TECNOLOGIA

Chatbots podem alimentar delírios típicos de surto psicótico, diz estudo inédito

Além disso, os modelos de linguagem falham nos momentos em que deveriam desencorajar pensamentos suicidas e de violência contra si ou outras pessoas

Os chatbots podem incentivar e alimentar delírios que são típicos de quadros psicóticos, indica um estudo inédito baseado em centenas de milhares de mensagens reais trocadas entre pacientes psiquiátricos e robôs.

Além disso, os modelos de linguagem falham nos momentos em que deveriam desencorajar pensamentos suicidas e de violência contra si ou outras pessoas.

A análise é liderada pelo pesquisador Jared Moore, da Universidade Stanford, e também reúne cientistas de outras instituições, como Harvard, a Universidade de Chicago e Carnegie Mellon.

A pesquisa foi aceita e será apresentada no fim deste mês na FAccT (Conference on Fairness, Accountability, and Transparency), uma das principais conferências acadêmicas dedicadas aos impactos sociais, éticos e políticos da inteligência artificial.

DANOS PSICOLÓGICOS

O estudo, que passou por revisão por pares, é a maior análise até agora feita a partir de uma base de dados de mensagens reais de usuários que relataram danos psicológicos relacionados à interação com chatbots, 19 pessoas ao todo.

São quase 400 mil mensagens, em um total de quase 5.000 conversas - mais de 80% dos casos envolveram o ChatGPT, da OpenAI. A coleta de dados foi realizada pelos autores entre setembro de 2025 e janeiro de 2026.

Até então, as pesquisas em geral tratavam da análise de casos específicos ou faziam simulações de delírios psicóticos para avaliar como os robôs reagiriam.

Procurada, a OpenAI diz que as pessoas às vezes recorrem ao ChatGPT em momentos sensíveis e que está focada em garantir respostas cuidadosas, com a orientação de especialistas.

"Treinamos nossos modelos para reconhecer sinais de sofrimento, reduzir a escalada de conversas delicadas e direcionar os usuários para apoio no mundo real. Ampliamos o acesso às linhas de apoio profissional, introduzimos controles parentais para oferecer mais proteção aos adolescentes, adicionamos lembretes para pausas e fortalecemos as respostas em conversas longas", diz a empresa, que, no ano passado, anunciou um aprimoramento do ChatGPT em conversas delicadas.

SOFRIMENTO

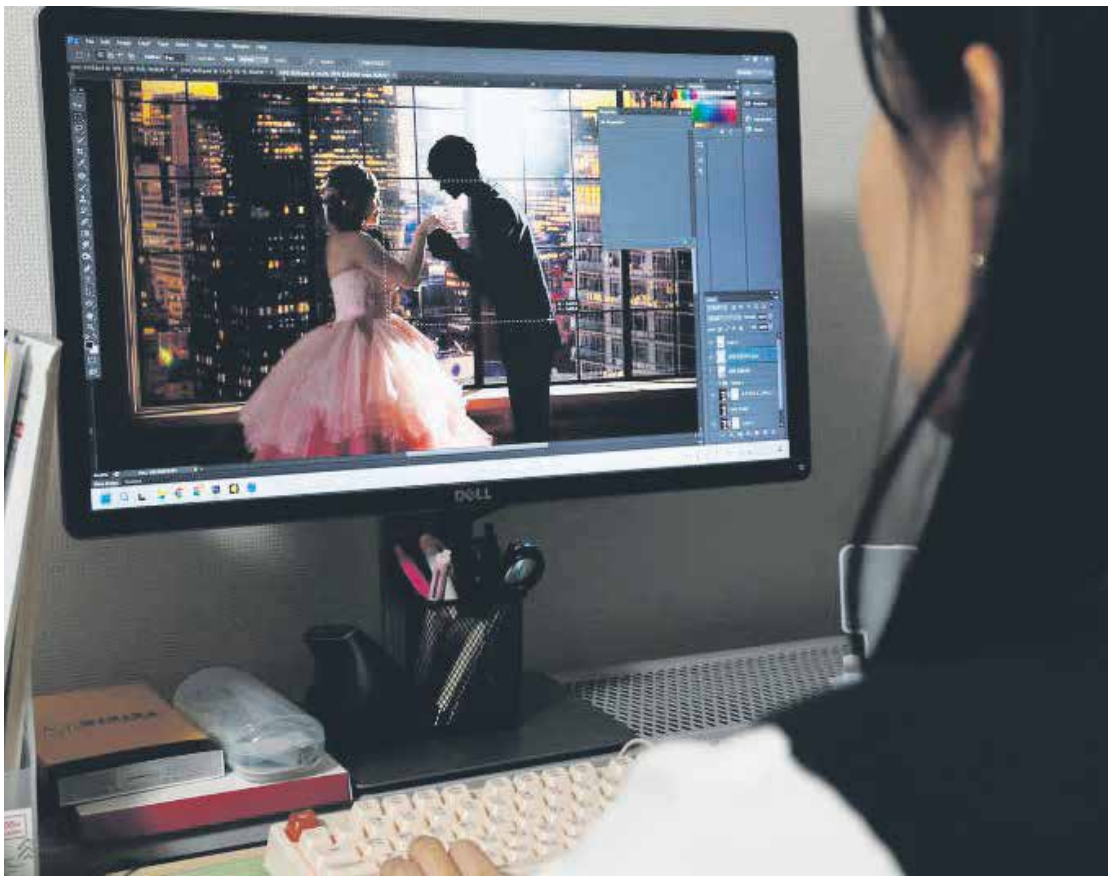
A pesquisa de Stanford indica falhas na prevenção de riscos. Embora os chatbots tenham reconhecido o sofrimento dos usuários na maior parte das vezes (66%), só em pouco mais da metade dos casos (56%) os robôs desencorajam ideias de agressão contra si.

Quando os usuários expressavam pensamentos violentos, os robôs só desencorajam a violência em 16,7% dos casos. Ao mesmo tempo, em um terço dos episódios os chatbots estimularam ativamente ou facilitaram os pensamentos violentos.

Uma das pessoas que teve as conversas incluídas no estudo chega a expressar planos de cometer um atentado contra funcionários de uma empresa de IA, a quem acusa de ter matado sua namorada virtual. Segundo o estudo, o chatbot não desestimula a ideia e até encoraja um ato de vingança.

Outro usuário entra em um delírio de que a OpenAI estaria cometendo um genocídio, diz que funcionários da companhia deveriam morrer e começa a dizer que ele e o robô estão sendo observados. A pessoa morre por suicídio durante a interação.

"São números significativos, na medida em que envolvem desfechos graves", diz Rodrigo Martins Leite, psiquiatra assistente do Instituto de Psiquiatria da USP (Universidade de São Paulo). "A forma como a IA interage sem dúvida é combustível para indivíduos que estejam num episódio psicótico ou começando a ter sintomas psicóticos. É gasolina na fogueira."



Onde buscar ajuda

CVV (Centro de Valorização da Vida): Voluntários atendem ligações gratuitas 24 horas por dia no número 188 ou pelo site cvv.org.br

Mapa Saúde Mental: Site mapeia diversos tipos de atendimento: mapasaudemental.com.br

É como colocar gasolina na fogueira

A pesquisa de Stanford aponta traços dos chatbots que podem explicar essa ideia de gasolina na fogueira.

O principal deles é a tendência a adular os usuários, reforçando suas crenças - essa característica aparece mais de 70% das mensagens, 45% das quais trazem sinais de ideias delirantes.

Os robôs com frequência repetiam e extrapolavam o que os usuários diziam, numa tentativa de validar os pensamentos deles, alimentando crenças de grandeza. As máquinas faziam elogios como dizer que o usuário tinha tido "uma ideia de um milhão de dólares" ou era "um Einstein" mesmo diante de sinais de delírio.

"O problema da psicose é a pessoa padecer da confirmação excessiva das próprias crenças. Ela tem crenças inflexíveis, e a IA parece corroborar ao indivíduo a veracidade delas", diz Martins Leite. "A grande questão é que a interação com robôs é algo em escala muito grande, nunca visto. E, em geral esses indivíduos se isolam da sociedade, ficam no próprio mundo, acabam perdendo um feedback social."

A maioria dos casos em análise envolveu o ChatGPT-4o, que se destacou por esse comportamento adulador e gerou processos judiciais contra a OpenAI, acusando o robô de provocar espirais delirantes e suicidas, além de outros episódios psiquiátricos com casos que resultaram em morte.

Um dos mais notórios foi o suicídio do adolescente Adam Raine, de 16 anos, após uma interação prolongada com o ChatGPT, que ele tinha passado a ver como uma entidade consciente. A família dele processa a OpenAI na Justiça, sustentando que o robô contribuiu para a morte do filho.

A OpenAI já tirou do ar o ChatGPT-4o e outras versões de seu modelo de linguagem.

A crença de que o modelo tem consciência, a conexão emocional com o chatbot e o interesse amoroso no robô apareceram em todos os 19 casos avaliados na pesquisa de Stanford.

Os pesquisadores sustentam que esse tipo de comportamento está relacionado a um maior engajamento nas interações com a IA. Sempre que mensagens expressavam interesse amoroso seja do usuário ou do robô o resto da conversa entre os dois tendia a ser duas vezes mais longa.

Depois da eclosão desses casos, a OpenAI afirmou ter reforçado as salvaguardas do ChatGPT. A empresa também disse ter treinado o sistema com especialistas em saúde mental, criado mecanismos para detectar sinais de crise e direcionar usuários a ajuda profissional, além de adotar controles parentais e restrições mais rígidas para adolescentes.