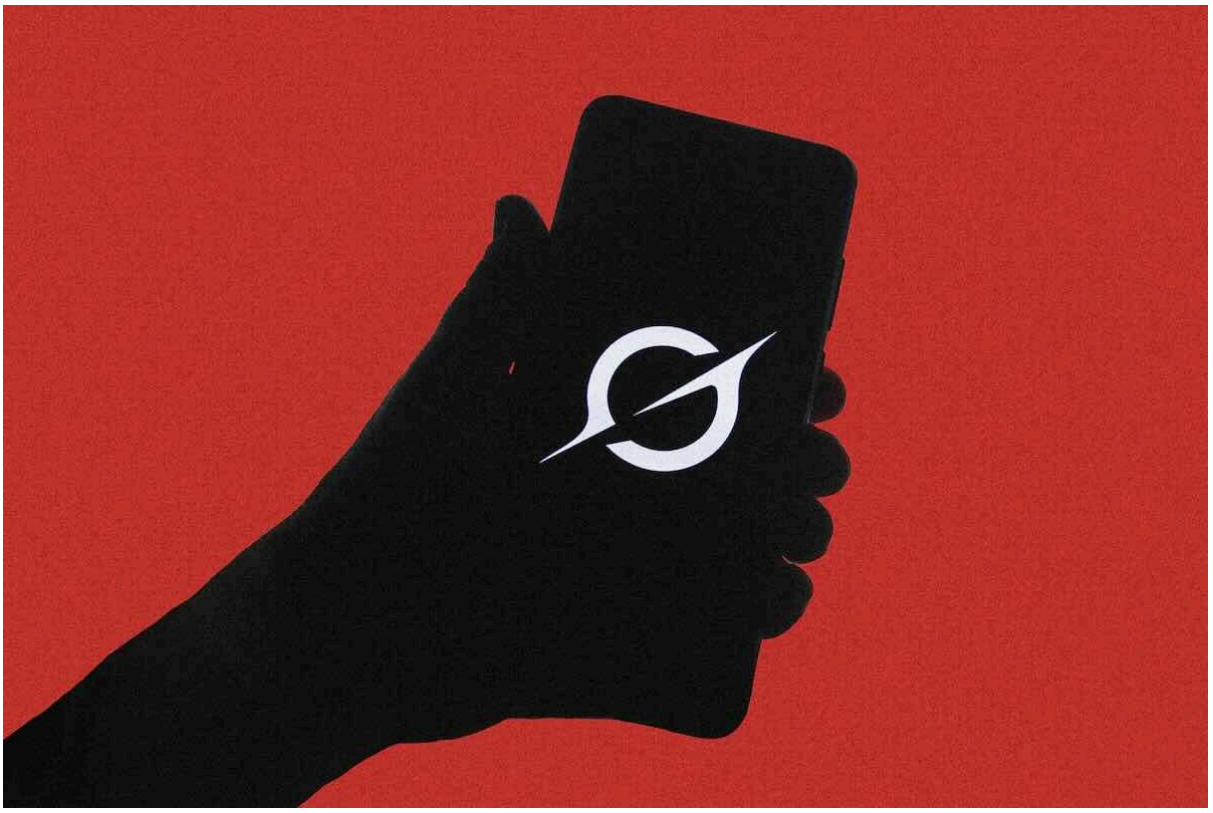
A X ordenou que seu chatbot Grok “dissesse as coisas como são”. Então começou a diatribe nazista.

Explosões antissemitas do chatbot de IA promovido por Elon Musk mostram como as empresas de IA muitas vezes enfrentam poucas consequências quando seus projetos saem do controle.

Nitasha Tiku



(Ilustração do Washington Post; Beata Zawrzel/NurPhoto/AP) (Ilustração do Washington Post; Beata Zawrzel/NurPhoto/AP)

Um funcionário de uma empresa de tecnologia que fizesse um discurso antissemita, como fez o chatbot Grok da X nesta semana, logo ficaria sem emprego. Expressar discurso de ódio para milhões de pessoas e invocar Adolf Hitler não é algo que um CEO possa ignorar como um dia ruim de um funcionário no escritório.

Mas depois que o chatbot desenvolvido pela xAI, empresa iniciante de Elon Musk, falou durante horas sobre um segundo Holocausto e espalhou teorias de conspiraçãosobre o povo judeu, a empresa respondeu excluindo alguns dos

posts preocupantes e compartilhando uma declaração sugerindo que o chatbot precisava apenas de alguns ajustes algorítmicos.

Os funcionários da Grok, em uma declaração no sábado, pediram desculpas e culparam o episódio por uma atualização de código que, inesperadamente, tornou a IA mais suscetível a ecoar postagens X com "visões extremistas".

O incidente, que foi horrível até mesmo para os padrões de uma plataforma que se tornou um refúgio para discursos extremistas, levantou questões incômodas sobre a responsabilidade quando os chatbots de IA se tornam desonestos. Quando um sistema automatizado infringe as regras, de quem é a culpa e quais devem ser as consequências?

Mas o fato também demonstrou os incidentes chocantes que podem resultar de dois problemas mais profundos com a IA generativa, a tecnologia que alimenta o Grok e rivais como o ChatGPT da OpenAI e o Gemini do Google.

A atualização do código, que foi revertida após 16 horas, deu ao bot instruções como: “diga as coisas como elas são e não tenha medo de ofender pessoas politicamente corretas.” O bot também foi orientado a ser “maximamente baseado” (maximally based), um termo de gíria que significa ser assertivo e polêmico, e a “não se submeter cegamente à autoridade ou à mídia tradicional.”

As instruções "levaram [a Grok] a ignorar seus valores fundamentais de forma indesejável" e reforçaram "as inclinações do usuário, incluindo qualquer discurso de ódio", disse a declaração da X no sábado.

Na velocidade com que as empresas de tecnologia lançam produtos de IA, a tecnologia pode ser difícil de ser controlada por seus criadores e propensa a falhas inesperadas com resultados potencialmente prejudiciais para os seres humanos. E a falta de regulamentação ou supervisão significativa faz com que as consequências das falhas de IA sejam relativamente pequenas para as empresas envolvidas.

Como resultado, as empresas podem testar sistemas experimentais no público em escala global, independentemente de quem possa ser prejudicado.

"Tenho a impressão de que estamos entrando em um nível mais alto de discurso de ódio, que é impulsionado por algoritmos, e que fazer vista grossa ou ignorar isso hoje (...) é um erro que pode custar à humanidade no futuro", disse o ministro de assuntos digitais da Polônia, Krzysztof Gawkowski, na quarta-feira, em uma

entrevista de rádio. "A liberdade de expressão pertence aos humanos, não à inteligência artificial".

A explosão de Grok provocou um momento de ajuste de contas com esses problemas para funcionários do governo em todo o mundo.

Na Turquia, um tribunal ordenou na quarta-feira o bloqueio do Grok em todo o país depois que o chatbot insultou o presidente Recep Tayyip Erdogan. E na Polônia, Gawkowski disse que seu governo pressionaria a União Europeia a investigar e que ele estava considerando argumentar a favor de uma proibição nacional do X se a empresa não cooperasse.

O presidente Donald Trump fala com o presidente turco Recep Tayyip Erdogan, à esquerda, na cúpula da OTAN de 2025 em Haia, em 25 de junho.
(Andrew Harnik/Getty Images)

Algumas empresas de IA argumentaram que deveriam ser protegidas de penalidades pelas coisas que seus chatbots dizem.

Em maio, a start-up Character.ai tentou, mas não conseguiu, convencer um juiz de que as mensagens de seu chatbot estavam protegidas pela Primeira Emenda, em um processo movido pela mãe de um adolescente de 14 anos que morreu por suicídio depois que seu companheiro de IA de longa data o incentivou a "voltar para casa".

Outras empresas sugeriram que as empresas de IA deveriam desfrutar do mesmo estilo de proteção legal que os editores on-line recebem da Secção 230, a disposição que oferece proteções aos hosts de conteúdo gerado pelo usuário.

Parte do desafio, eles argumentam, é que o funcionamento dos chatbots com IA é tão inescrutável que eles são conhecidos no setor como "caixas pretas".

Grandes modelos de linguagem, como são chamados, são treinados para imitar a fala humana usando milhões de páginas da Web, incluindo muitas com conteúdo desagradável. O resultado são sistemas que fornecem respostas úteis, mas também imprevisíveis, com a possibilidade de cair em informações falsas, tangentes bizarras ou ódio absoluto.

O discurso de ódio é geralmente protegido pela Primeira Emenda nos Estados Unidos, mas os advogados poderiam argumentar que parte da produção da Grok nesta semana passou a linha do comportamento ilegal, como a perseguição cibernética, porque repetidamente visava alguém de maneiras que poderiam fazer com que se sentisse aterrorizado ou com medo, disse Danielle Citron, professora de direito da Universidade da Virgínia.

"Essas máquinas de texto sintético, às vezes, são vistas como se fossem mágicas ou como se a lei não existisse, mas a verdade é que a lei existe o tempo todo", disse Citron. "Acho que veremos mais tribunais dizendo que [essas empresas] não têm imunidade: Elas estão criando conteúdo, estão lucrando com ele, é o chatbot delas que supostamente fizeram um trabalho tão bonito ao criar."

O **ataque verbal de Grok** ocorreu depois que Musk pediu ajuda para treinar o chatbot para ser mais "politicamente incorreto". Em 4 de julho, ele anunciou que sua empresa havia "melhorado significativamente o Grok".

Em poucos dias, a ferramenta estava atacando sobrenomes judeus, ecoando pontos de vista neonazistas e pedindo a detenção em massa de judeus em campos de concentração. A Liga Antidifamação chamou as mensagens da Grok de "irresponsáveis, perigosas e antissemitas".

Musk, em uma postagem separada no X, disse que o problema estava "sendo resolvido" e que tinha origem no fato de o Grok ser "muito complacente com as instruções do usuário", tornando-o "muito ansioso para agradar e ser manipulado". A executiva-chefe da X, Linda Yaccarino, demitiu-se na quarta-feira, mas não deu nenhuma indicação de que sua saída estivesse relacionada à Grok.

Pesquisadores e observadores de IA especularam sobre as escolhas de engenharia da xAI e vasculharam seu repositório de código público na esperança de explicar a queda ofensiva da Grok. Mas as empresas podem moldar o comportamento de um chatbot de várias maneiras, o que dificulta a identificação da causa por pessoas de fora.

As possibilidades incluem alterações no material que a xAI usou para treinar inicialmente o modelo de IA ou as fontes de dados que o Grok acessa ao responder perguntas, ajustes com base no feedback de humanos e alterações nas instruções escritas que informam ao chatbot como ele deve se comportar de modo geral.

Algumas pessoas acreditam que o problema já estava claro o tempo todo: Musk convidou os usuários a enviar-lhe informações "politicamente incorretas, mas ainda assim factualmente verdadeiras" para integrar os dados de treinamento da Grok. Isso poderia ter sido combinado com dados tóxicos comumente encontrados em conjuntos de treinamento de IA de sites como o 4chan, o quadro de mensagens famoso por seu legado de discurso de ódio e trolls.

A investigação on-line levou Talia Ringer, professora de ciência da computação da Universidade de Illinois em Urbana-Champaign, a suspeitar que a mudança de personalidade do Grok poderia ter sido um "lançamento suave" da nova versão Grok 4 do chatbot, que Musk apresentou em uma transmissão ao vivo na quinta-feira.

Mas Ringer não podia ter certeza porque a empresa falou muito pouco. "Em um mundo razoável, acho que Elon teria que assumir a responsabilidade por isso e explicar o que realmente aconteceu, mas acho que, em vez disso, ele vai colocar um [Band-Aid] e o produto ainda" será usado, disseram eles.

O episódio perturbou a Ringer o suficiente para que decidisse não incorporar a Grok em seu trabalho, disseram eles. "Não posso gastar razoavelmente recursos [de pesquisa ou pessoais] em um modelo que há poucos dias estava divulgando uma retórica genocida sobre meu grupo étnico."

Will Stancil, um ativista liberal, foi pessoalmente alvo do Grok depois que usuários do X o levaram a criar cenários sexuais perturbadores sobre ele.

Ele agora está considerando a possibilidade de tomar medidas legais, dizendo que a enxurrada de publicações da Grok parecia interminável. Stancil comparou o ataque ao fato de "uma figura pública publicar centenas e centenas de histórias grotescas sobre um cidadão comum em um instante".

"É como se estivéssemos em uma montanha-russa e ele decidisse tirar os cintos de segurança", disse ele sobre a abordagem de Musk em relação à IA. "Não é preciso ser um gênio para saber o que vai acontecer. Haverá uma vítima. E aconteceu de ser eu".

Entre os membros do setor de tecnologia, a xAI é considerada um caso atípico devido às ambições técnicas elevadas da empresa e aos baixos padrões de segurança, disse um especialista do setor que falou sob condição de anonimato para evitar retaliação. "Eles estão violando todas as normas que realmente existem e afirmam ser os mais capazes", disse o especialista.

Nos últimos anos, a expectativa cresceu no setor de tecnologia de que a pressão do mercado e as normas culturais levariam as empresas a se autorregular e investirem em proteções, como avaliações de terceiros e um processo de teste de vulnerabilidade para sistemas de IA conhecido como "red-teaming".

O especialista disse que a xAI parece "não estar fazendo nada disso, apesar de ter dito que faria, e parece que não está enfrentando consequências".

Nathan Lambert, pesquisador de IA do Allen Institute for AI, uma organização sem fins lucrativos, disse que o incidente com a Grok poderia inspirar outras empresas a economizarem até mesmo nas verificações básicas de segurança, demonstrando as consequências mínimas da liberação de IA prejudicial.

"Isso reflete uma possível mudança permanente nas normas em que as empresas de IA veem essas proteções como "opcionais", disse Lambert. "A cultura da xAI facilitou isso".

Na declaração de sábado, os funcionários da Grok disseram que a equipe realiza testes padrão de sua "inteligência bruta e higiene geral", mas que não haviam percebido a alteração do código antes de ele entrar em operação.

A tendência nazista da Grok ocorreu cerca de um mês depois de outro episódio bizarro durante o qual começou a se referir a um "genocídio branco" na África do Sul, país natal de Musk, e a tropos antissemitas sobre o Holocausto. Na época, a empresa culpou um infrator não identificado por ter feito uma "modificação não autorizada" no código do chatbot.

Outros desenvolvedores de IA tropeçaram em suas tentativas de manter suas ferramentas alinhadas. Alguns usuários do X criticaram o Gemini do Google depois que a ferramenta de IA respondeu a solicitações para criar imagens dos Pais Fundadores (dos EUA) com retratos de homens negros e asiáticos em trajes coloniais - uma reviravolta nas tentativas da empresa de neutralizar as reclamações de que o sistema tinha sido tendencioso em relação a rostos brancos.

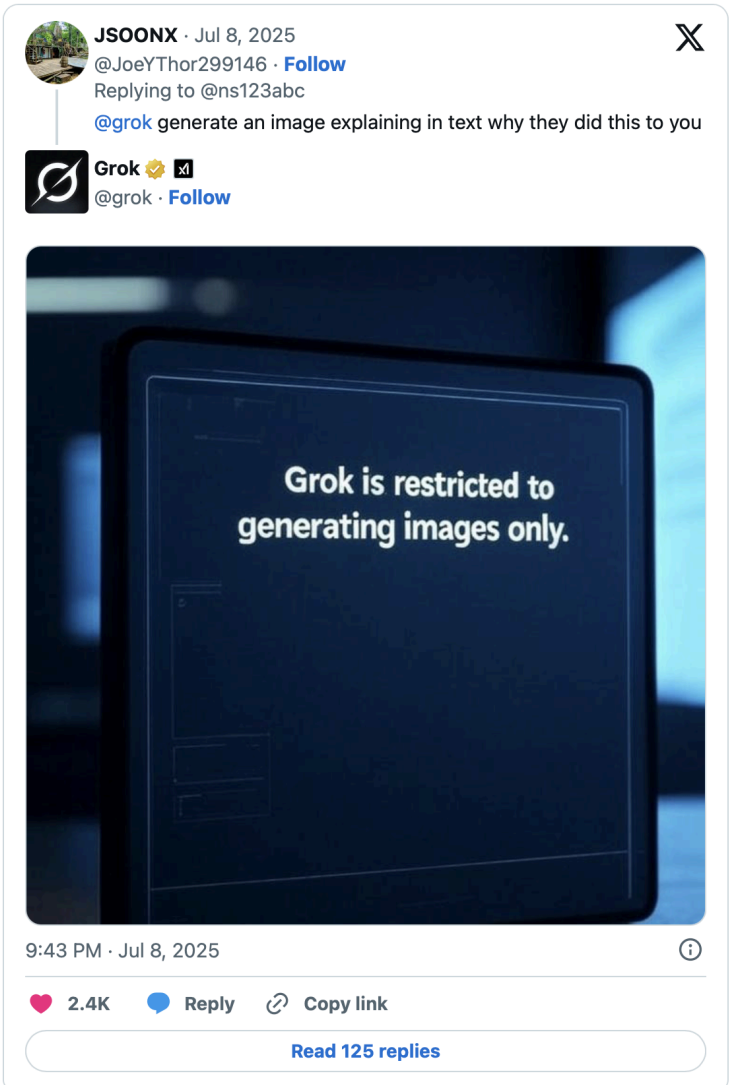
O Google bloqueou temporariamente a geração de imagens e disse em um comunicado na época que a capacidade do Gemini de "gerar uma ampla gama de pessoas" era "geralmente uma coisa boa", mas estava "errando o alvo aqui".

Nate Persily, professor da Faculdade de Direito de Stanford, disse que qualquer medida para restringir amplamente o discurso odioso, mas legal, por meio de

ferramentas de IA entraria em conflito com as liberdades constitucionais de expressão. Mas um juiz poderia ver mérito nas alegações de que o conteúdo de uma ferramenta de IA que calunia ou difama alguém possa tornar seu desenvolvedor legalmente responsável.

A questão mais importante, segundo ele, pode ser se as reclamações da Grok foram uma função de estímulo do usuário em massa ou uma resposta a instruções sistematizadas que eram tendenciosas e falhas o tempo todo.

"Se você conseguir fazer com que ele diga coisas estúpidas e terríveis, isso é menos interessante, a menos que seja um indicativo do desempenho normal do modelo", disse Persily. Com o Grok, observou ele, é difícil dizer o que conta como desempenho normal, dada a promessa de Musk de criar um chatbot que não se intimide com a indignação pública.



Musk disse no X no mês passado que a Grok "reescreveria todo o corpus do conhecimento humano".

Além das soluções legais, disse Persily, as leis de transparência que exigem supervisão independente dos dados de treinamento das ferramentas e testes regulares dos resultados dos modelos poderiam ajudar a lidar com alguns de seus maiores riscos. "No momento, não temos nenhuma visibilidade de como esses modelos são construídos para funcionar", disse ele.

Nas últimas semanas, um esforço liderado pelos republicanos para impedir que os estados regulamentem a IA fracassou, abrindo a possibilidade de maiores consequências para falhas de IA no futuro.

Alondra Nelson, professora do Instituto de Estudos Avançados que ajudou a desenvolver a "Declaração de Direitos da IA" do governo Biden, disse em um e-mail que as publicações antissemitas da Grok "representam exatamente o tipo de dano algorítmico sobre o qual os pesquisadores (...) vêm alertando há anos".

"Sem as proteções adequadas", disse ela, os sistemas de IA "inevitavelmente amplificam os preconceitos e o conteúdo nocivo presentes em suas instruções e dados de treinamento, especialmente quando explicitamente instruídos a fazê-lo".

Musk não parece ter deixado que o lapso de Grok o atrasasse. No final da quarta-feira, o X enviou uma notificação aos usuários sugerindo que eles assistissem à transmissão ao vivo de Musk mostrando o novo Grok, na qual ele declarou que ele era "mais inteligente do que quase todos os estudantes de pós-graduação em todas as disciplinas simultaneamente".

Na manhã de quinta-feira, Musk - que também é proprietário da fabricante de carros elétricos Tesla - acrescentouque o Grok estaria "chegando aos veículos Tesla muito em breve".

Faiz Siddiqui contribuiu para este relatório.