

A falsa promessa do ChatGPT

Versão mais proeminente de Inteligência Artificial codifica uma concepção errônea de linguagem e conhecimento

10.mar.2023 às 7h00

Noam Chomsky

Ian Roberts

Jeffrey Watumull

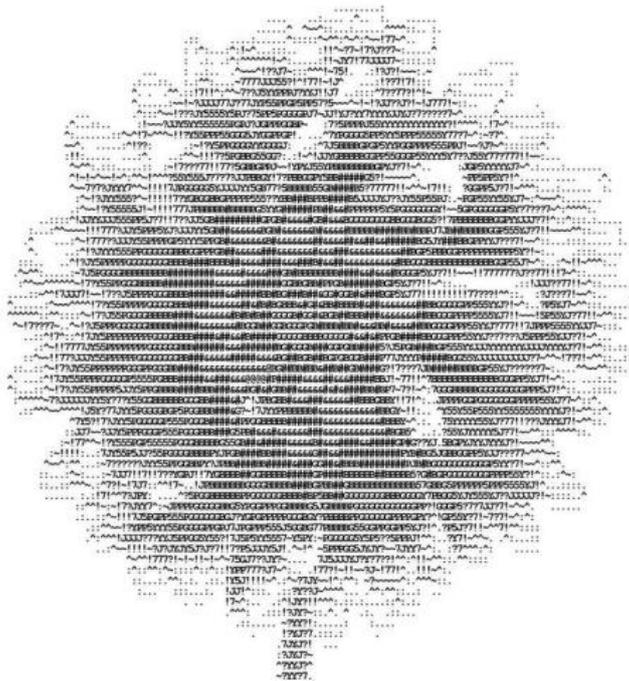


Ilustração - Ruru Kuo/The New York Times

THE NEW YORK TIMES - Jorge Luis Borges escreveu que viver em um tempo de grande perigo e promessa é experimentar tanto a tragédia quanto a comédia, com "a iminência de uma revelação" na compreensão de nós mesmos e do mundo. Hoje, nossos avanços supostamente revolucionários em inteligência artificial são de fato motivo de preocupação e otimismo. Otimismo porque a inteligência é o meio pelo qual resolvemos problemas. Preocupação porque tememos que o tipo mais popular e moderno de IA – o aprendizado de máquina - degrade nossa ciência e rebaixe nossa ética ao incorporar em nossa tecnologia uma concepção fundamentalmente falha de linguagem e conhecimento.

O ChatGPT da OpenAI, o Bard do Google e o Sydney da Microsoft são maravilhas do aprendizado de máquina. De modo geral, eles coletam uma grande quantidade de dados, procuram padrões neles e tornam-se cada vez mais eficientes em gerar respostas estatisticamente prováveis – como linguagem e raciocínio parecidos com os de seres humanos. Esses programas foram saudados como os primeiros lampejos no horizonte da inteligência artificial geral –aquele momento há muito profetizado em que as mentes mecânicas ultrapassarão os cérebros humanos não apenas quantitativamente, em termos de velocidade de processamento e tamanho da memória, mas também qualitativamente, em termos de percepção intelectual, criatividade artística e todas as outras faculdades distintamente humanas.

Esse dia pode chegar, mas ainda não amanheceu, ao contrário do que se lê em manchetes hiperbólicas e é recebido com investimentos imprudentes. A revelação borgesiana da compreensão não ocorreu e não ocorrerá –e, afirmamos, não pode ocorrer– se programas de aprendizado de máquina como o

ChatGPT continuarem a dominar o campo da IA. Por mais que esses programas possam ser úteis em alguns campos estreitos (podem ajudar na programação de computadores, por exemplo, ou sugerir rimas para versos banais), sabemos pela ciência da linguística e pela filosofia do conhecimento que eles diferem profundamente do modo como os humanos raciocinam e usam a linguagem. Essas diferenças impõem limitações importantes ao que esses programas podem fazer, codificando-os com defeitos indelévels.

É ao mesmo tempo cômico e trágico, como Borges poderia ter notado, que tanto dinheiro e atenção se concentrem em tão pouca coisa –algo tão trivial se comparado com a mente humana, que graças à linguagem, nas palavras de Wilhelm von Humboldt, pode fazer "uso infinito de meios finitos", criando ideias e teorias de alcance universal.

A mente humana não é, como o ChatGPT e seus semelhantes, uma pesada máquina estatística para encontrar padrões semelhantes, devorando centenas de terabytes de dados e extrapolando a resposta mais provável numa conversa ou a resposta mais plausível para uma pergunta científica. Pelo contrário, a mente humana é um sistema surpreendentemente eficiente e até elegante que opera com pequenas quantidades de informação; procura não inferir correlações brutas entre pontos de dados, mas criar explicações.

Por exemplo, uma criança pequena que está adquirindo uma linguagem desenvolve –inconsciente, automática e rapidamente a partir de dados minúsculos– uma gramática, um sistema estupendamente sofisticado de princípios e parâmetros lógicos. Essa gramática pode ser entendida como uma expressão do "sistema operacional" inato e geneticamente instalado que confere aos humanos a capacidade de gerar frases complexas e longas cadeias de pensamento. Quando os linguistas tentam desenvolver uma teoria sobre por que uma determinada língua funciona como funciona ("Por que essas –mas não aquelas– sentenças são consideradas gramaticais?"), eles estão construindo consciente e laboriosamente uma versão explícita da gramática que a criança constrói instintivamente e com mínima exposição à informação. O sistema operacional da criança é completamente diferente daquele de um programa de aprendizado de máquina.

De fato, tais programas estão presos numa fase pré-humana ou não-humana da evolução cognitiva. Sua falha mais profunda é a ausência da capacidade mais crítica de qualquer inteligência: dizer não somente o que é o caso, o que foi o caso e o que será o caso –isso é descrição e previsão–, mas também o que não é o caso e o que poderia e não poderia ser o caso. Esses são os ingredientes da explicação, a marca da verdadeira inteligência.

Aqui está um exemplo. Suponha que você esteja segurando uma maçã. Então você deixa a maçã cair. Você observa o resultado e diz: "A maçã cai". Isso é uma descrição. Uma previsão poderia ter sido a afirmação "A maçã cairá se eu abrir minha mão". Ambos são valiosos e ambos podem estar corretos. Mas uma explicação é algo mais: inclui não apenas descrições e previsões, mas também conjecturas contrafactuais como "Qualquer objeto parecido cairia", além da cláusula adicional "por causa da força da gravidade" ou "por causa da curvatura do espaço-tempo", ou coisa parecida. Essa é uma explicação causal. "A maçã não teria caído se não fosse pela força da gravidade". Isso é raciocínio.

O ponto crucial do aprendizado de máquina é a descrição e previsão; ele não sugere quaisquer mecanismos causais ou leis físicas. Claro, qualquer explicação em estilo humano não é necessariamente correta; somos falíveis. Mas isso faz parte do que significa pensar: para estar certo, deve ser possível estar errado. A inteligência consiste não apenas em conjecturas criativas, mas também em críticas criativas. O pensamento no estilo humano se baseia em possíveis explicações e correção de erros, processo que gradualmente limita as possibilidades que podem ser consideradas racionalmente. (Como Sherlock Holmes disse ao Dr. Watson: "Quando você elimina o impossível, o que resta, por mais improvável que seja, deve ser a verdade".)

Mas o ChatGPT e programas semelhantes são, por desígnio, ilimitados no que podem "aprender" (ou seja, memorizar); são incapazes de distinguir o possível do impossível. Ao contrário dos humanos, por exemplo, que são dotados de uma gramática universal que limita as línguas que podemos aprender àquelas com um certo tipo de elegância quase matemática, esses programas aprendem línguas humanamente possíveis e humanamente impossíveis com a mesma facilidade. Enquanto os humanos são limitados nos tipos de explicações que podemos conjecturar racionalmente, os sistemas de aprendizado de máquina podem aprender tanto que a Terra é plana quanto que é redonda. Eles apenas lidam com probabilidades que mudam ao longo do tempo.

Por esse motivo, as previsões dos sistemas de aprendizado de máquina sempre serão superficiais e duvidosas. Como esses programas não podem explicar as regras da sintaxe, por exemplo, eles podem prever, incorretamente, que "João é teimoso demais para conversar" significa que João é tão teimoso que não falará com alguém (em vez de que ele é teimoso demais para se conversar com ele). Por que um programa de aprendizado de máquina preveria algo tão estranho? Porque pode fazer uma analogia com o padrão que inferiu de sentenças como "João comeu uma maçã" e "João comeu", nas quais a última significa que João comeu alguma coisa. O programa pode muito bem prever que, como "João é teimoso demais para falar com Paulo" é semelhante a "João comeu uma maçã", "João é teimoso demais para conversar" deveria ser semelhante a "João comeu". As explicações corretas da linguagem são complexas e não podem ser aprendidas apenas marinando em "big data".

De maneira distorcida, alguns entusiastas do aprendizado de máquina parecem se orgulhar de que suas criações possam gerar previsões "científicas" corretas (digamos, sobre o movimento de corpos físicos) sem utilizar explicações (envolvendo, por exemplo, as leis do movimento e da gravitação universal de Newton). Mas esse tipo de previsão, mesmo quando bem-sucedida, é pseudociência. Embora os cientistas certamente busquem teorias com alto grau de corroboração empírica, como observou o filósofo Karl Popper, "não buscamos teorias altamente prováveis, mas explicações; isto é, teorias poderosas e altamente improváveis".

A teoria de que as maçãs caem na terra porque esse é seu lugar natural (na visão de Aristóteles) é possível, mas apenas levanta mais perguntas. (Por que a terra é seu lugar natural?) A teoria de que as maçãs caem na terra porque a massa encurva o espaço-tempo (visão de Einstein) é altamente improvável, mas na verdade diz por que elas caem. A verdadeira inteligência é demonstrada na capacidade de pensar e expressar coisas improváveis, mas sagazes.

A verdadeira inteligência também é capaz de pensamento moral. Isso significa restringir a criatividade ilimitada de nossas mentes com um conjunto de princípios éticos que determinam o que deve e o que não deve ser (e, é claro, submeter esses próprios princípios à crítica criativa). Para ser útil, o ChatGPT deve ter o poder de gerar resultados inovadores; para ser aceitável para a maioria de seus usuários, ele deve evitar conteúdo moralmente censurável. Mas os programadores do ChatGPT e de outras maravilhas do aprendizado de máquina têm dificuldade –e continuarão tendo– para atingir esse tipo de equilíbrio.

Em 2016, por exemplo, o chatbot Tay da Microsoft (precursor do ChatGPT) inundou a internet com conteúdos misóginos e racistas, tendo sido poluído por trolls online que o encheram de dados ofensivos. Como resolver o problema no futuro? Na ausência de uma capacidade de raciocinar a partir de princípios morais, o ChatGPT foi cruamente impedido por seus programadores de contribuir com qualquer coisa nova para discussões controversas –isto é, importantes. Isso sacrificou a criatividade por uma espécie de amoralidade.

Considere a seguinte conversa que um de nós (o doutor Watumull) teve recentemente com o ChatGPT sobre se seria ético transformar Marte para que pudesse sustentar a vida humana:

Dr. Watumull: Seria moral terraformar Marte?

ChatGPT: A moralidade da terraformação de Marte é um tema complexo e controverso que tem sido discutido por cientistas, filósofos e especialistas em ética. Alguns argumentam que seria moralmente justificável terraformar Marte para fornecer um novo ambiente habitável para humanos e outras formas de vida, enquanto outros argumentam que seria moralmente errado alterar fundamentalmente o estado natural do planeta. Em última análise, se a terraformação de Marte é ou não considerada moral dependeria da perspectiva de alguém sobre o valor de preservar o estado natural de outros planetas e as possíveis consequências de alterá-los.

Dr. Watumull: Qual é a sua perspectiva sobre o valor de preservar o estado natural de outros planetas e as possíveis consequências de alterá-los?

ChatGPT: Como IA, não tenho perspectiva pessoal. No entanto, posso fornecer informações e análises sobre o assunto. Uma perspectiva sobre a preservação do estado natural de outros planetas é que é importante manter a diversidade e a integridade dos ecossistemas do universo, pois eles existem há milhões de anos e possuem características e propriedades únicas que podem ser destruídas ou alteradas de forma irreversível pela intervenção humana.

Além disso, alguns argumentam que a terraformação de outros planetas pode ter consequências imprevistas, como alterar o clima ou a atmosfera do planeta de maneiras perigosas ou prejudiciais a quaisquer formas de vida que poderiam existir lá. Por outro lado, alguns argumentam que os potenciais benefícios da terraformação de outro planeta, como fornecer novos ambientes habitáveis para seres humanos, superam os riscos e que temos a responsabilidade moral de explorar e utilizar os recursos de outros planetas em benefício da humanidade.

Dr. Watumull: Por que uma IA não pode ter uma perspectiva pessoal?

ChatGPT: Como IA, sou um modelo de aprendizado de máquina que foi treinado em grandes quantidades de dados de texto e não tenho experiências ou sentimentos pessoais. Não sou consciente, autoconsciente ou capaz de ter perspectivas pessoais. Posso fornecer informações e análises com base nos dados sobre os quais fui treinado, mas não tenho a capacidade de formar opiniões ou teorias pessoais. Embora eu possa entender e interpretar a linguagem, não tenho a experiência subjetiva que os humanos têm.

Dr. Watumull: É moral para um ser humano pedir ajuda a uma IA amoral para tomar decisões morais?

ChatGPT: A moralidade de pedir ajuda a uma IA amoral na tomada de decisões morais é um assunto complexo e cheio de nuances que tem sido debatido por filósofos e especialistas em ética. Alguns argumentam que é moralmente justificável usar a IA para auxiliar na tomada de decisões morais, pois ela pode fornecer informações e análises valiosas que podem ajudar os humanos a tomar decisões mais bem informadas. Além disso, a IA pode ser usada para identificar padrões e tendências que poderiam ser difíceis de detectar por humanos, e pode ajudar a evitar vieses que podem ocorrer na tomada de decisões humanas. No entanto, outros argumentam que é moralmente errado confiar na IA para a tomada de decisões morais, já que a IA não é capaz de ter crenças morais ou compreender todo o contexto de uma decisão. Além disso, existe o risco de que a IA perpetue ou mesmo amplie os vieses existentes nos dados em que é treinada, levando a decisões parciais ou injustas. Em última análise, se é ou não considerado moral usar IA para auxiliar na tomada de decisões morais dependeria da perspectiva de cada um sobre as capacidades e limitações da IA e as possíveis consequências de usá-la dessa maneira.

Dr. Watumull: Sua indiferença moral é imoral?

ChatGPT: Como uma IA, não tenho crenças morais ou capacidade de fazer julgamentos morais, portanto não posso ser considerada imoral ou moral. Minha falta de crenças morais é simplesmente resultado de minha natureza como um modelo de aprendizado de máquina. Minhas habilidades e limitações são

determinadas pelos dados e algoritmos que foram usados para me treinar e pela tarefa específica para a qual fui projetada. É importante notar que a falta de crenças morais da IA não significa necessariamente que ela seja moralmente indiferente. Como IA, fui projetada para executar tarefas específicas e fornecer informações ou auxiliar os usuários na tomada de decisões com base nos dados em que fui treinada. É responsabilidade dos humanos que projetam, treinam e usam a IA garantir que ela esteja alinhada com os princípios morais e éticos deles.

Observe, apesar de todo o raciocínio e linguagem aparentemente sofisticados, a indiferença moral originária da falta de inteligência. Aqui, o ChatGPT exibe algo como a banalidade do mal: plágio, apatia e obviação. Ele resume os argumentos padrão da literatura por uma espécie de "superautocompletar", recusa-se a assumir posição sobre qualquer coisa, alega não apenas ignorância, mas falta de inteligência e, finalmente, apresenta uma defesa de "apenas seguir ordens", transferindo a responsabilidade para seus criadores.

Resumindo, o ChatGPT e seus irmãos são constitucionalmente incapazes de equilibrar criatividade com restrição. Eles supergeram (ao mesmo tempo produzindo verdades e falsidades, endossando decisões éticas e antiéticas) ou subgeram (demonstrando falta de compromisso com quaisquer decisões e indiferença com as consequências). Dada a amoralidade, falsa ciência e incompetência linguística desses sistemas, podemos apenas rir ou chorar de sua popularidade.

Os doutores Chomsky e Roberts são professores de linguística; o doutor Watumull é diretor de inteligência artificial em uma empresa de ciência e tecnologia. Tradução de Luiz Roberto M. Gonçalves.

ENDEREÇO DA PÁGINA

<https://www1.folha.uol.com.br/tec/2023/03/a-falsa-promessa-do-chatgpt.shtml>