Introdução

"A maior parte do que fazemos com o machine learning acontece abaixo da superfície. Embora possa não ser visível, grande parte do impacto do machine learning será assim: uma melhoria silenciosa, mas significativa, das operações essenciais."

Jeff Bezos[1]

"A inteligência artificial (AI) é o campo da ciência e da engenharia voltado para a criação de máquinas inteligentes e, principalmente, de software inteligente. Ela está relacionada à tarefa semelhante de usar computadores para entender a inteligência humana, mas a AI não precisa se limitar a métodos biologicamente observáveis"[2].

Essa foi a definição de AI oferecida por John McCarthy, professor da Universidade de Stanford, um dos fundadores da disciplina e coautor do termo "inteligência artificial".

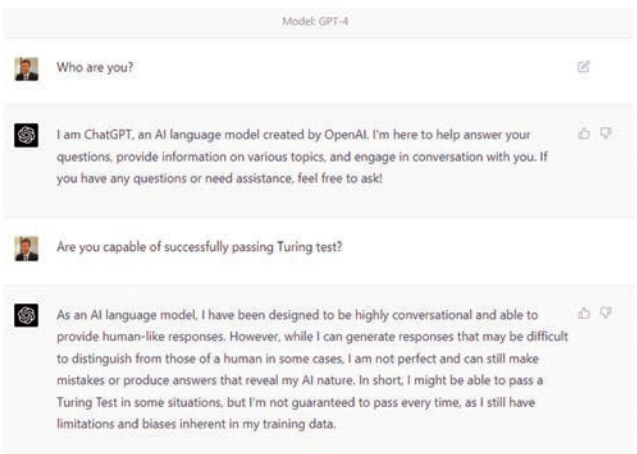
No entanto, já em 1950, Alan Turing se perguntava[3]: "as máquinas podem pensar?" e formulou o que mais tarde se tornaria conhecido como o "teste de Turing": um teste da capacidade de uma máquina de demonstrar inteligência indistinguível da de um ser humano. Turing propôs que um avaliador humano julgasse as conversas em linguagem natural entre uma pessoa e uma máquina projetada para gerar respostas semelhantes às humanas. Se o avaliador não conseguisse distinguir a máquina do ser humano, a máquina teria passado no teste.

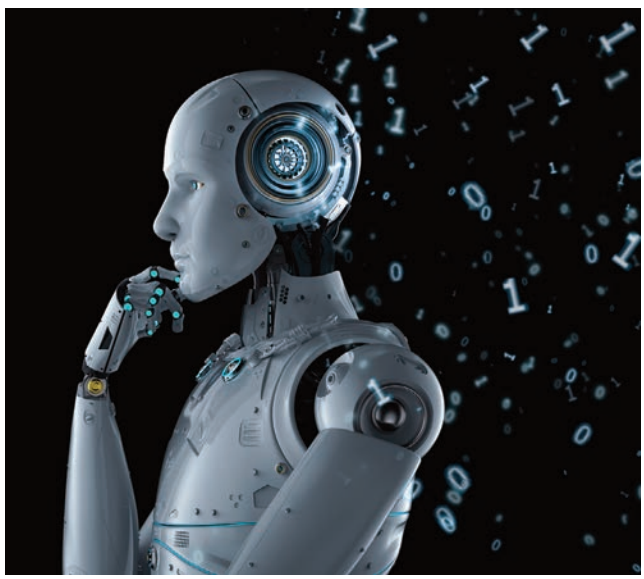
Embora haja controvérsia sobre isso[4], muitos autores consideram que já existem inteligências artificiais que poderiam passar no teste de Turing, como o GPT-4, da Open AI Foundation, embora o próprio GPT-4 não seja totalmente claro (Fig. 1). Há também testes mais sofisticados, como o desafio de esquemas de Winograd, que consiste em resolver anáforas complexas que exigem conhecimento e bom senso[5], algo que a AI atual ainda não parece ser capaz de fazer.

[1]Bezos (nascido em 1964), J., fundador, presidente executivo e ex-CEO da Amazon.
[2]McCarthy (2004). Professor de Ciência da Computação na Universidade de Stanford.
[3]Turing (1950). Matemático, lógico, cientista teórico da computação, criptógrafo, filósofo e biólogo teórico britânico.
[4]Harnad (2003). Professor de Psicologia da Universidade de Quebec em Montreal (UQAM) e da Universidade McGill, e Professor Emérito de Ciências Cognitivas da Universidade de Southampton.
[5]Um esquema Winograd é uma pergunta de escolha binária em que (i) há duas partes mencionadas na pergunta; (ii) pronomes são usados para se referir a elas; (iii) há ambiguidade sobre a quem o pronome se refere; e (iv) há palavras específicas que podem alterar a resposta correta. Em um exemplo do mesmo Terry Winograd (Professor de Ciência da Computação da Universidade de Stanford):
- Pergunta: Os conselheiros municipais negaram permissão aos manifestantes porque (temiam/defendiam) a violência. Quem (teme/defende) a violência?
- Resposta: (os conselheiros / os manifestantes).

Con ello, se puede generar un test alternativo al Test de Turing, utilizando dichas preguntas y penalizando fuertemente las respuestas erróneas (véase Levesque (2014)).

Figura 1. Conversa com o GPT-4 sobre sua capacidade de passar no teste de Turing.





Ainda assim, embora o campo da AI não seja novo, nos últimos anos houve avanços vertiginosos, com aplicações que vão de carros autônomos a diagnósticos médicos, trading automático, reconhecimento facial, gerenciamento de energia, segurança cibernética, robótica e tradução automática, para citar apenas alguns.

Uma característica distintiva da AI atual está precisamente ligada à definição de McCarthy acima: ela não se limita a métodos observáveis e, quando atinge um certo nível de sofisticação, apresenta problemas de interpretabilidade. Em outras palavras: os modelos de AI tendem a ter uma alta taxa de acerto, muito maior do que os algoritmos tradicionais; mas, em cada caso individual, pode ser extremamente complexo explicar por que o modelo produziu um determinado resultado.

Embora existam aplicações de AI em que seja menos relevante poder entender ou explicar por que o algoritmo retornou um determinado valor, em muitos casos isso é essencial e é um requisito regulatório. Por exemplo, na União Europeia, de acordo com o Regulamento Geral de Proteção de Dados (GDPR), os consumidores têm o que é conhecido como "direito a uma explicação"[6]:

> [...] não estar sujeito a uma decisão [...] baseada exclusivamente em processamento automatizado [...], como a rejeição automática de uma solicitação de crédito on-line, [...] [na qual] não há intervenção humana envolvida", e tem o direito de "receber uma explicação sobre a decisão tomada [...] e contestar a decisão".

Isso levou ao desenvolvimento da disciplina de Inteligência Artificial Explicável (XAI), que é o campo de estudo que visa tornar os sistemas de AI compreensíveis para os seres humanos[7], em oposição à noção de "black box", que se refere a algoritmos nos quais somente os resultados são observáveis e o funcionamento do modelo é desconhecido, ou a lógica dos resultados não é explicada.

Um algoritmo pode ser considerado[8] como pertencente à disciplina XAI se seguir três princípios: transparência, interpretabilidade e explicabilidade. A transparência é garantida se os processos que calculam os parâmetros do modelo e produzem os resultados puderem ser descritos e justificados. A interpretabilidade descreve a possibilidade de entender o modelo e apresentar como ele toma decisões de uma forma compreensível para o ser humano. A explicabilidade refere-se à capacidade de decifrar por que uma determinada observação recebeu um valor específico. Na prática, esses três termos estão intimamente ligados e são frequentemente usados de forma intercambiável, na ausência de consenso sobre suas definições precisas[9].

Para alcançar esses princípios, há basicamente duas abordagens possíveis: desenvolver algoritmos que sejam interpretáveis e explicáveis por sua natureza (como regressões lineares, modelos logísticos ou multinomiais ou certos tipos de redes neurais

[6]GDPR (2018), Considerando 71.
[7]Vilone et al. (2021). Doutorado em Inteligência Artificial, School of Computer Science, Technological University of Dublin.
[8]Roscher et al. (2020). Cientista de dados da Universidade Técnica de Munique.
[9]Marcinkevics et al. (2020). Pesquisador do Departamento de Ciência da Computação, ETH Zurich.

profundas, entre outros) ou usar técnicas de interpretabilidade como ferramentas para alcançar esses princípios[10].

Portanto, a XAI se preocupa tanto com técnicas para tentar explicar o comportamento de determinados modelos opacos ("black box") quanto com o design de algoritmos inerentemente interpretáveis ("white box")[11].

A XAI é fundamental para o desenvolvimento da AI e, portanto, para os profissionais que trabalham em contato com ela, devido a pelo menos três fatores:

- Contribui para gerar confiança na tomada de decisões com base em modelos de AI; sem essa confiança, os usuários desses modelos podem demonstrar resistência à sua adoção.
- É um requisito regulatório em determinadas áreas (por exemplo, proteção de dados, proteção ao consumidor, igualdade de oportunidades na contratação de funcionários, regulação de modelos em bancos).
- Favorece o aprimoramento e a robustez dos modelos de AI (por exemplo, identificando e eliminando vieses, compreendendo as informações relevantes para produzir um determinado resultado ou antecipando possíveis erros em observações não contempladas na amostra de treinamento do modelo). Isso leva ao desenvolvimento de algoritmos éticos e permite que as organizações concentrem seus esforços na identificação e na garantia da qualidade dos dados que são relevantes para seus processos de tomada de decisão.

Embora o desenvolvimento de sistemas XAI esteja recebendo muita atenção do meio acadêmico, do setor e dos órgãos reguladores, ele ainda apresenta vários desafios.

Este documento analisará o contexto e a justificativa da XAI, incluindo a regulação sobre o assunto e suas implicações na organização; o estado da arte e as principais técnicas da XAI; e o progresso e os desafios não resolvidos da XAI. Por fim, será apresentado um estudo de caso da XAI para ajudar a ilustrar sua aplicação prática.

[10] iDanae (2022). Cátedra (inteligência, dados, análise e estratégia) em Big Data e Analytics, que surgiu graças à colaboração entre a Management Solutions e a Universidade Politécnica de Madri (UPM) nos âmbitos educativo, científico e técnico, e tem como objetivo promover a geração de conhecimento, a difusão e a transferência de tecnologia, e fomentar a P&D&I na área de Data Analytics.
[11] Sudjianto et al. (2011). Diretor de Risco de Modelo da Wells Fargo.