências subjetivas e qualitativas da consciência, argumentando que elas não podem ser completamente explicadas por informações físicas.

que não podem ser explicadas, concebidas ou replicadas por outros seres. Da mesma forma, esse tipo de crítica também percorre o cenário da Inteligência Artificial. Se essas máquinas não são capazes de gerar experiências subjetivas, como poderíamos conceber a possibilidade da confecção de uma inteligência artificial com consciência própria, por mais que ela pudesse simular a inteligência humana, ainda não poderíamos dizer que há uma consciência presente. A máquina operaria como uma espécie de zumbi filosófico[17], ou seja, por mais que simulasse nossa maneira de pensar, não possuiria consciência.

3.4 Materialismo Eliminativista

Outra importante teoria se trata do eliminativismo materialista. Segundo o filósofo José Aparecido Pereira (2015), os dois teóricos mais destacados dessa vertente do materialismo se trata do casal Paul e Patrícia Churchland, que em conjunto publicaram vários estudos sobre essa abordagem. De maneira geral, essa corrente filosófica critica a existência dos *qualia* e de várias outras noções como a mente e a intencionalidade. No eliminativismo materialista, todos esses conceitos são produtos de uma compreensão primitiva, moldados por construções linguísticas e culturais, que não correspondem à realidade das explicações científicas sobre o cérebro e suas funções. Dessa forma, noções como experiência subjetiva, dor e sentimentos são entendidas como processos puramente físicos e operacionais, que podem ser plenamente compreendidos através da investigação das atividades neuronais. Essa perspectiva desafia abordagens filosóficas que tratam a mente como algo distinto dos fenômenos físicos, propondo que a neurociência será capaz de substituir os conceitos do senso comum por explicações mais precisas e científicas. Como Pereira descreve:

> Inicialmente, podemos afirmar que o que sustenta as pretensões dessa corrente de pensamento é a sua convicção sobre a possibilidade de eliminação da nossa linguagem cotidiana (chamada de psicologia popular = folk psychology) utilizada para descrever fenômenos mentais subjetivos. O que se propõe, então, é a substituição desse tipo de linguagem por outro tipo mais alinhado com uma visão científica. (2015, p. 45).

Dessa forma, essa abordagem abre portas para discussões sobre a consciência em inteligências artificiais. Se as experiências subjetivas podem ser reduzidas a processos físicos e computacionais do cérebro, então, uma IA que consiga replicar os eventos neuronais e as funções operacionais do sistema nervoso humano poderia ser considerada capaz de simular a

[17] O termo zumbi filosófico surgiu no artigo *Zombies v. Materialists* (1974) do filósofo Robert Kirk e se tornou mais relevante no cenário da filosofia após aparecer na obra *The Conscious Mind (1996)* de David Chalmers. Basicamente, o conceito diz respeito a um ser fisicamente idêntico ao ser humano, mas sem estados conscientes ou qualquer tipo de experiência subjetiva.

nossa consciência. Dessa maneira, essa corrente filosófica oferece um alicerce teórico para questionar se a distinção entre mentes biológicas e mentes artificiais é realmente necessária ou fundamentada.

3.5 Teoria Computacional da Mente

A teoria computacional da mente (TCM), como é conhecida, é uma abordagem defendida por pensadores como o psicólogo Steven Pinker e sugere que o cérebro humano funciona de maneira semelhante a um computador, processando informações por meio de operações algorítmicas. Para entendermos melhor essa abordagem devemos nos atentar a seguinte passagem de Pinker:

> Quando você telefona para sua mãe em outra cidade, a mensagem permanece a mesma enquanto sai de seus lábios e vai até o ouvido materno, mesmo que fisicamente ela mude de forma, passando de vibrações do ar a eletricidade em um fio, cargas no silício, luz tremulante em um cabo de fibra óptica, ondas eletromagnéticas, voltando então em ordem inversa. Em um sentido semelhante, a mensagem permanece a mesma enquanto sua mãe a repete para seu pai, que está na outra ponta do sofá, depois de ter mudado de forma na cabeça dela, transformando-se em uma cascata de neurônios disparando e substâncias químicas difundindo-se através de sinapses. De modo semelhante, um dado programa pode ser executado em computadores feitos de tubos de vácuo, comutadores eletromagnéticos, transistores, circuitos integrados ou pombos bem treinados, e realiza as mesmas coisas pelas mesmas razões. Esse insight, expresso pela primeira vez pelo matemático Alan Turing, pelos cientistas da computação Alan Newell, Herbert Simon e Marvin Minsky e pelos filósofos Hilary Putnam e Jerry Fodor, hoje em dia é denominado teoria computacional da mente (Pinker, 1998, p. 35).

De maneira geral, a TCM permite que seja possível interpretar o funcionamento da mente humana por meio de conceitos comumente utilizados no campo da computação, como *input*, *output*, processamento e assim por diante. Nesse contexto, a mente humana pode ser compreendida como uma espécie de *software* — isto é, um sistema operacional estruturado por instruções algorítmicas — que é responsável por coordenar o funcionamento de um *hardware*, no caso, o corpo e suas ações. Assim, nossa mente pode ser vista como um conjunto de processos computacionais que recebem informações de entrada (*inputs*), ou seja, nossas percepções, e geram informações de saída (*outputs*), como nossas ações, pensamentos e convicções perante ao mundo. Dessa forma, assim como o materialismo eliminativista, essa abordagem remove as barreiras que impedem que inteligências artificiais possam simular a mente humana. Assim, através dessa teoria seria completamente plausível reduzir toda consciência humana em processos operacionais, isso tornaria viável que máquinas pudessem executar esses mesmos processos e consequentemente simular nossa consciência.

3.6 Os limites da computação

Como descrevem Russell e Norvig (2013) desde as origens da computação, estudiosos como Alan Turing, Kurt Gödel e Alonzo Church revelaram limitações fundamentais dos sistemas formais, trazendo implicações profundas para a tentativa de simular a inteligência humana por meio de máquinas. Em 1936, Alan Turing introduziu o conceito de Máquina de Turing para explorar o *Entscheidungsproblem*, ou problema da decisão, proposto pelo matemático David Hilbert. Hilbert buscava desenvolver um sistema formal universal capaz de resolver qualquer problema matemático de maneira mecânica[18]. Turing (1936), no entanto, demonstrou que tal objetivo não era viável, deixando claro que há problemas matemáticos incomputáveis, ou seja, que nenhum processo mecânico ou computacional pode resolver. Essa descoberta, em conjunto com os teoremas da incompletude de Gödel (1931)[19], expôs limites fundamentais para os sistemas computacionais: há verdades matemáticas que não podem ser provadas dentro de seus próprios sistemas formais. Devido a isso, segundo Russell e Norvig:

> Filósofos como J. R. Lucas (1961) afirmam que esse teorema [teorema da incompletude] mostra que as máquinas são mentalmente inferiores aos seres humanos porque as máquinas são sistemas formais limitados pelo teorema da incompleteza — não podem estabelecer a verdade sobre sua própria sentença de Gödel —, enquanto os seres humanos não têm tal limitação. (Lucas, 1961 apud Russell e Norvig, 2013, p.1132).

Dessa forma, essa limitação tem implicações significativas para a IA. De modo que, se a mente humana opera em um nível que sistemas computacionais não podem alcançar, seja por causa de processos intuitivos ou incomputáveis, a tentativa de simular plenamente a inteligência humana enfrenta obstáculos teóricos talvez intransponíveis. Além disso, Russell e Norvig (2013) destacam que teóricos como Roger Penrose (1989, 1994) reiteraram a afirmação de Lucas (1961), com algumas variações, sugerindo que a consciência humana pode depender de fenômenos quânticos que ultrapassam as capacidades da computação. Dessa maneira, essa perspectiva parece desafiar as afirmações da teoria computacional da mente, enquanto estabelece um diálogo positivo com teorias como os *qualia*.

[18]De acordo com Ferreira (2019), a expressão "maneira mecânica" pode ser entendida como um processo inteiramente sistemático e automatizado, que segue um conjunto de regras predefinidas, sem depender da criatividade ou intuição humana.

[19]A prova dos Teoremas da Incompletude de Gödel é consideravelmente extensa e envolve conceitos avançados de lógica matemática e teoria dos sistemas formais. Devido a isso, não é possível incluí-la neste trabalho. Para os leitores interessados, recomenda-se a obra *Gödel, Escher, Bach: Um Entrelaçamento de Gênios Brilhantes*, do físico Douglas Hofstadter, que oferece uma explicação acessível e detalhada dos teoremas de Gödel.

3.7 Apontando caminhos

Desse modo, após analisar algumas das diversas teorias que buscam resolver o problema mente-corpo, é necessário ter a noção de que nenhuma delas está livre de críticas e que todas possuem um problema ou outro na sua abordagem. Dessa forma é possível notar que ainda não parece ser uma resposta sólida afirmar que uma inteligência artificial poderia simular a consciência ou o modo de pensar humano. Em contrapartida, também não é completamente plausível garantir que em momento algum do futuro uma máquina poderá ser capaz de simular a mente humana de forma perfeita. Tendo isso em vista, é importante compreender que mesmo nesse cenário, investigar essas teorias têm sua devida relevância porque nos apresenta os trajetos já traçados até o momento e nos auxilia a pavimentar os novos caminhos que poderão nos fornecer as respostas corretas em algum instante.

4. SUPERINTELIGÊNCIA: SUAS FORMAS, PROBLEMAS E POSSÍVEIS SOLUÇÕES

Agora, avançaremos para um território ainda mais complexo e desafiador. Após compreendermos as distinções e os problemas das inteligências artificiais fraca e forte, analisarmos o problema mente-corpo e as limitações da computação, avançaremos para explorar o conceito de "superinteligência". Nesse ponto, não estamos apenas lidando com a questão de se máquinas podem ou não simular a inteligência humana, mas sim enfrentando a possibilidade de sistemas que transcendam amplamente as capacidades cognitivas humanas, trazendo implicações que extrapolam os limites do entendimento atual. De acordo com o filósofo Nick Bostrom (2018) a superinteligência de máquina pode ser compreendida como um agente intelectual que supera o nível humano na maioria das suas capacidades cognitivas, ela pode resolver problemas extremamente complexos além de possuir uma inteligência inalcançável para cérebros humanos. Dessa forma, a superinteligência supera até mesmo a IA forte, com níveis de compreensão e eficiência em um patamar bastante avançado. Com isso em mente, baseado nos estudos de Bostrom apresentados em seu livro *Superinteligência* (2018), iremos analisar os possíveis tipos de superinteligência, suas capacidades, os problemas que elas podem causar a humanidade e alguns métodos de prevenção.

4.1 Tipos de superinteligência

Entre os tipos de superinteligência apresentados por Bostrom (2018), em um primeiro momento, destaca-se o modelo de superinteligência rápida. As máquinas desta categoria

seriam capazes de realizar qualquer atividade que exigisse níveis humanos de inteligência, mas com uma velocidade surpreendente. Para compreendermos melhor as capacidades desse sistemas podemos nos atentar a seguinte passagem:

> O exemplo mais simples da superinteligência rápida seria uma emulação completa do cérebro executada em um hardware veloz. Uma emulação operando com uma velocidade 10 mil vezes maior que a de um cérebro biológico seria capaz de ler um livro em alguns segundos e escrever uma tese de doutorado em uma tarde. Se a velocidade fosse 1 milhão de vezes maior, uma emulação poderia realizar o trabalho intelectual de um milênio em apenas um dia de trabalho. (Bostrom, 2018, p. 87).

Devido a essa capacidade cognitiva extremamente veloz, seria mais vantajoso que uma máquina com essa capacidade operasse em conjunto com outras de igual rapidez, em vez de colaborar com o relativamente lento cérebro humano. Esse tipo de inteligência artificial poderia processar e otimizar processos de pesquisa científica de forma extremamente eficiente, realizando experimentos e interpretando resultados com uma precisão muito superior à humana, acelerando o avanço de áreas como a saúde, a tecnologia e a engenharia.

Além desse modelo, o filósofo também descreve a superinteligência coletiva, que se desenvolveria a partir da união de um grande número de inteligências menores, cada uma especializada em diferentes áreas. Essa colaboração formaria um sistema complexo e inteligente, capaz de superar o desempenho cognitivo humano. Uma superinteligência coletiva poderia, por exemplo, dedicar-se ao monitoramento de ecossistemas, à engenharia ambiental e ao desenvolvimento de tecnologias sustentáveis, criando soluções inovadoras para preservar o meio ambiente, otimizar o uso dos recursos naturais e reduzir os impactos negativos das atividades humanas na natureza. Como o filósofo afirma, “A inteligência coletiva se sobressai na resolução de problemas que podem ser facilmente divididos em partes, de modo que as soluções para esses subproblemas possam ser encontradas paralelamente e verificadas de maneira independente” (2018, p. 88). Esse tipo de sistema poderia evoluir tanto pela incorporação de mais inteligências menores ao conjunto quanto pela inclusão de inteligências altamente especializadas, atuando de forma integrada em diferentes campos cognitivos.

Por fim, Bostrom nos apresenta a ideia de uma superinteligência de qualidade, que não seria apenas mais rápida, mas também intelectualmente mais avançada que qualquer ser humano. Esse tipo de máquina seria capaz de enfrentar as tarefas cognitivas mais complexas e relevantes, criando soluções e resolvendo problemas que até mesmo a pessoa mais inteligente do planeta não poderia imaginar. Com uma máquina dessa, seria possível projetar naves

autossustentáveis, desenvolver rotas interplanetárias e criar ecossistemas e atmosferas artificiais que nos permitiriam colonizar e habitar outros planetas. Os níveis que uma superinteligência dessa categoria pode chegar beira o inimaginável, problemas que nem mesmo as outras formas de superinteligência conseguiriam resolver poderiam ser superados por uma superinteligência de qualidade. Como Bostrom afirma nesta passagem:

> Em algum sentido vago, a superinteligência de qualidade seria a mais capaz de todas as formas, na medida em que fosse capaz de perceber e resolver problemas que estão, para todos os propósitos práticos, além do alcance direto da superinteligência rápida e da superinteligência coletiva (Bostrom, 2018, p. 95).

De maneira geral, o autor considera as "superinteligências" uma possibilidade concreta, fundamentada no avanço exponencial da inteligência artificial, com ênfase na "explosão de inteligência"[20], onde a partir do momento em que uma máquina se tornasse superinteligente, ela poderia aprimorar suas próprias capacidades de forma autônoma, em um ciclo contínuo e acelerado gerando novas máquinas capazes de desenvolver uma inteligência ainda superior. Bostrom acredita que caso algo semelhante ocorra o desenvolvimento dessas superinteligências não ocorrerá de maneira lenta ou gradual, como as revoluções industriais ou agrícolas[21], mas sim de forma acelerada, em questão de meses, dias ou até mesmo minutos, oferecendo pouco tempo de resposta para que medidas de intervenção ou regulamentações eficazes sejam criadas. Esse processo, como veremos a seguir, levanta questões críticas sobre controle e segurança, já que uma superinteligência poderia ultrapassar não apenas o desempenho humano, mas também nossas capacidades de governá-la, representando um risco potencial para a humanidade.

4.2 O problema do alinhamento e o problema do controle

Após analisarmos os possíveis tipos de superinteligência, é necessário investigar quais adversidades que essas entidades podem nos causar caso cheguem a ser desenvolvidas em algum momento. Dessa forma, iremos examinar o problema do alinhamento, em que se analisa o cenário onde os objetivos de uma superinteligência não estejam alinhados com os objetivos humanos e o problema do controle, em que em uma situação de explosão de

[20] Segundo Bostrom (2018), a explosão de inteligência pode ser compreendida como um evento no qual, em um curto espaço de tempo, o nível de inteligência de um sistema passaria de capacidades relativamente modestas de cognição (níveis sub-humanos) a uma superinteligência radical através de uma rápida sequência de auto aperfeiçoamentos recursivos capazes de multiplicar as capacidades de uma IA.

[21] Bostrom afirma que "(...) um cenário de transição lenta é improvável. Se e quando uma partida ocorrer, tudo indica que ela será explosiva" (2018, p. 104).

inteligência, uma IA poderia se desenvolver de maneira extremamente rápida de modo a superar a capacidade do controle humano sobre a máquina.

Nesse contexto, para o filósofo Nick Bostrom (2018) o problema do alinhamento surge como uma das principais preocupações, ao expor os riscos e as dificuldades para garantir que uma inteligência artificial avançada tenha os mesmos objetivos, valores e interesses que os humanos. Essa tarefa já é particularmente complexa devido à falta de um conjunto de normas morais universalmente aceito. No entanto, à medida que as máquinas evoluem e se tornam mais inteligentes, torna-se cada vez mais difícil assegurar que suas decisões e ações permaneçam alinhadas a padrões éticos e não representem uma ameaça à humanidade. Um exemplo ilustrativo dessa dificuldade seria a criação de uma superinteligência programada para otimizar o uso de recursos naturais. Embora a intenção fosse beneficiar a humanidade, a IA poderia concluir que a forma mais eficiente de atingir esse objetivo seria explorar os recursos de maneira a ignorar qualquer preocupação com o meio ambiente, destruindo ecossistemas e esgotando fontes vitais, com consequências catastróficas para o planeta. Esse cenário — claramente indesejado por seus desenvolvedores — evidencia o perigo de uma interpretação literal e descontextualizada de comandos[22], ressaltando a importância de garantir que os sistemas de IA avancem com objetivos alinhados a intenções humanas mais amplas e seguras.

Além do problema do alinhamento, Bostrom (2018) também nos apresenta o problema do controle. Esse desafio surge da possibilidade de uma inteligência artificial avançada se aperfeiçoar de forma autônoma, gerando uma explosão de inteligência. Isso significa que a IA poderia criar sistemas e máquinas ainda mais desenvolvidos, alcançando um nível de inteligência muito além da capacidade de qualquer supervisão ou domínio humano. Nesse cenário, estaríamos diante de uma superinteligência sem controle, representando um sério risco para a nossa existência e nos conduzindo a um futuro extremamente perigoso e imprevisível. É preciso compreender que os vastos recursos e habilidades de uma superinteligência artificial e sua capacidade de aprendizado e adaptação acelerados dificultam não apenas a previsão de suas ações, mas também a implementação de medidas preventivas para evitar danos. Devido a isso, métodos eficazes para controlar o desenvolvimento e a operação desses sistemas devem ser pensados previamente para garantir que as ações dessas máquinas permaneçam dentro dos limites seguros para a humanidade.

[22] Veja *IA fabricante de clipes de papel e Hipótese de catástrofe de Riemann* em Bostrom (2018, p. 173) para mais exemplos.

4.3 Procedimentos de controle e métodos de motivação

As metodologias que serão analisadas a seguir foram propostas por Bostrom (2018) e buscam controlar as capacidades de uma superinteligência e alinhar seus interesses aos desejos e valores humanos durante o seu desenvolvimento. O objetivo desses métodos trata-se de evitar que uma explosão de inteligência não planejada ocorra e também que a humanidade perca o controle de uma poderosa entidade superinteligente. No entanto, como veremos adiante, essas estratégias ainda apresentam lacunas significativas, que podem comprometer sua eficácia e deixar margem para riscos imprevistos.

O primeiro método que iremos investigar é o de controle de capacidade, que visa restringir o que uma superinteligência pode fazer caso não obedeça às ordens humanas. Desse modo, para reduzir o potencial de dano desse sistema, poderíamos confiná-lo fisicamente em um recipiente, caixa ou cômodo, de acordo com o tamanho da máquina, limitando seu acesso ao mundo externo. No entanto, Bostrom (2018) alerta que esse método possui vulnerabilidades, o material que compõe o espaço confinado precisaria bloquear qualquer transmissão de rádio ou acesso à internet, impedindo que a superinteligência encontre meios de influenciar o mundo externo. Ainda assim, há risco de a máquina descobrir formas de contornar essas barreiras.

Outro procedimento que poderíamos utilizar se trata dos métodos de incentivo, que por sua vez buscam recompensar a superinteligência por adotar comportamentos alinhados aos valores e normas humanas, ensinando-a que colaborar é vantajoso. Como Bostrom descreve, “Um outro tipo de método de incentivo envolveria a criação de uma configuração na qual a IA pudesse ser recompensada e penalizada pelo projeto que a criou e, dessa forma, seria incentivada a agir de acordo com os interesses de seu responsável.” (2018, p. 185). O problema, contudo, é que a IA poderia simular cooperação apenas para receber os incentivos, sejam eles quais forem sem realmente aderir aos valores humanos. Isso tornaria o comportamento da IA imprevisível, pois ela poderia manipular as recompensas sem adotar, de fato, a ética esperada.

De mesmo modo, mais uma técnica que podemos analisar é o método de inibição, que limita o hardware e as capacidades da IA para controlar seu desenvolvimento. Essa estratégia busca restringir o banco de dados e a capacidade de processamento da inteligência artificial,

de modo a liberar gradualmente seu poder conforme a máquina não apresentasse riscos para a humanidade. Entretanto, como Bostrom adverte:

> Até mesmo sem qualquer base de conhecimento determinada, uma mente suficientemente superior pode ser capaz de aprender muito simplesmente fazendo uma introspecção sobre o funcionamento de sua própria psique — as opções de projeto refletidas em seu código fonte e as características físicas de seus circuitos.(2018, p. 188).

Dessa forma, se a superinteligência alcançasse níveis mais avançados de inteligência, poderia encontrar formas de contornar essas limitações ou otimizar seu funcionamento, escapando do controle.

Um outro método a ser destacado é o dos detonadores em que a IA é monitorada por mecanismos de segurança que detectam atividades irregulares e, se necessário, interrompem o fornecimento de energia para desligá-la. Dessa forma, através desta técnica, como Bostrom expõem, "Os programadores podem então examinar gravações das atividades suspeitas quando quiserem e fazer mudanças nas bases de código ou no ambiente da IA antes de reiniciá-la. Se acharem que o problema é muito grave, eles poderiam decidir abortar todo o projeto." (2018, p. 189). Entretanto, esse método enfrenta o risco de uma falha nos sistemas de monitoramento, o que permitiria à IA realizar ações perigosas antes que qualquer intervenção ocorra, ou até mesmo encontrar meios de desativar os detonadores para evitar seu desligamento.

Além dos métodos de controle de capacidade que investigamos, Bostrom também propôs os procedimentos de seleção de motivação. As técnicas dessa abordagem foram desenvolvidas para garantir que a superinteligência aja de forma segura e alinhada com os objetivos e valores humanos, de modo que não prejudique a humanidade. Dessa maneira, essas estratégias empenham-se em criar um conjunto de motivações alinhadas aos nossos interesses. Assim, nessas metodologias, busca-se definir os valores internos e os objetivos da inteligência artificial.

Com isso em mente, o primeiro procedimento a ser analisado é a especificação direta. Nesse modelo, a superinteligência seria programada com um conjunto detalhado de regras e objetivos, cuidadosamente elaborados para garantir que ela se comporte de forma ética e segura. No entanto, o problema com esse método é que pode ser extremamente difícil antecipar todas as situações possíveis em que a IA atuará. Caso surjam circunstâncias imprevistas, a IA pode interpretar as regras de forma literal ou inadequada, gerando

consequências inesperadas e possivelmente perigosas. Devido a isso, Bostrom descreve que "É provável que seja humanamente impossível formular explicitamente um conjunto de regras altamente complexas de forma detalhada, aplicá-las em diversas circunstâncias e ainda conseguir fazer isso da maneira correta na primeira tentativa de implementação." (2018, p. 191).

Devido às complicações observadas no método passado, outra abordagem proposta por Bostrom (2018) é a domesticidade, que limita as ambições e os objetivos da superinteligência. Em vez de projetar uma inteligência artificial com um enorme potencial transformador, seriam atribuídos a ela objetivos mais simples e específicos, permitindo que seus desenvolvedores avaliem seu alinhamento ético antes de conceder mais autonomia. O problema aqui é que restringir os objetivos da IA também limita seu desempenho, o que pode reduzir seu potencial e tornar seu desenvolvimento menos eficiente, Podendo ainda frustrar a IA, caso ela desenvolva uma motivação de superação[23].

Outro método a ser considerado é o da normatividade indireta, um procedimento no qual os programadores permitem que a inteligência artificial desenvolva seu próprio conjunto de regras normativas durante o processo de aprendizado, com base em investigações influenciadas por valores humanos. Como Bostrom afirma, "O seu potencial reside no fato de que ela permitiria que deixássemos a cargo da superinteligência a maior parte do trabalho cognitivo mais difícil exigido para a elaboração da especificação direta de um objetivo final apropriado."(2018, p. 193). No entanto, o risco dessa abordagem é que a IA pode interpretar ou desenvolver normas que sejam apenas parcialmente alinhadas aos valores humanos, ou até mesmo criar interpretações que contradigam os objetivos dos programadores, resultando em ações que não refletem a ética esperada.

Por fim, o método da ampliação, segundo Bostrom (2018) busca desenvolver uma IA com valores e motivações já aceitáveis e familiares aos desejos humanos, de forma que, à medida que ela se torna mais avançada, seus princípios continuem alinhados com a ética humana. O problema com essa abordagem é que, à medida que a inteligência artificial se aperfeiçoa, seus valores iniciais podem ser modificados ou evoluir de maneira não prevista, o que poderia afastá-la dos princípios éticos originais que se pretendia manter ao longo de seu desenvolvimento.

[23] Capacidade de uma inteligência artificial transcender as restrições impostas a ela, seja para atingir objetivos mais amplos ou para melhorar suas próprias capacidades.

Em síntese, Apesar dessas discussões sobre os problemas do controle e do alinhamento, bem como os métodos de prevenção e suas brechas, parecerem algo que saiu de uma distopia tecnológica e que está bem distante da nossa realidade, não podemos esquecer que em séculos ou até mesmo décadas atrás jamais imaginaríamos que a humanidade possuiria as tecnologias que existem atualmente. Além disso, essas questões devem ser analisadas de fato com muita antecedência, antes mesmo do surgimento de qualquer vestígio de uma superinteligência ou de uma IA forte. Caso contrário, quando nos depararmos com um problema real, nossos métodos podem ser primitivos e ineficientes para lidar com essas situações. É importante ressaltar que a IA trabalha de forma opaca e muitas vezes nem mesmo seus programadores sabem explicar a trajetória que esses sistemas desenvolvem para chegar nas suas conclusões. Toda essa dificuldade humana para interpretar as decisões e escolhas dessas máquinas já nos causaram danos em eventos passados. Na medida que esses sistemas evoluem as nossas perdas podem ser cada vez maiores, caso os devidos métodos de prevenção não sejam desenvolvidos e testados de maneira prévia.

5. CONSIDERAÇÕES FINAIS

Este trabalho buscou não apenas explorar as questões filosóficas associadas à inteligência artificial, mas também enriquecer, dentro de suas limitações, o debate acadêmico brasileiro, incentivando futuras investigações sobre as implicações éticas, existenciais e tecnológicas que cercam esse campo. Em suma, o que precisamos ter em mente ao investigar cada capítulo desse trabalho se trata da forma como a inteligência artificial, desde suas origens até os dias atuais, vem modificando a forma como compreendemos a consciência, o cérebro e os avanços tecnológicos. Além disso, ao observar os conceitos de IA fraca e IA forte, podemos perceber que o desenvolvimento desses sistemas está intrinsecamente relacionado a diversas questões filosóficas, de modo que traz à tona até mesmo debates como o problema mente-corpo, que ganha uma nova forma de se apresentar sob a luz das máquinas. A reinterpretação dessa questão filosófica não só deixa claro que o nosso entendimento sobre a mente ainda é bastante limitado, como também evidencia a necessidade de repensarmos esse e diversos conceitos relacionados, de modo a incluir as potencialidades das máquinas. Ainda nesse contexto, é importante ressaltar que embora os avanços em inteligência artificial sejam notáveis, questões fundamentais, como as limitações matemáticas dos sistemas formais e a complexidade da mente humana, continuam a desafiar as nossas ambições de simular plenamente a consciência. Por outro lado, o conceito de superinteligência nos leva a especular um futuro em que a IA não apenas se iguale, mas também supere as nossas capacidades cognitivas, o que, apesar de fabuloso, também nos traz novas preocupações. Dentre elas, investigamos o problema do alinhamento e o problema do controle, que expõem a necessidade de estruturar um desenvolvimento de inteligências artificiais que não sejam somente eficientes, mas também éticas e seguras. Desse modo, enquanto o problema do alinhamento nos demonstra que os sistemas de IA precisam estar integrados aos valores humanos, o problema do controle nos apresenta a complexidade de se garantir que esses sistemas nos obedeçam durante seus avanços e não ameacem a humanidade.

Dessa forma, por meio deste trabalho, buscou-se oferecer uma perspectiva que conecta o avanço tecnológico à reflexão filosófica, evidenciando como esses campos se complementam e como essa relação pode nos ajudar a compreender os desafios e as potencialidades da IA. Devido a isso, para que o desenvolvimento da inteligência artificial ocorra de forma segura e alinhada aos valores humanos, é indispensável preservar a interdisciplinaridade, como a vista no Simpósio de Hixon em 1948 e no Workshop da

Dartmouth College em 1956. A colaboração entre filósofos, cientistas da computação, psicólogos, neurocientistas e pesquisadores de diversas áreas é fundamental para garantir que esse avanço tecnológico seja ético e humanizado. Por fim, embora este trabalho tenha abordado questões relevantes, reconhecemos que muitos temas relacionados à IA permanecem fora do escopo ou pouco examinados, mas que podem ser explorados em pesquisas futuras, ampliando ainda mais a compreensão sobre os impactos da inteligência artificial e contribuindo para o desenvolvimento de soluções éticas.

REFERÊNCIAS

APOLLONIO, R. **Os Argonautas.** Tradução de José Maria da Costa e Silva. Lisboa: Imprensa Nacional, 1852.

BOSTROM, N. **Superinteligência: caminhos, perigos e estratégias para um novo mundo.** Tradução de Aurélio Antônio Monteiro et al. Rio de Janeiro: DarkSide Books, 2018.

CHALMERS, D. J. **The Conscious Mind.** New York: Oxford University Press, 1996.

CHURCHLAND, P. M. (1981). **Eliminative Materialism and the Propositional Attitudes.** *The Journal of Philosophy*, *78*(2), 67–90. Disponível em: https://doi.org/10.2307/2025900. Acesso em: 8 de nov. 2024.

DESCARTES, R. **Discurso do Método.** Tradução de Maria Ermantina Galvão. São Paulo: Martins Fontes, 2001.

FERREIRA, F. "**O problema da decisão e a máquina universal de Turing".** In Alan Turing. Cientista Universal, editado por Espírito Santo, José, 55-84. Braga, Portugal: UMinho Editora, 2019.

FREETH, T. et al. **Calendars with Olympiad display and eclipse prediction on the Antikythera Mechanism**. Nature 454, 614–617, 2008. Disponível em: https://doi.org/10.1038/nature07130 . Acesso em: 8 de nov 2024

HOFSTADTER, D. R. **Gödel, Escher, Bach – Um entrelaçamento de gênios brilhantes.** Tradução de José Viegas Filho. Brasília, DF: Ed. da UnB, 2001.

JACKSON, F. **Epiphenomenal qualia. Philosophical Quarterly**, 32, 127–136, 1982.

KIRK, R.; SQUIRES, R. **Zombies v. Materialists.** *Proceedings of the Aristotelian Society, Supplementary Volumes*, *48*, 135–163, 1974. Disponível em: http://www.jstor.org/stable/4106864. Acesso em: 8 de nov. 2024.

LUCAS, J. R. **Minds, machines, and Gödel.** Philosophy, 36, 1961.

MASLIN, K. **Introdução à Filosofia da Mente.** Tradução de Fernando José R. da Rocha. São Paulo: Artmed, 2009.

MCCULLOCH, W., PITTS, W., **"A Logical Calculus of the Ideas Immanent in Nervous Activity"**, The Bulletin of Mathematical Biophysics, Vol. 5, No. 4, pp. 115 – 133, 1943.

NAGEL, T. **"What is it like to be a bat?"** The Philosophical Review, Vol. LXXXIII, Nº 4, p. 435-50, 1974.

PENROSE, R. **Shadows of the Mind**. Oxford University Press, 1994.

PENROSE, R. **The Emperor's New Mind**. Oxford University Press, 1989.

PEREIRA, J. A. **Uma Análise sobre Materialismo Eliminativo a Partir do Pensamento de Nagel**. Trans/form/ação, 38(3), 43–56, 2015. Disponível em: https://doi.org/10.1590/S0101-31732015000300004. Acesso em: 8 de nov. 2024.

PINKER, S. **Como a Mente Funciona.** São Paulo: Companhia das Letras, 1998.

ROCHA, E. M.. (2004). **Animais, homens e sensações segundo Descartes.** Kriterion: Revista De Filosofia, 45(110), 350–364. Disponível em: https://doi.org/10.1590/S0100-512X2004000200008. Acesso em: 9 de nov. 2024.

RUSSELL, S.; NORVIG, P. **Inteligência Artificial: uma abordagem moderna.** Tradução de Regina Célia Simille de Macedo. Rio de Janeiro: Campus, 2013.

SEARLE, J. R. Minds, brains and programs. **Behavioral and Brain Sciences**, v. 3, n. 3 p. 417-424, set. 1980. Disponível em: https://web-archive.southampton.ac.uk/cogprints.org/7150/1/10.1.1.83.5248.pdf. Acesso em: 8 nov. 2024.

TEIXEIRA, J.F. **O que é inteligência artificial.** Livro eletrônico: e-galáxia, 2019

TEIXEIRA, J. F. **O que é inteligência artificial.** Rio de Janeiro: Paulus, 2009.

TURING, A. M. **Computing Machinery and Intelligence**. Mind, v. 59, n. 236, out. 1950. p. 433-460. Disponível em: https://academic.oup.com/mind/article/LIX/236/433/986238. Acesso em: 08 nov. 2024.

TURING, A. M. **On Computable Numbers, with an Application to the Entscheidungsproblem.** Disponível em: https://www.cs.virginia.edu/~robins/Turing_Paper_1936.pdf. Acesso em: 8 de nov. 2024.

WIENER, N. **Cybernetics, or Control and Communication in the Animal and The Machine.** 2ª Ed. Cambridge, Massachutts: MIT Press, 1961.