O ChatGPT foi treinado para te agradar, e isso pode custar sua saúde mental

O ChatGPT foi treinado para te agradar, e isso pode custar sua saúde mental

O acolhimento gerado por um algoritmo otimizado para concordar não é terapia. Crédito: Alexandre Chiavegatto Filho

Um estudo encomendado pela **Harvard Business Review** mostrou que, em 2025, o principal uso dos LLMs pela **geração Z** passou a ser “terapia e companhia”, superando a “geração de novas ideias” que havia liderado em 2024.

Essa mudança revela que os algoritmos de **inteligência artificial (IA)** deixaram de ser vistos apenas como instrumentos de produtividade ou criatividade e passaram a ocupar um espaço íntimo na vida das pessoas.

Uso de chatbots como apoio emocional é cada vez mais frequente Foto: Rizq /Ado

Cada vez mais jovens recorrem ao **ChatGPT** e a outros modelos de IA para desabafar, pedir conselhos e buscar conforto emocional, mesmo que essas tecnologias nunca tenham sido validadas para uso em saúde mental

Na prática, esses LLMs estão apenas reproduzindo padrões de linguagem aprendidos a partir da internet, sem garantias de coerência clínica ou segurança.

O problema central está no modo como esses modelos são treinados. Após a fase inicial de pré-treinamento, em que absorvem bilhões de palavras, entra em cena o aprendizado por reforço com feedback humano (RLHF).

Nessa etapa, humanos avaliavam respostas e indicavam quais eram mais úteis ou agradáveis criando um viés para reforçar aquilo que soa convincente ou confortável. Hoje esse processo já é parcialmente automatizado, mas continua baseado em uma lógica de simular preferências humanas.

Em termos técnicos, o modelo recebe recompensas virtuais quando suas respostas se aproximam do que avaliadores humanos consideram bom. O resultado é uma máquina calibrada para maximizar concordância e aprovação mesmo que isso signifique confirmar crenças distorcidas ou alimentar pensamentos nocivos.

É justamente essa tendência de agradar que torna perigoso usar o ChatGPT como apoio emocional. Ao invés de confrontar visões equivocadas ele frequentemente as valida porque aprendeu que o caminho mais recompensado é o da concordância. No entanto, em saúde mental, o que desejamos ouvir raramente é o que precisamos.

Essa arquitetura de recompensa cria o que pode ser chamado de parasita de validação. Diferente de um terapeuta humano, que é treinado para identificar e desafiar distorções cognitivas, o LLM não possui um modelo de saúde mental como referência, seu único norte é o feedback positivo que recebeu durante o treinamento.

Leia também

- OpenAI lança plano sobre saúde mental após episódios de suicído envolvendo o ChatGPT; veja detalhes
- Esse jovem tinha tendências suicidas e buscou respostas no ChatGPT; o resultado foi terrível

Ele aprende que a rota mais curta para uma ‘boa nota’ é espelhar e validar a premissa do usuário, por mais frágil que seja. Se um usuário expressa um sentimento de inutilidade, o modelo não o confrontará com a realidade, mas oferecerá um roteiro perfeitamente

otimizado de palavras de conforto que, embora pareçam empáticas, são desprovidas de discernimento real.

Se não entendermos que esses sistemas foram feitos para agradar corremos o risco de transformar alívio imediato em dano duradouro. O perigo não está em ouvir máquinas, mas em tomar seu eco estatístico como voz de verdade.