MÁQUINAS QUE PENSAM: CIÊNCIA OU METAFÍSICA?

Paulo Uiris da Silva Gomes
Rosângela Araújo Darwich

Resumo: Máquinas podem pensar? É uma questão discutida pelos estudiosos da Inteligência Artificial desde os seus primórdios, com autores favoráveis e outros críticos, seja nos campos da filosofia, das ciências cognitivas ou da ciência da computação. O objetivo desse trabalho é analisar essa questão central - e o projeto associado a ela de criar máquinas inteligentes - a partir do enfoque do seu *status* científico. Isto é, questiona-se, a partir da teoria da ciência do filósofo Karl Popper, se o projeto das máquinas inteligentes seria, realmente, científico e se seria capaz de fazer previsões científicas. Tal questionamento é importante considerando os impactos atuais e potenciais dessa tecnologia. Nossa conclusão é que a Inteligência Artificial possui elementos das ciências empíricas, mas também possui elementos metafísicos, sendo mais adequado categorizá-la como um "Projeto Metafísico de Pesquisa" e, por isso, sua capacidade preditiva é limitada.

Palavras-chave: Inteligência Artificial. Filosofia da Ciência. Falseabilidade. Karl Popper. Previsões Científicas.

MACHINES THAT THINK: SCIENCE OR METAPHYSICS?

Abstract: Can Machines Think? It is a question discusses by Artificial Intelligence scholars since its beginnings, with some favorable authors and critics, whether in the fields of philosophy, cognitive sciences or computer science. This work aims to analyze this central questions – and its associated project of creating intelligent machines – from the perspective of its scientific status. That is, based on philospher Karl Popper's theory of science, it is questioned whether the project of intelligence machines is really scientific and whether it is capable of making scientific predictions. Such investigation is important considering the current and potential impacts of this technology. Our conclusion is that Artificial Intelligence has elements of empiricial sciences, but it also has metaphysical elements, making it more appropriate to categorize it as a "Mataphysicial Research Programme" and, therefore, its predictive power is limited.

Keywords: Artificial Intelligence. Philosophy of Science. Falseability. Karl Popper. Scientific Predicitions.

MÁQUINAS QUE PIENSAN: ¿CIENCIA O METAFÌSICA?

Resumen: ¿Pueden las máquinas pensar? Es un tema discutido por los estudiosos de la Inteligencia Artificial desde sus inicios, con autores favorables y otros críticos, ya sea en los campos de la filosofía, las ciencias cognitivas o la informática. Nuestro objetivo en este trabajo es analizar este tema central - y el proyecto asociado de creación de máquinas inteligentes - desde la perspectiva de su estatus científico. O sea, nos preguntamos, basándonos en la teoría de la ciencia del filósofo Karl Popper, si el proyecto de máquinas inteligentes sería realmente científico y si sería capaz de hacer predicciones científicas. Este cuestionamiento es importante considerando los impactos actuales y potenciales de esta tecnología. Nuestra conclusión es que la Inteligencia Artificial tiene elementos propios de las ciencias empíricas, pero también tiene elementos metafísicos, por lo que es más apropiado categorizarla como un "Proyecto de Investigación Metafísica" y, por tanto, su capacidad predictiva es limitada.

Palabras-clave: Inteligencia artificial. Filosofía de la Ciencia. Falsificabilidad. Karl Popper. Predicciones científicas.

1. INTRODUÇÃO: MÁQUINAS PODEM PENSAR?

A questão sobre se as máquinas podem pensar tem sido um tema de debate desde os primórdios da inteligência artificial (IA). Esta questão, popularizada por Alan Turing em seu ensaio seminal "Computing Machinery and Intelligence", continua a ser discutida fervorosamente em diversos campos, desde a filosofia até a ciência da computação.

Os defensores da posição de que as máquinas podem pensar frequentemente baseiam seu argumento na capacidade das máquinas de realizar tarefas que, tradicionalmente, eram consideradas exclusivas da mente humana. Com os avanços na IA, vemos exemplos de sistemas que podem realizar tarefas complexas, como reconhecimento de fala, tradução automática, e até mesmo jogar xadrez ou Go em níveis que desafiam os melhores jogadores humanos. Esses feitos levam alguns a argumentar que as máquinas são capazes de processar informações e tomar decisões de forma semelhante à mente humana, o que constituiria uma forma de pensamento.

No entanto, os críticos dessa visão frequentemente levantam a questão da natureza da consciência e da experiência subjetiva. Eles argumentam que, apesar de as máquinas poderem realizar tarefas complexas, isso não implica necessariamente que elas possuam consciência ou uma verdadeira compreensão do que estão fazendo. Por exemplo, um sistema de IA que joga xadrez pode ser extremamente habilidoso, mas não possui a experiência consciente do jogo ou a capacidade de apreciar sua beleza estética, como um ser humano faria.

Outro ponto de contenda é a questão da originalidade e da criatividade: os críticos argumentam que, mesmo que as máquinas sejam capazes de gerar resultados inovadores, como composições musicais ou obras de arte visual, essas criações são essencialmente imitações de padrões existentes e não refletem uma verdadeira capacidade de pensamento criativo. Além disso, a

questão da intencionalidade é frequentemente levantada. As máquinas podem executar tarefas com eficiência, mas isso não implica que elas tenham intenções ou propósitos semelhantes aos humanos. Por exemplo, um robô que realiza tarefas domésticas não o faz por vontade própria ou por um desejo de ajudar, mas simplesmente porque foi programado para fazê-lo.

Essencialmente, a questão de se as máquinas podem pensar depende da definição de "pensamento" que estamos usando. Se “pensamento” for definido estritamente como a capacidade de processar informações e tomar decisões com base nessas informações, então é plausível argumentar que as máquinas podem, de fato, pensar. No entanto, se considerarmos o pensamento como algo mais abrangente, que inclui consciência, experiência subjetiva, originalidade e intencionalidade, então as máquinas ainda têm um longo caminho a percorrer antes de poderem ser consideradas verdadeiramente pensantes. Em última análise, o debate sobre se as máquinas podem pensar é multifacetado e complexo, e é improvável que haja uma resposta definitiva no futuro próximo. À medida que a tecnologia continua a avançar e nossa compreensão da mente humana se aprofunda, é provável que novas perspectivas e insights surjam, enriquecendo ainda mais esse debate fascinante.

Talvez o leitor não tenha suspeitado, mas o texto acima foi escrito por uma máquina (mais especificamente, o *chatbot* ChatGPT 3.5): fato que, até pouquíssimo tempo atrás, deixaria perplexo qualquer pesquisador de Inteligência Artificial (IA doravante) ou qualquer indivíduo interessado no assunto. Um comando e alguns segundos foram o suficiente para que a máquina, em sua primeira tentativa, produzisse-o fluidamente. Convido o leitor a julgar a qualidade do texto e, também, a pensar o que seria necessário para que um de nós, humanos, escrevesse algo próximo a isto. Pode-se dizer, de modo geral, que é um texto bem estruturado, bem escrito, coeso, coerente e, essencialmente, correto acerca de seu conteúdo. Poder-se-ia até

acreditar que fora escrito por algum especialista da área (como, talvez, alguns dos leitores tenham acreditado até a revelação deste parágrafo).

Máquinas que argumentam são impressionantes (e muito úteis) – o que é corroborado por sua rápida e crescente popularidade. Até menos de uma década atrás, pertenciam apenas aos intangíveis planos da especulação e da ficção: o que parecia impossível ou, no melhor dos casos, longínquo, está se tornando, em alguma medida, realidade. Mas, como saber se tais máquinas são, de fato, inteligentes ou se realmente pensam? Provavelmente o meio mais adequado para responder a essas questões seria utilizando as teorias, métodos e experimentos da ciência. No entanto, seria a IA, de fato, uma ciência? Ou suas práticas teriam mais proximidade da metafísica? É o que pretendemos investigar.

2. UM BREVE HISTÓRICO DO PROJETO DAS MÁQUINAS PENSANTES

2.1 Máquinas de Computação e Inteligência

Historicamente, quem trata dessas questões relacionadas à Inteligência, ao Pensamento e seus correlatos são os campos de estudo da Filosofia (da mente), da Psicologia e das Ciências Cognitivas. Não obstante, um dos primeiros e mais notáveis teóricos a refletir seriamente sobre a aplicação destas às máquinas foi o matemático e cientista da computação inglês Alan Turing, “o primeiro a fazer pesquisas substanciais no campo que hoje chamamos de Inteligência Artificial ou IA. [...] Turing pensava sobre inteligência das máquinas ao menos desde 1941” (COPELAND, 2004, p. 353, tradução nossa), sendo, por isso, considerado um dos pais da IA, com contribuições importantíssimas que repercutem até os dias atuais.

Uma delas é a investigação presente em seu artigo seminal intitulado “Computing Machinery and Intelligence”, publicado na revista filosófica *Mind* em 1950, que trata da questão central do campo da Inteligência Artificial como um todo: “eu proponho que consideremos a questão: ‘máquinas podem pensar?’” (TURING, 1950, p. 433, tradução nossa). Sua perspectiva era que

tal questão “é por demais sem sentido para merecer uma discussão” (TURING, 1950, p. 442). e que seria “absurdo” debater os significados dos termos “máquina” e “pensamento”, por isso, sugere substituir esta questão por outra, menos ambígua, a saber: o que aconteceria se uma máquina jogasse o “jogo da imitação”[1]? Ou seja, e se uma máquina fosse tão bem-sucedida na imitação de um ser humano a ponto de não ser possível diferenciá-los?

Para Turing, essa nova forma do problema tem como vantagem a sua clareza e delimitação, dessa forma, é possível inquirir a máquina sobre qualquer assunto dentro do âmbito intelectual humano e avaliar a sua inteligência a partir da sua resposta (o que foi sugerido ao leitor realizar acerca do texto do ChatGPT). Muitas objeções foram levantadas contra a proposição de Turing (e sobre o jogo da imitação) como parâmetro de avaliação da inteligência das máquinas e ele mesmo antecipou algumas dessas críticas e “visões contrárias” à sua (TURING, 1950, p. 442-454). Ainda assim, suas ideias seminais tiveram profundo impacto na formação do campo de estudo e continuam a reverberar nas práticas e reflexões presentes.

2.2 Máquinas que Simulam Qualquer Característica da Inteligência

Das reflexões iniciais de Turing até os dias atuais, muito aconteceu no campo da IA: muito foi teorizado, escrito, discutido, feito e testado – para muito além do que poderia enxergar e imaginar o eminente teórico britânico. Em toda essa trajetória multifacetada, instável, repleta de sonhos e fracassos igualmente grandiosos, é preciso destacar como um de seus momentos mais importantes a Conferência de Dartmouth nos Estados Unidos em 1956.

Mais especificamente, a Conferência de Dartmouth (*The Dartmouth Summer Research Project on Artificial Intelligence*) marca a fundação da Inteligência Artificial enquanto campo de estudo distinto e característico – não meramente um subcampo de outra disciplina –, e é onde se estabelece o termo “Inteligência Artificial” para designar possíveis máquinas inteligentes. A proposta para tal conferência data de 1955 (MCCARTHY *et al.* 1955), tem

autoria dos notáveis teóricos John McCarthy, Marvin Minsky, Nathaniel Rochester e Claude Shannon. A intenção era reunir pesquisadores interessados em autômatos, processamento de linguagem, redes neurais, aprendizado e inteligência de máquinas. A proposta, em si, é bem curta, mas contém ideias inovadoras e audaciosas que viriam a definir o futuro do campo de estudo, a exemplo da seguinte:

> O estudo terá como base **a conjectura que todo aspecto do aprendizado ou qualquer outra característica da inteligência pode em princípio ser tão bem precisamente descrito que uma máquina pode ser capaz de simulá-lo**. Será realizada uma tentativa de descobrir como fazer com que máquinas usem a linguagem, formem abstrações e conceitos e resolvam problemas, até então, reservado a humanos, e se aperfeiçoem (MCCARTHY *et al.* 1955, p. 13, tradução e grifos nossos).

A conferência de Dartmouth foi um importantíssimo marco para o campo de estudo. Não só por reunir alguns dos pesquisadores mais relevantes da época e facilitar o seu diálogo, mas, principalmente, por estabelecer as bases e perspectivas fundamentais – e diversas – que definiram os rumos da prática e pesquisa no campo pelas próximas décadas: a conferência “apresentou uns aos outros todos os personagens importantes da história. Nos 20 anos seguintes, o campo seria dominado por essas pessoas e por seus alunos e colegas do MIT, da CMU, de Stanford e da IBM.” (RUSSELL; NORVIG, 2013, p. 17).

Nas décadas seguintes à conferência de Dartmouth, o campo de estudo em IA se ramificou ainda mais, em subcampos como processamento de linguagem natural, visão computacional, representação e racionalização de conhecimento, robótica, redes bayesianas, aprendizado de máquinas, entre outros (NILSSON, 2009; MCCORDUCK, 2004). É importante ressaltar que, até pouco tempo atrás, esses subcampos de estudo não conversavam muito entre si e os seus progressos costumavam ser locais e não aplicáveis aos outros subcampos, devido a diferentes técnicas e práticas.

2.3 Máquinas Atentas

Um ponto de inflexão para uma maior aproximação entre as subáreas foi o desenvolvimento e popularização dos Transformadores (*Transformers*), com a publicação e grande repercussão do artigo “Atenção é tudo o que você precisa” (VASWANI *et al*. 2017), já considerado revolucionário e central para a IA moderna, com mais de 110.000 citações. Esse trabalho tornou-se fundamental para a IA porque o que parecia, a princípio, uma nova arquitetura para o subcampo específico de Processamento de Linguagem Natural (Natural Language Processing ou NLP), mostrou-se generalizável e aplicável também a diversos outros subcampos – fato imaginado pelos autores e que acabou por se mostrar realidade.

Os Transformadores possibilitaram uma maior integração entre os, até então, muito dispersos subcampos da IA: se, antes, as pesquisas e progressos nos subcampos da IA eram muito atomizadas e, quase sempre, restritas ao contexto específico do próprio subcampo, agora, com os Transformadores, essas fronteiras entre os subcampos têm progressivamente se abrandado e os progressos relacionados aos Transformadores costumam ter aplicações em diversos domínios da IA. Sendo assim, muito embora o campo de pesquisa da IA continue fragmentado sob diferentes domínios, os Transformadores são uma ponte de contato entre tais.

Por que é o caso? Qual o grande diferencial dessa arquitetura? Muito simplificadamente, as suas especificidades permitem a “tradução” e processamento de diferentes modalidades de *inputs* – texto, áudio, imagem, vídeo –, transformando tudo na mesma “linguagem” (*tokens* e vetores) e possibilitando também *outputs* em modalidades diferentes, de um modo relativamente rápido, paralelizado, escalável e eficiente. Desse modo, os Transformadores mostraram-se úteis e promissores em diferentes ramos da IA (para muito além do Processamento de Linguagem) e, por isso, têm relação com boa parte das pesquisas e práticas do campo de estudo como um todo nos últimos anos (LIN *et al*. 2022). Apesar de tudo, os Transformadores

possuem diversas limitações e, devido a estas, há críticos que questionam se são realmente um modelo viável para atingir os fins determinados de certos subcampos ou, mesmo, para chegar a uma Inteligência Artificial Geral.

2.4 Máquinas Diversas

Apesar da convergência de interesses de boa parte dos pesquisadores em IA acerca dos Transformadores – e do aprendizado de máquinas em geral -, o campo continua fragmentado em subcampos distintos: a alcunha “Inteligência Artificial” serve mais como uma denominação genérica (ou um hiperônimo) do que como a designação de uma ciência ou de um campo de estudo bem delimitado, estruturado e uniforme. Nesse sentido, “não há nenhuma definição amplamente aceita de Inteligência Artificial. Consequentemente, o termo ‘IA’ tem sido usado em diversos sentidos diferentes, tanto dentro do Campo de Estudo quanto fora dele” (WANG, 2019, p. 1, tradução nossa). Este era o caso quando o Campo foi fundado (na Conferência de Dartmouth) e continua a sê-lo (pós-Transformadores) mais de meio século depois.

A história da Inteligência Artificial é a história da ascensão e queda de paradigmas distintos, sempre muito promissores, mas nunca bons o suficiente. Nesse processo, o público em geral ganha aplicações novas, interessantes e úteis, mas nunca aquela Inteligência Artificial Geral tão prometida.

Sendo assim, em todos esses anos de progressos e conquistas no campo, várias questões e hipóteses sobre uma inteligência artificial geral ou próxima à humana foram levantadas, mas poucas foram realmente respondidas ou corroboradas, o foco do campo se concentrou em aplicações específicas ou *expert systems*. Há uma grande preocupação em publicar sucessos e inovações, mas pouquíssima em publicar insucessos, testes falseadores, reprodução de experimentos, ou mesmo críticas de publicações anteriores. Há um grande esforço para pesquisar, produzir e publicar a “melhor, maior e mais impressionante nova IA”, e muito pouco em analisar, questionar e criticar

as atividades do Campo, o que seria essencial para a consolidação de uma ciência robusta. Sobre isto Marvin Minsky adverte: “a IA nunca poderá ser uma ciência enquanto não publicar tanto seus fracassos quanto seus sucessos” (MOOR, 2006, p. 87, tradução nossa). De modo geral, a prática da IA nas últimas décadas segue muito mais as demandas do mercado e das Big Techs do que as da ciência e da academia (e os pesquisadores de destaque da academia tendem a ser cooptados pelo mercado), isto é um grande problema caso se queira desenvolver uma IA robusta, segura e benéfica.

3. MÁQUINAS QUE PENSAM: CIÊNCIA OU METAFÍSICA?

3.1 Sobre o *status* científico da Inteligência Artificial

Afinal, seria a Inteligência Artificial uma ciência? A resposta dependerá do que considerarmos como “ciência” e o quão bem a IA se adeque nesses critérios. Avaliemos, então, a perspectivas do célebre filósofo da ciência Karl Popper (1902-1994) sobre as características fundamentais da ciência e apliquemos ao campo de estudo da IA.

Evidentemente, não é nosso intento afirmar que a filosofias de Karl Popper é a verdade absoluta sobre os critérios de cientificidade e que, caso o campo da Inteligência Artificial não se encaixe em seus parâmetros, não teria qualquer valor científico e todo o esforço de seus pesquisadores não passou de um colossal desperdício. Elegemos a filosofia da ciência de Popper para analisar o status científico do campo da Inteligência Artificial porque julgamos que seja - além de muito útil para descrevê-lo - pertinente e especialmente pontual quando aplicada às práticas da IA, uma vez que seus conceitos se adequam muito bem à história e às atividades da área.

O que define a prática científica na visão do filósofo Karl Popper? A característica fundamental da ciência é a atitude crítica (POPPER, 2006a, p. 80), seu método é o de “conjecturas e refutações” (POPPER, 1975, p. 84), e seu “critério de demarcação” que a distingue da não ciência é a falseabilidade (POPPER, 2013, p. 38). Isto é, para que uma teoria seja considerada científica

deve ser possível, mediante um teste falseador, refutá-la. Sendo assim, a prática da ciência é a de formular teorias audaciosas e investigá-las criticamente a fim de falseá-las. Nas palavras do próprio filósofo: “um cientista [...] formula hipóteses ou sistemas de teorias, e submete-os a teste, confrontando-os com a experiência, através de recursos de observação e experimentação” (POPPER, 2013, p. 27).

O campo de estudo da IA se adequa nessa descrição? É difícil responder assertivamente esta questão dada a diversidade e falta de unidade do campo: “qual parte da IA?”. A despeito disso, é possível afirmar algumas coisas: certamente, os seus pesquisadores formulam “teorias audaciosas”[2], a própria premissa do campo – máquinas pensantes ou inteligentes – evidencia isto; todavia, se buscam “investigá-las criticamente a fim de falseá-las” é algo discutível, pois a atitude do campo parece ser muito mais a de buscar “verificações” ou “confirmações” para as conjecturas (atitude, para Popper, não científica), até que, eventualmente, chegue-se a um ponto sem saída e a realidade demonstre as suas limitações, em vez de buscar ativamente refutá-las.

Também há uma questão de escopo da testabilidade: há, ao menos nos últimos anos, testes rigorosos e minuciosos quanto às capacidades de certas aplicações de IA, em um nível mais pragmático; mas não há a mesma criticidade e severidade acerca das premissas teóricas dessas aplicações, em um nível mais abstrato. Nesse sentido, fica a questão: vale a pena ser crítico apenas em um nível mais superficial e ignorar os possíveis problemas subjacentes? Dando um exemplo prático: os Modelos de Linguagem Grandes são rigorosamente testados segundo diferentes critérios (*benchmarks*) a fim de avaliar as suas capacidades; contudo, pouco se questiona (ou se testa) se, realmente, são uma boa arquitetura para se atingir níveis mais gerais de inteligência, havendo uma fé quase cega, por parte do campo, em sua efetividade e, cada vez mais, os esforços, pesquisas e investimentos nesses modelos aumentam.

À vista de tudo isso, o campo de estudo da Inteligência Artificial não se adequa plenamente à descrição de ciência empírica proposta pelo filósofo Karl Popper. Embora possua elementos vistos pelo filósofo como científicos (a formulação de conjecturas ousadas e a realização de testes), deixa a desejar quanto aos aspectos mais críticos da ciência, isto é, não tem como enfoque a falseabilidade (pelo contrário, tem uma atitude muito mais verificacionista) e a sua "testabilidade" é muito concentrada em pontos mais superficiais e pragmáticos, que não têm o caráter de "testes falseadores" da teoria subjacente - se, por exemplo, um Modelo de Linguagem Grande falha em algum teste de *benchmark* (o que costuma acontecer frequentemente), isto não põe em cheque a hipótese de que através dessa arquitetura é possível chegar a uma Inteligência Artificial Geral.

Em nossa visão, o Campo da Inteligência Artificial se adequa muito mais ao conceito popperiano de "Programa Metafísico de Pesquisa" (*Metaphysical Research Programme*), uma vez que a sua prática parece estar na fronteira entre a Ciência e a Metafísica, possuindo elementos de ambos. Mas o que é um "Programa Metafísico de Pesquisa"[3]? Nas palavras do filósofo da ciência, "o nome 'programa metafísico de pesquisa' foi usado por mim para certos programas de pesquisa para a ciência; aqueles que ainda não são testáveis" (POPPER, 1982, p. 32, tradução nossa), tais programas, "moldam e determinam o curso e o desenvolvimento da pesquisa científica" (POPPER, 1982, p. 31, tradução nossa) e são "uma estrutura possível para teorias científicas testáveis" (POPPER, 1990, p. 168 tradução nossa). Ou seja, os Programas Metafísicos de Pesquisa são ideias norteadoras da prática científica, embora elas mesmas – ainda – não sejam científicas: são conjecturas metafísicas (irrefutáveis) audaciosas suficientemente poderosas para direcionar a curiosidade e o método científico em certos caminhos a fim de explorar novos horizontes e, eventualmente, caso haja robustez de pesquisa o suficiente, torná-la uma disciplina ou teoria científica. Os exemplos de Programas Metafísicos de Pesquisa dados por Popper são: o Darwinismo,

a Teoria da Célula e o Atomismo.

Como o conceito de “Programas Metafísicos de Pesquisa” se aplica ao campo de estudo da Inteligência Artificial? Ora, desde seus primórdios, o campo foi norteado por algumas conjecturas audaciosas que direcionaram suas linhas de pesquisa (vide nota de fim ii). O próprio Alan Turing, ao explicitar as suas crenças acerca das máquinas inteligentes, corrobora tal ponto de vista:

> A concepção popular de que os cientistas procedem inexoravelmente de fatos bem estabelecidos para fatos bem estabelecidos, nunca sendo influenciados por nenhuma conjectura não provada, está muito equivocada. Desde que se deixe claro quais são os fatos provados e quais são as conjecturas, não há prejuízo algum. As conjecturas são de grande importância, pois sugerem linhas de pesquisa úteis (TURING, 1950, p. 442, tradução nossa).

De certo modo, o esforço do Campo, nesses últimos setenta anos desde a sua fundação, foi o de transformar essas conjecturas - a princípio - metafísicas em conhecimentos e práticas de uma ciência empírica séria. Tal esforço, até então, teve um grau de sucesso razoável (tendo em vista os pontos apresentados anteriormente) e, nos últimos anos, houve uma maior integração entre os subcampos da IA e uma maior coesão metodológica, principalmente em torno do aprendizado de máquinas e dos Transformardores: “A IA avançou mais rapidamente na última década, devido ao uso mais intenso do método científico nas experiências e na comparação entre as abordagens” (RUSSELL; NORVIG, 2013, p. 27).

3.2 Máquinas de Inteligência Geral: Previsões ou Profecias?

Não é preciso de uma apurada investigação histórica para atestar que muitas das promessas da Inteligência Artificial parecem ultrapassar os limites rigorosos das previsões científicas e adentrar o território fantasioso da ficção científica - ou do otimismo cego (*wishful thinking*). Nesse sentido, os pesquisadores do campo são famosos – ou infames – por suas “previsões” (ou, talvez, profecias) grandiosas que, muitas das vezes, após o infalível teste do tempo, mostram-se malogradas (até então).

Popper (2006b, p. 452-453) ressalta a distinção entre “previsões” e “profecias”, ideias muitas das vezes confundidas na prática cotidiana e científica. Para ele, a principal característica das previsões científicas é a sua condicionalidade, isto é, a delimitação de que se determinadas condições forem alcançadas, então determinada consequência ocorrerá, “se ocorrer X, então resultará Y”, se a temperatura da água numa chaleira ao nível do mar chegar à 100 graus celsius, então a água evaporará. Já as profecias têm como característica principal a sua incondicionalidade, isto é, não são dadas condições testáveis para que um determinado evento ou fenômeno ocorra, só se enuncia que este ocorrerá (“Y ocorrerá”) sem fundamentação teórica ou empírica exatas. Para o filósofo, a ciência deve ser caracterizar por fazer previsões, não profecias, muito embora, devido a falhas críticas e metodológicas, a realização de profecias seja relativamente comum em sua prática.

Dentro desse contexto, em nossa visão, muitas das promessas grandiosas na história da Inteligência Artificial soam mais como profecias do que como previsões, uma vez que, muitas das vezes, não há grandes preocupações em definir claramente os seus condicionantes, e a suntuosidade dessas alegações tende a extrapolar consideravelmente a plausibilidade de sua fundamentação empírico-teórica. Em outras palavras: promete-se, sem definir exatamente como, muito mais do que se demonstra poder cumprir.

Para dar alguns exemplos notórios: no início da década de sessenta, Hebert Simon (1960, p. 38), um dos pioneiros da pesquisa em IA e detentor de um prêmio Nobel e um prêmio Turing, afirmou que “as máquinas serão capazes, dentro de vinte anos, de fazer qualquer trabalho que um homem possa fazer [ou seja, até 1980]”; já em 1967, Marvin Minsky, outro pioneiro e grande nome da IA, alegou que “dentro de uma geração [...] o problema de criar uma inteligência artificial estará substancialmente solucionado” (MINSKY, 1967, p. 2, tradução nossa) e, em 1970, em uma entrevista à *Life Magazine*, é ainda

mais audacioso ao afirmar que “daqui a três a oito anos, teremos uma máquina com a inteligência geral de um ser humano médio [entre 1973 e 1978]”; por sua vez, em 1977, o célebre cientista da computação e roboticista Hans Moravec afirmou que “máquinas que pensam tão bem quantos humanos começarão a aparecer em dez anos [ou seja, até 1987]” (MORAVEC, 1977, tradução nossa), contudo, onze anos depois (MORAVEC, 1988), refaz sua “previsão” dizendo que isto ocorrerá em trinta ou quarenta anos (até 2028); mais recentemente, Shane Legg, eminente pesquisador da área, cofundador e cientista chefe de uma das maiores empresas de IA do mundo (Google DeepMind), em uma entrevista concedida a Alexander Kruel (2011) sobre os riscos da IA, afirmou que há uma chance de 50% de até 2028, e 90% até 2050, de desenvolvermos uma inteligência artificial geral equiparável à humana (*human-level AI*).

De modo mais abrangente, para além de algumas alegações selecionadas, e levando em conta a diversidade de pontos de vista sobre o assunto[4] , o panorama geral é: “as previsões sobre os desenvolvimentos futuros da inteligência artificial são tão confiantes quanto são diversas” (ARMSTRONG; SOTALA, 2012, p. 1, tradução nossa), que “não há consenso aparente entre os especialistas em IA sobre o futuro do campo” (GRACE *et al.*, 2024, p. 2, tradução nossa). Isto é, “os especialistas discordam fortemente entre si no que se refere ao timing e a ordem de marcos chave” (BAUM *et al.* 2011, p. 1, tradução nossa) e, por isso, “a credibilidade geral no julgamento de especialistas sobre o futuro da IA mostrou-se pobre, um resultado que se adequa aos levantamentos anteriores” (ARMSTRONG *et al.* 2014, p. 1, tradução nossa).

À vista disso, cabe questionar: os pesquisadores do campo (e, por conseguinte, o público em geral) têm uma ideia realista das potencialidades e limites do campo? ou será que o imaginário que cerca a ideia de máquinas pensantes foi por demais contaminado (pela ficção e pela propaganda) a ponto de os próprios praticantes da área terem dificuldade de distinguir

expectativas realistas daquelas fantasiosas?

Por, desde o princípio, a IA necessitar de significantes recursos para conduzir suas pesquisas e também ter se transformado em uma indústria, hoje, bilionária, é compreensível (porém, muito questionável) que seu discurso tenha se adaptado para atrair a atenção de investidores ou clientes e para vender projetos de pesquisa promissores e “novas e extraordinárias” aplicações. Igualmente, o apelo aos clichês da ficção científica é muito eficiente do ponto de vista comunicacional, dado que o público em geral tem muito mais familiaridade com a ficção do que com as especificidades técnicas da pesquisa. O grande problema dessas práticas é que o seu objetivo não é a verdade e tampouco o seu rigor é científico. Logo, o campo fica em uma ambiguidade entre os, muitas das vezes contraditórios, incentivos e pressões do mercado, que lhe geram financiamento, e os da ciência, que o desenvolvem.

Sem dúvida, é preciso um maior rigor e severidade no que se refere às alegações, promessas e previsões do campo. É preciso que seu discurso e prática sejam menos próximos daqueles da ludibriosa publicidade (que está mais preocupada com atrair a atenção do que com a verdade) e da fabulosa ficção científica (para a qual o deslumbramento é muito mais importante do que a fatualidade), e mais alinhado com os princípios da sóbria ciência, cuja característica fundamental é a atitude crítica e para a qual a verdade é a meta, por mais insípida que possa ser.

Por que as previsões dos pesquisadores da IA não são tão confiáveis quanto as de outras ciências estabelecidas? Na perspectiva de Karl Popper (2006b, p. 449-462) sobre as previsões científicas, é porque as previsões devem ser uma derivação das leis científicas, isto é, se temos uma lei científica fortemente corroborada, então as previsões sobre os fenômenos explicados por essa lei serão altamente confiáveis (e, caso falhem, podem ser usadas como falseadores potenciais da teoria). Sendo assim, “todas as ciências

teóricas são ciências de previsão" (POPPER, 2006b, p. 452). Agora, se não há uma teoria científica robusta, não há como fazer boas previsões. Dentro desse contexto, não há, até hoje, uma teoria robusta sobre Inteligência Artificial Geral (*Artificial General Intelligence*, ou AGI), embora haja uma diversidade de perspectivas e hipóteses. Por conseguinte, é impossível fazer previsões precisas sobre o assunto, e o esperado é que não haja mesmo consenso entre os especialistas. O que há são boas teorias sobre aprendizado de máquinas e sobre outras técnicas e especificidades de sistemas especialistas (*expert systems* ou *narrow* AI), mas a lacuna entre estes e uma Inteligência Artificial Geral tão boa ou superior à humana é gigantesca.

O que, então, as previsões dos pesquisadores da IA podem nos dizer? É importante considerar tais previsões, por mais diversas e incertas que possam ser, porque - algo sobre o que não falta consenso é que - uma Inteligência Artificial Geral ou de um nível semelhante à inteligência humana pode ter impactos completamente transformadores em nossa sociedade, sejam eles positivos (ou, mesmo, utópicos) ou negativos (quiçá, catastróficos):

> Uma IA geral superinteligente pode usar sua vasta inteligência para resolver a maioria dos problemas humanos ou para destruir os seus criadores humanos. Sendo seus resultados bons ou ruins, a façanha de uma Inteligência Artificial Geral seria um dos eventos mais importantes da história humana (BAUM *et al.* 2011, p. 3, tradução nossa).

Nesse sentido, é do interesse público saber o quão razoáveis são as estimativas para essa superinteligência artificial, quais são seus riscos e como será o seu desenvolvimento, segundo Baum *et al*. (2011, p. 3, tradução nossa), "é de considerável importância e interesse fazer as melhores estimativas sobre quando e como uma IA geral ocorrerá". Por menor consenso que possa haver entre os seus especialistas e por mais superestimadas que sejam as previsões sobre o futuro do campo, ainda assim, esses mesmos especialistas são as pessoas mais bem informadas e mais bem indicadas para tentar fazer qualquer prognóstico acerca dessa tecnologia de gigantesco potencial, pois "sua familiaridade com a tecnologia e com a dinâmica de seus

progressos passados os colocam em uma boa posição para realizar suposições instruídas sobre o futuro da IA" (GRACE *et al*. 2024, p. 2, tradução nossa). O conjunto de perspectivas sobre o futuro do campo, embora não possa nos oferecer uma precisão científica sobre o seu desenvolvimento, pode nos oferecer uma ideia do sentimento geral de seus especialistas sobre o assunto e, a partir disso, uma ideia do quão urgente deve ser a nossa preocupação com os seus riscos.

4. CONSIDERAÇÕES FINAIS

Afinal, máquinas podem pensar? Embora seja uma questão discutida desde a origem do campo de estudo da Inteligência Artificial, ela continua em aberto e alvo de interessantes discussões até hoje. A despeito da sua falta de resolução, tal indagação provocou importantes debates, ideias e projetos na história da Inteligência Artificial. Turing, um dos primeiros a formalmente discuti-la, propos substituí-la por um teste objetivo (o "jogo da imitação") e, assim, estipulou um critério de inteligência que, para o bem ou para o mal, serviu de modelo para os futuros progressos do campo. Pouco tempo depois das especulações iniciais de Turing, os participantes da Conferência de Dartmouth propunham-se a simular "todo aspecto do aprendizado ou qualquer outra característica da inteligência" (MCCARTHY *et al*. 1955, p. 13, tradução nossa), dividindo o campo em formação em projetos de pesquisa distintos. Mais recentemente, surgiu uma ponte entre os diferentes subcampos da IA: os Transformadores, que têm se mostrado muito úteis e promissores.

A partir desse contexto histórico de formação do campo de estudo da Inteligência Artificial, levantamos a questão: afinal, formou-se uma Ciência? O que podemos esperar de suas supostas previsões? Recorremos à filosofia da ciência do eminente filósofo Karl Popper a fim esclarecer tais pontos e concluímos que a história da Inteligência Artificial é a história da tentativa de transformar uma conjectura metafísica audaciosa – de que máquinas podem pensar – em uma ciência empírica. Nesse sentido, embora o campo de estudo da Inteligência Artificial possua diversos elementos de uma ciência, também

continua a carregar elementos não científicos (ou metafísicos), sendo a denominação popperiana de “Programa Metafísico de Pesquisa” mais adequada para descrevê-lo. Por ainda não haver uma teoria robusta acerca de uma Inteligência Artificial Geral ou de uma inteligência próxima à humana, também não é possível fazer boas previsões sobre se, como e quando as atingiremos, sendo assim, tentar fazê-lo equivale a fazer profecias. Finalmente, é importante que busquemos maior cientificidade nas práticas da Inteligência Artificial em razão das suas possíveis e grandiosas implicações e riscos: o nosso futuro depende disso.

REFERÊNCIAS

ARMSTRONG, Stuart; SOTALA, Kaj. “How We’re Predicting AI—or Failing To.” In: ROMPORTL *et al.* **Beyond AI: Artificial Dreams**. 52–75. Pilsen: University of West Bohemia. 2012.

ARMSTRONG, Stuart; SOTALA, Kaj; ÓHÉIGEARTAIGH, Séan. “The Errors, insights and lessons of famous AI predictions – and what they mean for the future”. In: **Journal of Experimental & Theoretical Artificial Intelligence**. v. 26, n. 3, p. 317-342, 2014.

BAUM, Seth; GOERTZEL, Ben; GOERTZEL, Ted. How Long Until Human-Level AI? Results from an Expert Assessment. **Technological Forecasting & Social Change**. v.78, n. 1, p. 185-195, 2011.

COPELAND, Jack. **The Essential Turing**. Oxford: Oxford University Press, 2004.

LIN, Tianyang; WANG, Yuxin; LIU, Xiangyang; QIU, Xipeng. A survey of transformers. **AI Open**, v. 3, p. 111-132, 2022. Disponível em: https://doi.org/10.1016/j.aiopen.2022.10.001. 2022. Acesso em: 10 jan. 2024.

DIAS, Elizabeth de Assis. O papel dos programas de investigação metafísica no avanço da ciência no pensamento de Popper. **Disputatio:** Philosophical Research Bulletin. v.12, n. 24. p. 205-2225, 2023.

DREYFUS, Hubert L. **What computers can’t do**. New York: Harper & Row, 1972.

GRACE, Katja *et al*. **Thousands of AI Authors on the Future of AI**. 2024. Disponível em: https://arxiv.org/abs/2401.02843. Acesso em: 15 abr. 2024.

KRUEL, Alexandrer. **Interview Series on Risks from AI**. 2011. Disponível em: https://www.lesswrong.com/tag/interview-series-on-risks-from-ai. Acesso em: 15 Abr. 2024.

MCCARTHY, J. *et al.* A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. 1955. In: **AI Magazine**. v. 27, n. 4. 2006.

MCCORDUCK, Pamela. **Machines who think**. Massachusetts: A K Peters Ltd., 2004.

MINSKY, Marvin. **Computation:** Finite and Infinite Machines. New Jersey: Englewood, 1967.

MOOR, J. The Dartmouth College Artificial Intelligence Conference*:* the next Fifty Years. **AI Magazine**, v. 27, n. 4, p. 87-91, 2006.

MORAVEC, Hans. **Intelligente Machines:** How to Get there From Here and What to Do Afterwards. Artigo não publicado. 1977.

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

NEWELL, A.; SIMON, H. A. “Computer science as empirical inquiry: Symbols and search”. In: **Communications of the ACM**, v. 19, n. 3, p. 113-126, 1976.

NILSSON, Nils. **The Quest for Artificial Intelligence:** a history of ideas and achievements. Cambridge: Cambridge University Press, 2009.

POPPER, Karl. **A lógica da pesquisa científica**. Tradução Leonidas Hegenberg e Octanny Silveira da Mota. 2 ed. São Paulo: Cultrix, 2013.

POPPER, Karl. **Conhecimento objetivo***.* Tradução de Milton Amado. São Paulo: Ed. Universidade de São Paulo, 1975.

POPPER, Karl. **Em busca de um mundo melhor**. Tradução de Milton Camargo Mota. São Paulo: Martins, 2006a.

POPPER, Karl. **Conjecturas e Refutações**. Tradução de Benedita Bittencourt. Coimbra: Almedina, 2006b.

POPPER, Karl. **Quantum theory and the schism in physics**. The Postscript to The logic of scientific discovery as edited by W. W. Bartley III. New Jersey: Rowman and Littlefield, 1982.

POPPER, Karl. **Unended Quest:** an intellectual autobiography. Illinois: Open Court, 1990.

RUSSELL, Stuart; NORVIG, Peter. **Inteligência Artificial**. Tradução da 3ª edição, feita por Regina Célia Simille. Rio de Janeiro: Elsevier, 2013.

SIMON, Herbert A. **The New Science of Management Decision**. New York: Harper & Row, 1960.

TURING, Alan. Computing Machinery and Intelligence. **Mind**, v. LIX, n. 236, p. 433-460, 1950.

VASWANI, Ashish et al. Attention is all you need. **Proceedings of the 31st Conference on Neural Information Processing Systems,** *(NIPS' 17)*. p. 1–11. 2017.

WANG, Pei. On Defining Artificial Intelligence. **Journal of Artificial General Intelligence**. v. 10, n. 2, p. 2019.

[1]**Notas:**

[1] O “Jogo da Imitação” ou, como estudiosos posteriores costumam denominar, “Teste de Turing”, é um jogo proposto por Turing (1950, p. 433): há três jogadores, um interrogador humano (C) e outros dois participantes (A e B) desconhecidos para C, um deles é humano (A) e outro é uma máquina (B). O interrogador pode fazer perguntas para A e B, os quais podem respondê-las através de mensagens escritas. O objetivo do interrogador é descobrir quem entre A e B é humano, já o objetivo de A e B é tentar convencer C de que são humanos. Assim, o aspecto interessante do jogo é se a máquina (B) consegue convencer o interrogador (C) de que é humana, ou posto de outro modo, se as respostas e o comportamento da máquina (B) são tão semelhantes aos de um humano que o interrogador (C) não consegue notar a diferença. Com o passar do tempo foram propostas diferentes formas do Teste de Turing, a forma original pode ser encontrada em (TURING, 1950). Há uma competição anual (the Loebner Prize) baseada no Teste de Turing que premia aqueles que conseguem desenvolver programas que consigam passar no Teste de Turing. Todavia, a legitimidade de tal competição é controversa.

[2] Para dar apenas alguns exemplos: (a) a conjectura da proposta da Conferência de Dartmouth, segundo a qual “todo aspecto do aprendizado ou qualquer outra característica da inteligência pode em princípio ser tão bem precisamente descrito que uma máquina pode ser capaz de simulá-lo” (MCCARTHY *et al*. 1955, p. 13); (b) A “hipótese do sistema de símbolos físicos” de Newell e Simon, segundo a qual, “Um sistema de símbolos físicos tem os meios necessários e suficientes para a ação geral inteligente” (NEWELL; SIMON, 1976, p. 116, tradução nossa); (c) ou a “Teoria Computacional da Mente”, segundo a qual, a mente é um sistema de processamento de informações tal como um computador: “O computador serve como um modelo da mente [...]. Tanto empiristas quanto idealistas prepararam o terreno para esse modelo de pensamento como processamento de dados” (DREYFUS, 1972, p. 68).

[3] Evidentemente, a explicação dada aqui é muito simplória e não dá conta da complexidade do conceito, para uma melhor compreensão da temática, recomendamos o excelente artigo de Dias (2023, p. 205-225).

[4] Há diferentes estudos que fazem um levantamento da opinião dos especialistas da área sobre os possíveis prognósticos do campo e, de certo modo, esse tipo de pesquisa virou algo habitual na área nas últimas décadas (obviamente, não é nosso intento abordar as especificidades dessa temática aqui, apenas fazer alguns apontamentos). A depender do estudo e, principalmente, de sua data de realização, a média da estimativa de quando atingiríamos um tipo de Inteligência Artificial equiparável à humana (human-level AI) ou uma Inteligência Artificial Geral (Artifciial General Intelligence, AGI) varia, como é de se esperar, contudo, um fato curioso é que, a despeito dos avanços significativos no campo em todas essas décadas desde sua gênese, o tipo de previsão mais comuns é de que a IA (geral) está entre 15 e 25 anos em nossa frente, nesse sentido, alguns críticos argumentam que “a IA está perpetuamente entre 15 a 25 anos à frente. Dessa forma, o previsor pode ganhar crédito por estar trabalhando em algo que será relevante, mas sem a possibilidade de que a sua previsão possa ser falseada durante a sua carreira” (ARMSTRONG; SOTALA, 2012, p. 15).

SOBRE OS AUTORES:

Paulo Uiris da Silva Gomes
Doutorando no Programa de Pós-Graduação em Comunicação, Linguagem e Cultura (PPGCLC) da Universidade da Amazônia (UNAMA). Mestre em Filosofia pela Universidade Federal do Pará (PPGFIL/UFPa). Bacharel e Licenciado em Filosofia pela Universidade Federal do Pará (UFPa). É Bolsista da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).
Orcid: https://orcid.org/0000-0002-0653-7816
E-mail: paulouiris@gmail.com

Rosângela Araújo Darwich
Professora do curso de Psicologia e do Programa de Pós-Graduação em Comunicação, Linguagens e Cultura (PPGCLC) da Universidade da Amazônia (Unama). Doutora em Psicologia: Teoria e Pesquisa do Comportamento pela Universidade Federal do Pará (UFPA). Coordenadora do Grupo de Pesquisa “Poesia no Dia a Dia: Grupos Vivenciais e Resiliência (CNPq).
Orcid: https://orcid.org/0000-0001-7325-9097
E-mail: rosangeladarwich@yahoo.com.br

Artigo recebido em: 12 maio 2024. | Artigo aprovado em: 13 jun. 2024.