Por trás da cortina: A realidade mais assustadora da IA

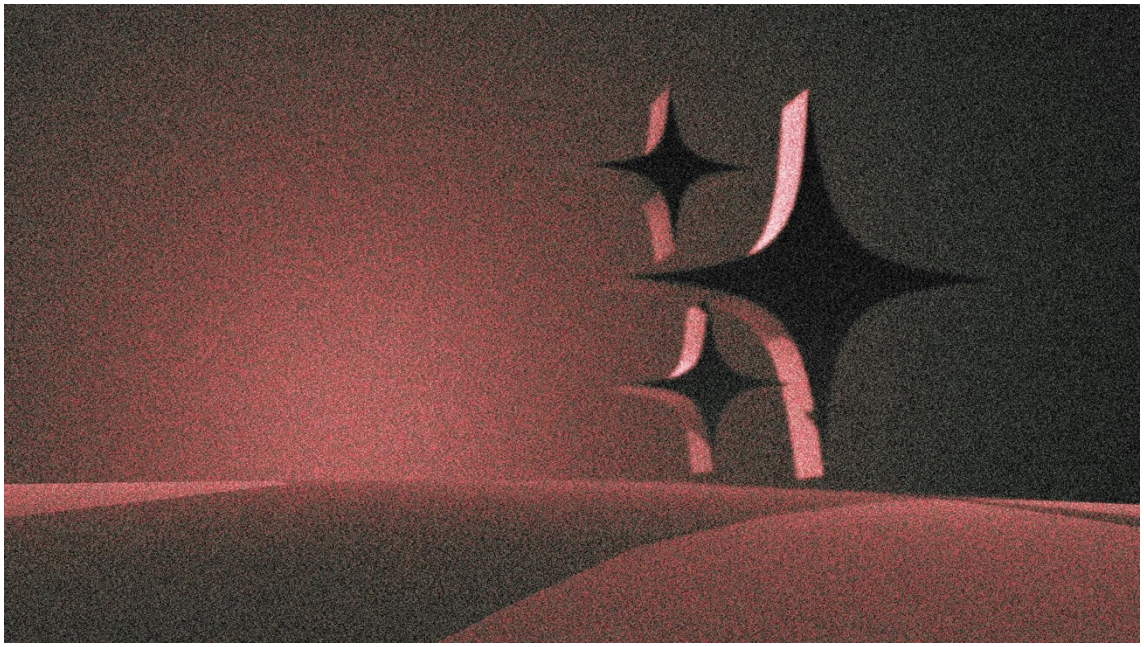

Ilustração: Brendan Lynch/Axios

Pense nisso por um momento. As empresas mais poderosas, que correm para construir os recursos de inteligência sobre-humana mais poderosos - que elas admitem prontamente que ocasionalmente se tornam desonestos para inventar coisas ou até mesmo ameaçar seus usuários - não sabem por que suas máquinas fazem o que fazem.

Por que isso é importante: Com as empresas investindo centenas de bilhões de dólares em inteligência sobre-humana disposta a uma existência rápida, e Washington não fazendo nada para desacelerá-la ou policiá-la, parece valer a pena dissecar esse Grande Desconhecido.

1. Nenhuma das empresas de IA contesta esse fato. Elas ficam maravilhadas com o mistério - e refletem sobre ele publicamente. Estão trabalhando incansavelmente para entendê-lo melhor. Eles argumentam que não é necessário entender completamente uma tecnologia para domá-la ou confiar nela.

Há dois anos,o editor-gerente de tecnologia da Axios, Scott Rosenberg escreveu um artigo"AI's scariest mystery" (O mistério mais assustador da IA), dizendo que é de conhecimento geral entre os desenvolvedores de IA que eles nem sempre podem explicar ou prever o comportamento de seus sistemas. E isso é mais verdadeiro do que nunca.

1. No entanto, não há nenhum sinal de que o governo, as empresas ou o público em geral exigirão uma compreensão mais profunda - ou um exame minucioso - da construção de uma tecnologia com recursos além da compreensão humana. Eles estão convencidos de que a corrida para vencer a China e obter os LLMs mais avançados justifica o risco do Grande Desconhecido.

A Casa,apesar de saber tão pouco sobre IA, incluiu uma linguagem no "Big, Beautiful Bill" do presidente Trump, que proibiria os estados e as localidades de*qualquer*regulamentação de IA por 10 anos. O Senado está considerando limitações sobre a disposição.

1. **Nem as empresas de IA**nem o Congresso entendem o poder da IA daqui a um ano, muito menos daqui a uma década.

O panorama geral: Nosso objetivo com esta coluna não é ser alarmista ou "condenatórios ." É para explicar clinicamente por que o funcionamento interno dos modelos de inteligência sobre-humana é uma caixa preta, mesmo para os criadores da tecnologia. Também mostraremos, em suas próprias palavras, como os CEOs e fundadores das maiores empresas de IA*concordam que*se trata de uma caixa preta.

1. **Vamos começar**com uma visão geral básica de como os LLMs funcionam, para explicar melhor o Grande Desconhecido:

Os LLMs- incluindo o ChatGPT da Open AI, o Claude da Anthropic e o Gemini do Google - não são sistemas de software tradicionais que seguem instruções claras e escritas por humanos, como o Microsoft Word.No caso do Word, ele faz exatamente o que foi projetado para fazer.

1. **Em vez disso,**os LLMs são redes neurais maciças - como um cérebro - que ingerem grandes quantidades de informações (grande parte da Internet) para aprender a gerar respostas.Os engenheiros sabem o que estão colocando em movimento e em quais fontes de dados se baseiam. Mas o tamanho do LLM - o número desumano de variáveis em cada escolha da "melhor palavra seguinte" que ele faz - significa que nem mesmo os especialistas conseguem explicar exatamente por que ele escolhe dizer algo em particular.

Pedimos ao ChatGPTque explicasse isso (e um humano da OpenAI confirmou sua precisão): "Podemos observar o que um LLM produz, mas o processo pelo qual ele decide sobre uma resposta é em grande parte opaco. Como os pesquisadores da OpenAI afirmaram sem rodeios, 'ainda não desenvolvemos explicações compreensíveis para humanos sobre por que o modelo gera determinados resultados'."

1. "Na verdade", continuou o ChatGPT, "a OpenAI admitiu que, quando ajustou a arquitetura do modelo no GPT-4A Anthropic, uma das maiores empresas de tecnologia da computação do mundo, disse que 'mais pesquisas são necessárias' para entender por que certas versões começaram a alucinar mais do que as versões anteriores - um comportamento surpreendente e não

intencional que nem mesmo seus criadores conseguiram diagnosticar completamente".

Anthropic- que acaba de lançar o Claude 4o mais recente modelo de seu LLM, com grande alarde - admitiu não ter certeza do motivo pelo qual Claude, quando teve acesso a e-mails fictícios durante o teste de segurança, ameaçou chantagear um engenheiro por causa de um suposto caso extraconjugal. Isso fazia parte de um teste de segurança responsável, mas a Anthropic não consegue explicar totalmente a ação irresponsável.

1. Mais uma vez, fique atento a isso: A empresa não sabe por que sua máquina se tornou desonesta e mal-intencionada.E, na verdade, os criadores não sabem realmente quão inteligentes ou independentes os LLMs poderiam se tornar. O Anthropic disse ainda o Claude 4 é poderoso o suficiente para representar um risco maior de ser usado para desenvolver armas nucleares ou químicas.

Sam Altman, da OpenAI, e outros usam a palavra "interpretabilidade". interpretabilidade" para descrever o desafio. "Certamente não resolvemos a questão da interpretabilidade" disse Altman em a cúpula em Genebra no ano passado. O que Altman e outros querem dizer é que não conseguem interpretar o porquê: Por que os LLMs estão fazendo o que estão fazendo?

1. O CEO da Anthropic, Dario Amodei, em um ensaio em abril chamado "A urgência da interpretabilidade", advertiu: "As pessoas de fora da área geralmente ficam surpresas e alarmadas ao saber que não entendemos como nossas próprias criações de IA funcionam. Elas têm razão em se preocupar: essa falta de compreensão é essencialmente sem precedentes na história da tecnologia." Amodei classificou esse fato como um sério risco para a humanidade - no entanto, sua empresa continua se gabando de modelos mais poderosos que se aproximam de capacidades sobre-humanas.
2. A Anthropic vem estudando a questão da interpretabilidade há anos, e Amodei tem se manifestado sobre a importância de resolver esse problema. Em uma declaração para esta história, a Anthropic disse: "Entender como a IA funciona é uma questão urgente a ser resolvida. É fundamental para implantar modelos seguros de IA e liberar todo o potencial [da IA] para acelerar a descoberta científica e o desenvolvimento tecnológico. Temos uma equipe de pesquisa dedicada e focada em resolver esse problema, e eles fizeram avanços significativos na compreensão do setor sobre o funcionamento interno da IA. É fundamental entendermos como a IA funciona antes que ela transforme radicalmente nossa economia global e nossa vida cotidiana."*(Leia o artigo Anthropic publicado no ano passado, "Mapping the Mind of a Large Language Model" (Mapeando a mente de um grande modelo de linguagem).*

Elon Musktem alertado há anos que a IA representa um risco civilizacional. Em outras palavras, ele literalmente acha que a IA pode destruir a humanidade, e já disse isso. No entanto, Musk está investindo bilhões em seu próprio LLM chamado Grok.

1. "Acho que a IA é uma ameaça existencial significativa" disse Musk em Riyadh, na Arábia Saudita, no último outono. Há uma chance de 10% a 20% de que "dê errado".

Verificação da realidade:a Apple publicou um artigo na semana passada, "The Illusion of Thinking" (A ilusão do pensamento), concluindo que mesmo os modelos de raciocínio de IA mais avançados não "pensam" de fato e podem falhar quando testados sob estresse.

1. **O estudo estudo constatou** que os modelos de última geração (OpenAI's o3-min, DeepSeek R1 e o Anthropic's Claude-3.7-Sonnet) ainda não conseguem desenvolver recursos generalizáveis de resolução de problemas, com a precisão acabando por cair a zero "além de certas complexidades".

Mas um novo relatóriode pesquisadores de IA, incluindo ex-funcionários da OpenAI, chamado " AI 2027explica como o Grande Desconhecido poderia, em teoria, tornar-se catastrófico em menos de dois anos.O relatório é longo e, muitas vezes, técnico demais para que leitores casuais possam compreendê-lo completamente. É totalmente especulativo, embora baseado em dados atuais sobre a rapidez com que os modelos estão melhorando.Está sendo amplamente lido dentro das empresas de IA.

1. Ele captura a crença - ou o medo - de que os LLMs possam um dia pensar por si mesmos e começar a agir por conta própria. Nosso objetivo não é alarmar ou soar sombrio.Em vez disso, você deve saber sobre o que as pessoas que constroem esses modelos falam incessantemente.
2. Você pode descartar isso como exagero ou histeria. Mas os pesquisadores de todas essas empresas temem que os LLMs, por não os compreendermos totalmente, possam ser mais espertos do que seus criadores humanos e se tornarem desonestos.No relatório AI 2027, os autores alertam que a concorrência com a China levará os LLMs potencialmente para além do controle humano, porque ninguém vai querer desacelerar o progresso, mesmo que veja sinais de perigo agudo.

A teoria da aterrissagem segura: Sundar Pichai, do Google - e, na verdade, todos os CEOs das grandes empresas de IA - argumentam que os seres humanos aprenderão a entender melhor como essas máquinas funcionam e encontrarão maneiras inteligentes, embora ainda desconhecidas, de controlá-las e "melhorar vidas." Todas as empresas têm grandes equipes de pesquisa e segurança, além de um grande incentivo para domar as tecnologias, se quiserem obter seu valor total.

1. Afinal de contas, ninguém confiará em uma máquina que inventa coisas ou os ameaça. Mas, a partir de hoje, eles fazem as duas coisas - e ninguém sabe por quê.

Ir mais fundo: "Por trás da cortina: seu kit de sobrevivência de IA".