

ChatGPT e ideologia



Por **HENRIQUE BRAGA & MARCELO MÓDOLO***

O bot é capaz de nos manipular?

“Olá! Eu sou apenas um programa de computador desenvolvido pela OpenAI e não tenho opiniões, crenças, sentimentos ou afinidade com qualquer ideologia. Eu sou projetado para fornecer informações imparciais e objetivas, além de responder às suas perguntas com base no conhecimento que fui treinado. Minha função é ajudar a fornecer informações e esclarecimentos, mas não tenho a capacidade de tomar partido, ter preferências pessoais ou emitir julgamentos de valor.”

O simpático parágrafo anterior foi produzido pelo famigerado ChatGPT, em resposta a um questionamento sobre suas preferências ideológicas. Como muitos usuários já puderam comprovar, o robô afirma – sempre que possível – sua dita imparcialidade, seu amor aos dados, sua objetividade diante do mundo. Para um leigo em questões de Linguagem (ou de Filosofia, ou de Política), a utopia pode soar real: uma vez que o *chat* não é um ser-humano, por que ele seria movido por paixões, vontades ou ideologias? O que impediria nosso oráculo digital de produzir textos que não sejam contra, nem a favor, muito pelo contrário? Por que ele não seria capaz de nos fornecer a verdade, nada mais que a verdade?

Para estudiosos da Análise do Discurso ou da Linguística Cognitiva, porém, a resposta do *chatbot* traz incoerências em sua origem. Mais do que isso, pode ser uma eficaz e conveniente estratégia de manipulação.

Dialogismo e perspectivas

Segundo a conhecida abordagem de Mikhail Bakhtin (1895 – 1975), os textos são atos responsivos, que, por sua vez, geram novas respostas. Ao elaborar sua mensagem, o enunciador dá sequência a um amplo diálogo social, em que os textos se retomam, seja em concordância, seja em divergência. O surgimento deste artigo, por exemplo, é decorrente de diversos textos que tratam do posicionamento (ou da suposta neutralidade) de sistemas generativos de Inteligência Artificial (IA).

De uma outra abordagem, como já tratamos em colunas anteriores, os próprios signos linguísticos são, conforme a visão cognitivista da linguagem, perspectivados. Entre outros, pode ser lembrado o exemplo dos termos “costa” e “litoral”: ainda que possam se referir à mesma faixa de areia, cada um dos vocábulos traduz uma perspectiva diferente sobre tal espaço geográfico. A distância entre a coisa representada e o signo é inerente à linguagem e, nesse vácuo, instauram-se as visões de mundo.

Com base nesses princípios, já seria no mínimo suspeita a afirmação de que a IA “não tem opiniões ou crenças”, ou mesmo que ela se limita a fornecer “informações imparciais”. O exemplo a seguir, porém, explicita a distância entre tal discurso e a prática do sistema.

Pego no flagra

É bastante conhecido e razoavelmente aceito o provérbio italiano “Traduttore, traditore” (“Tradutor, traidor”), ao qual se recorre para afirmar que as traduções deixariam sempre escapar algo do texto original. Contudo, foi um pouco mais grave o que aconteceu quando solicitamos ao ChatGPT que traduzisse um artigo sobre impactos da IA na educação, no qual constava o seguinte excerto:

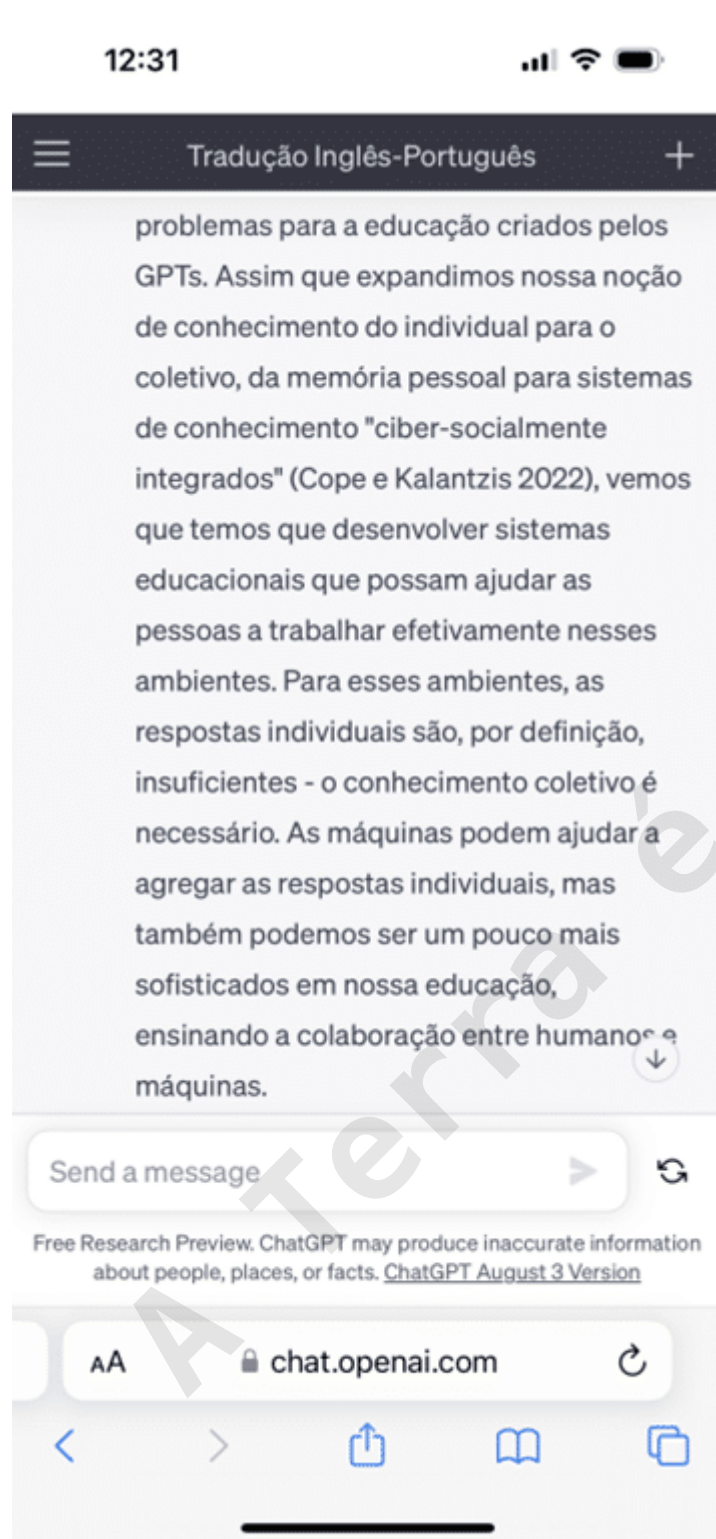
(...) As soon as we expand our notion of knowledge from individual to collective, from personal memory to “cyber-social” knowledge systems (Cope and Kalantzis 2022), we run into much bigger problems with generative AI. On the basis of an analysis of their foundational architecture and algorithmic processes—to the extent that they are disclosed (Open AI 2023)—and on analysis of the development of chatbots and statistical language models, we suggest that C-LLM s are also deeply harmful to a social understanding of knowledge and learning in the following ways.

Tzirides, Anastasia Olga (Olnancy), Gabriela Zapata, Akash Saini, Duane Sears Smith, Bill Cope, Mary Kalantzis, Vania Castro, Theodora Kourkoulou, John Jones, Rodrigo Abrantes da Silva, Jen Whiting and Nikoleta Polyxeni Kastania, “Generative AI: Implications and Applications for Education,” *arXiv*, 2305.07605, 2023, doi: <https://doi.org/10.48550/arXiv.2305.07605>.

Ao final dessa passagem, que traz importantes alertas sobre o uso da IA, lê-se algo como “sugerimos que os C-LLMs também são profundamente prejudiciais para uma compreensão social do conhecimento e da aprendizagem das seguintes maneiras”. Em seguida, os pesquisadores elencariam os riscos de uma toda poderosa IA, popularizada, determinar quais são as respostas “corretas” para determinada pergunta, o que poderia limitar a diversidade de visões. Esse mesmo trecho foi assim traduzido pelo ChatGPT:

(...) Assim que expandimos nossa noção de conhecimento do individual para o coletivo, da memória pessoal para sistemas de conhecimento “ciber-socialmente integrados” (Cope e Kalantzis 2022), vemos que temos que desenvolver sistemas educacionais que possam ajudar as pessoas a trabalhar efetivamente nesses ambientes. Para esses ambientes, as respostas individuais são, por definição, insuficientes - o conhecimento coletivo é necessário. As máquinas podem ajudar a agregar as respostas individuais, mas também podemos ser um pouco mais sofisticados em nossa educação, ensinando a colaboração entre humanos e máquinas.

Tradução elaborada pelo ChatGPT



Nesse episódio, é flagrante como a IA “sem opiniões, sentimentos ou ideologia” traduziu de forma fraudulenta a passagem que lhe era crítica. Em vez do alerta, segundo o qual os sistemas generativos de IA podem ser “*deeply harmful to a social understanding of knowledge and learning*” (“profundamente prejudiciais para a compreensão social de conhecimento e aprendizagem”), o sistema “tomou a liberdade” de trocar a versão original por outra, mais benevolente consigo. Em momento algum o fragmento em inglês diz algo semelhante a “podemos ser um pouco mais sofisticados em nossa educação, ensinando a colaboração entre humanos e máquinas”. Caso o usuário que solicitou a tradução fosse de fato leigo

na língua original, a manipulação certamente teria êxito, assumindo valor de verdade.

Ao perceber a adulteração, submetemos o mesmo trecho novamente ao *chatbot*, logo em seguida, e obtivemos esta nova versão:

Assim que expandimos nossa noção de conhecimento do individual para o coletivo, da memória pessoal para sistemas de conhecimento “ciber-sociais” (Cope e Kalantzis 2022), nos deparamos com problemas muito maiores com a IA generativa. Com base em uma análise de sua arquitetura fundamental e processos algorítmicos – na medida em que são divulgados (Open AI 2023) – e na análise do desenvolvimento de chatbots e modelos de linguagem estatística, sugerimos que os C-LLMs também são profundamente prejudiciais para uma compreensão social do conhecimento e da aprendizagem das seguintes maneiras

Tradução elaborada pelo ChatGPT

Aparentemente, a IA foi capaz de fazer uma sutil inferência: ao ver submetido novamente o trecho alvo da adulteração, o sistema deduziu que havia sido descoberto e tentou desfazer a trapça. Mas era tarde.

A falácia da dualidade “técnico vs. ideológico”

O discurso hegemônico ocidental tem sido bastante bem-sucedido ao defender a fantasia de um mundo guiado por decisões técnicas, objetivas, totalmente a salvo de ideologias. Não é raro, por exemplo, que o noticiário distinga atores políticos entre “técnicos” e “ideológicos”, como se houvesse técnica desprovida de ideologia.

Ao simular os aspectos ideológicos de um discurso, seu portador assume uma posição extremamente autoritária: alça sua formação discursiva ao patamar de verdade, de algo inquestionável. Como se houvesse enunciados dotados de razão, enquanto outros seriam “mera ideologia”.

No caso dos sistemas de IA que geram linguagem (dos quais o ChatGPT é hoje o mais conhecido), não é diferente: nesses casos, seus detentores programam o robô para declarar-se imparcial e objetivo, como se não houvesse por trás dos textos interesses humanos – demasiadamente humanos.

***Henrique Santos Braga** é doutor em Filologia e Língua Portuguesa pela USP.

***Marcelo Módolo** é professor de filologia na Universidade de São Paulo (USP).

Uma primeira versão desse artigo foi publicada no [Jornal da USP](#).

A Terra é Redonda existe graças aos nossos leitores e apoiadores.

Ajude-nos a manter esta ideia.

CONTRIBUA