

Demócrito Filho: ChatGPT é um sistema de “alto risco”?

A empresa estadunidense OpenAI [1] fez no último dia 14 de março, o lançamento da mais nova versão



O programa é uma ferramenta algorítmica que imita a

linguagem natural, um tipo de inteligência artificial conversacional, ou seja, um *chatbot* que conversa e estabelece diálogos com o usuário. O que impressiona nele é sua capacidade descomunal de produzir textos, responder a perguntas sobre praticamente todos os assuntos e estabelecer conversações com raciocínio lógico [3]. É capaz de escrever textos de natureza diversa, como poemas, crônicas e até letras de música, em diversos estilos. Também pode desempenhar outras funções, como elaborar códigos de programa de computador, escrever roteiros de filmes, ensaios e muito mais [4].

É da família dos grandes modelos de linguagem (*large language models*), que são programas treinados em vastos conjuntos de dados textuais para gerar linguagem natural, de forma semelhante a um texto produzido por uma pessoa humana. Os modelos de linguagem são usados para compreender e responder a perguntas em línguas naturais [5], como o inglês, o português, o francês, o espanhol etc. Para gerar textos em língua natural, os modelos de linguagem são treinados para aprender a prever a próxima palavra ou frase com base no contexto anterior. Eles podem ser usados para várias tarefas, como tradução automática, geração de texto, resumo automático [6] e resposta a perguntas. Os modelos de linguagem mais recentes, como o GPT-3 [7] da OpenAI, são baseados em redes neurais [8] profundas e apresentam desempenho impressionante em várias tarefas de processamento de linguagem.

A versão apresentada na terça-feira, o GPT-4, é ainda mais impressionante, pois é um modelo multimodal de linguagem e, ao contrário da versão anterior (o GPT-3.5), aceita perguntas não apenas em forma textual mas também decifra e dá respostas quando provocado por imagens alimentadas pelo usuário [9]. O GPT-4 é capaz de analisar imagens e compreendê-las como se fossem entradas em texto. Por ser um modelo multimodal, é capaz de entender imagens, o que lhe permite capturar e mesmo explicar em detalhes o conteúdo, por exemplo, de uma fotografia. Isso ocorre porque o GPT-4 foi treinado sobre uma base de dados que inclui imagens e textos, diferentemente da versão anterior, que só era treinada em dados textuais.

A capacidade de processamento da nova versão também foi aumentada exponencialmente [10] e certas falhas no seu funcionamento foram consertadas. Apesar de suas respostas serem coerentes e gramaticalmente bem formuladas, usuários relataram muitos erros na versão original. Vários testes mostraram o sistema inventando pessoas, biografias e fatos, falhando na identificação de datas e sendo enganado por perguntas mais capciosas. A própria OpenAI alertava os usuários para a possibilidade de erros e defasagem nas informações. Com a nova versão (o GPT-4), a empresa espera ter reduzido substancialmente esses erros do sistema algorítmico, apelidados de "alucinações".

O novo modelo também ficou mais "inteligente". Segundo a OpenAI, o GPT-4 passou em vários testes de admissão e exames de universidades dos Estados Unidos. A empresa afirma que ele superou o desempenho do antecessor em testes como o *Uniform Bar Exam* (o equivalente estadunidense ao exame da Ordem dos Advogados do Brasil-OAB) e nas Olimpíadas de Biologia. Os conhecimentos em programação do GPT-4 também se tornaram melhores. Ele é capaz de recriar softwares com agilidade, seguindo à risca as instruções dadas pelo usuário [11].

Se por um lado a nova versão foi aperfeiçoada e suas funções ampliadas, o lançamento do GPT-4 foi marcado pela falta de transparência sobre aspectos do funcionamento do novo sistema de linguagem natural, sobretudo quanto ao conjunto de dados utilizados para o treinamento do algoritmo e o procedimento do treinamento. A OpenAI se limita a informar que os dados utilizados no treinamento procedem de uma "*variedade de fontes de dados licenciadas, criadas e disponíveis publicamente, que podem incluir informações pessoais disponíveis publicamente*" [12]. Apesar de o lançamento da versão GPT-4 ter sido acompanhado da divulgação de um relatório de 98 páginas, esse documento não revela muito sobre o treinamento de dados do novo sistema de inteligência artificial generativa. A OpenAI justifica a ausência de informações em razão do cenário competitivo e por motivos de segurança, segundo o documento [13].

A falta de transparência sobre aspectos do funcionamento da versão GPT-4 repercutiu negativamente na comunidade acadêmica [14] e também serve de fator para apressar a regulamentação da IA. A opção por não divulgar praticamente nada sobre como o *ChatGPT* (na versão GPT-4) é treinado aumenta as apreensões quanto aos impactos da inteligência artificial generativa (*generative artificial intelligence*) sobre os direitos e segurança das pessoas.

A chegada do *ChatGPT* representa um salto no desenvolvimento da inteligência artificial, com implicações sociais profundas. Trata-se de um tipo de tecnologia não apenas revolucionária, mas que está se disseminando muito rapidamente. A tecnologia do *ChatGPT* já é empregada em grande número de diferentes serviços e aplicações [15]. No início de março de 2023, a OpenAI lançou sua API [16] para desenvolvedores poderem adicionar a tecnologia do *ChatGPT* em seus próprios serviços [17].

Os riscos associados com a proliferação de modelos de inteligência artificial de uso geral (*general purpose AI*) despertam a necessidade de haver um maior controle sobre sua utilização. A questão é como se estabelecer esse controle.

Desde abril de 2021, a União Europeia apresentou sua proposta para regulação das tecnologias de inteligência artificial. A proposta, que recebeu o nome de Artificial Intelligence Act (ou abreviadamente AI Act) [18], atualmente tramita no Parlamento Europeu. A abordagem regulatória tem como premissa principal a hierarquização dos riscos oferecidos por sistemas e tecnologias que usam IA. Segundo essa visão regulatória baseada nos riscos (*risk-based regulatory approach*), as restrições e exigências aumentam à medida que maiores sejam os riscos que os sistemas de IA possam oferecer a direitos e garantias fundamentais dos indivíduos. Em relação aos sistemas de alto risco (*high-risk AI systems*), o AI Act impõe severas restrições ao desenvolvimento, adoção e uso. Sistemas incluídos nessa categoria sujeitam o provedor a obrigações de documentação, rastreabilidade, supervisão humana e outras imposições indispensáveis para mitigar consequências danosas aos usuários [19].

A dificuldade é como enquadrar os sistemas de inteligência artificial generativa, a exemplo do *ChatGPT*, no esquema de categorização traçado no AI Act. O programa pode ser considerado um sistema de "alto risco", de maneira a que a empresa controladora possa ser submetida a maior supervisão e restrições regulamentares?

Para alguns eurodeputados, como o liberal romeno Dragoș Tudorache e o social-democrata italiano Brando Benifei, relatores do AI Act no Parlamento Europeu, a tecnologia que faz o *ChatGPT* funcionar tem o condão de enquadrá-lo como um sistema de IA que oferece elevado risco aos direitos fundamentais das pessoas [20]. No dia 14 de março, divulgaram uma minuta de emenda para impor obrigações aos desenvolvedores de sistemas de IA de uso geral, enquadrando na prática essa tecnologia na categoria de sistemas de "alto risco" [21]. Outros acreditam que classificar a inteligência artificial generativa como de "alto risco" pode embotar o desenvolvimento dessa tecnologia.

Em relação à inteligência artificial de uso geral (General Purpose AI), parece ser adequada uma regulamentação mais rigorosa. Especificamente no que diz respeito aos modelos de linguagem generativa, os riscos são de que aumentem a desinformação e o processo de disseminação de notícias falsas (fake news) que tem ameaçado governos democráticos. Os grandes modelos algorítmicos de linguagem natural (*large language models*) tornarão mais fácil a produção de milhares de notícias falsas, com versões diversas sobre um mesmo fato. Por meio da criação de perfis falsos em plataformas digitais, será mais fácil influenciar pessoas a tomar decisões equivocadas.

O *ChatGPT* e os grandes modelos de linguagem em geral, já que são treinados e coletam grandes quantidades de informações disponíveis livremente nas redes telemáticas, também levantam preocupações com aspectos ligados à privacidade e à proteção de dados pessoais.

O sistema coleta imensas quantidades de dados disponibilizadas na internet, podendo processar dados sensíveis das pessoas que se encontrem hospedados em outros serviços e plataformas digitais. Como não se sabem exatamente os parâmetros do seu funcionamento, não é descartado que o *ChatGPT* seja treinado não só com base em dados abertos, mas também sobre dados que lhe são compartilhados por outras plataformas e serviços digitais. Ademais, como se trata de um sistema conversacional, que se envolve em diálogo utilizando linguagem natural, isso pode levar a que o usuário forneça informações pessoais de cunho íntimo. Em face dessa característica de seu funcionamento, o usuário muitas vezes pode imaginar que está travando contato com uma pessoa humana, e não com um sistema de inteligência artificial. Engajando-se em um diálogo com o *ChatGPT*, o usuário pode revelar interesses, credos, preferências sexuais, ideológicas e dados relativos à saúde. O *chatbot* não só armazenará todas as mensagens recebidas do usuário, como, por ser um sistema de inteligência artificial altamente sofisticado, realizará cruzamento desses blocos de dados e extrairá inferências que lhe permitirão traçar um perfil completo da personalidade do usuário.

Mencionem-se ainda os acidentes de segurança que podem ocorrer, expondo dados pessoais de usuários. Na terça-feira, 21/3/2023, um *bug* no sistema do *ChatGPT* acabou revelando conversas de terceiros para os usuários do *chatbot* [22]. Além disso, como o programa é capaz de escrever códigos de programa para computador, teme-se que possa ser utilizado para elaboração de códigos maliciosos (*malwares*) e como ferramenta para ataques cibernéticos [23].

Como se observa, as implicações desses novos sistemas de inteligência artificial generativa são muitas, sobre diversas áreas e aspectos da vida humana, com o potencial de afetar direitos fundamentais, o que reforça a necessidade de regulamentação mais estrita do desenvolvimento e funcionamento dessas tecnologias. A ampla gama de aplicativos nos quais os sistemas de uso geral podem ser incorporados significa que qualquer falha pode ter efeitos abrangentes em muitos setores — uma única falha pode afetar centenas de aplicativos de IA integrados ao modelo de uso geral.

A formulação correta de uma política regulatória para os sistemas de IA de uso geral é essencial. Ainda que não se considerem os sistemas de inteligência artificial generativa, a exemplo do modelo sobre o qual funciona o *ChatGPT*, como uma tecnologia essencialmente periculosa, que traz "alto risco" para a segurança e saúde das pessoas, uma regulamentação específica é indispensável, como forma de atenuar os possíveis danos aos direitos fundamentais dos indivíduos.

[1] A OpenAI é uma companhia sediada na Califórnia especializada em soluções de inteligência artificial generativa. O site da OpenAI: <https://openai.com/>

[2] Do inglês *Chat Generative Pre-Trained Transformer*, que traduzido para o português seria algo como Transformador Pré-treinado Gerador de Diálogos.

[3] Para saber mais sobre as características do *ChatGPT* e seu impacto no mercado de tecnologia, recomendamos a leitura de nosso artigo inicial sobre o assunto, publicado na revista *Jus Navigandi* em 11.03.23, sob o título "O fenômeno do ChatGPT desperta a necessidade da regulamentação da Inteligência Artificial", acessível em: <https://jus.com.br/artigos/102919>

[4] Para fazer um teste no ChatGPT, acesse: <https://chat.openai.com/auth/login>

[5] *Língua natural* é qualquer linguagem desenvolvida naturalmente pelo ser humano, como o português, o francês, o inglês etc.

[6] Os modelos de linguagem podem ser usados para resumir automaticamente documentos longos ou artigos, tornando-os mais acessíveis e fáceis de ler.

[7] *Generative Pre-Training Transformer 3* (GPT-3) – que traduzido para o português seria algo como "Transformador Generativo Pré-Treinado 3". A versão completa do GPT-3 tem capacidade para 175 bilhões de parâmetros de aprendizado de máquina.

[8] Em ciência da computação, **redes neurais artificiais (RNAs)** são modelos computacionais inspirados pelo sistema nervoso central de um animal (em particular o cérebro) que são capazes de realizar o aprendizado de máquina (aprendizagem a partir dos dados) bem como o reconhecimento de padrões. (Cf. Wikipedia, https://pt.wikipedia.org/wiki/Rede_neural_artificial).

[9] Conforme notícia publicada no site *Olhar Digital*, em 16.03.23, acessível em: <https://olhardigital.com.br/2023/03/16/pro/novo-chatgpt-cria-site-do-zero-descreve-imagem-e-entende-humor/>

[10] O novo modelo consegue entregar respostas de até 25 mil palavras, bem mais do que as 8 mil do GPT-3.5

[11] Conforme notícia publicada no site Canal Tech, em 21.03.23, acessível em: <https://canaltech.com.br/inteligencia-artificial/coisas-que-o-chatgpt-4-faz-melhor-que-o-chatgpt-35/>

[12] No texto em inglês divulgado pela OpenAI, lê-se: "*training data came from a variety of licensed, created, and publicly available data sources, which may include publicly available personal information*".

[13] Diz o seguinte trecho do documento: *"Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar"*. O relatório completo pode ser acessado em: <https://cdn.openai.com/papers/gpt-4.pdf>

[14] Ver notícia publicada no site *FastCompany*, em 16.03.23, acessível em: <https://www.fastcompany.com/90866190/critics-denounce-a-lack-of-transparency-around-gpt-4s-tech>

[15] Essa é, aliás, a característica marcante dos sistemas de inteligência artificial de uso geral: a possibilidade de serem aproveitados em sistemas de IA mais especializados. Os modelos baseados em inteligência artificial generativa permitem larga variedade de uso. Podem ser utilizados para diferentes tarefas, em diversos campos, geralmente sem necessidade de modificações substanciais na sua programação. Por isso esses sistemas são às vezes chamados de "modelos de fundação" (*foundation models*), devido ao seu uso generalizado como modelos pré-treinados para outros sistemas de IA mais especializados.

[16] API significa *Application Programming Interface* (Interface de Programação de Aplicação). A interface pode ser entendida como uma ponte ou elo de ligação entre duas aplicações ou sistemas informáticos diferentes.

[17] Ver notícia publicada em 03.03.23, acessível em: <https://mundoconectado.com.br/noticias/v/32007/openai-anuncia-api-do-chatgpt-diversas-empresas-ja-estao-testando>

[18] Para saber mais sobre o AI Act, sugerimos a leitura do nosso artigo publicado na Revista Jus Navigandi, em 27.05.21, sob o título: "A proposta regulatória da União Europeia para a inteligência artificial (1ª parte): a hierarquização dos riscos". Disponível em: <https://jus.com.br/artigos/90816>

[19] Sistemas incluídos na categoria de "alto risco" (*high-risk*) se sujeitam a requisitos de **conformidade e avaliação prévia de impacto**, antes de serem colocados no mercado. Para saber mais sobre os sistemas de IA de "alto risco", recomendamos a leitura do nosso artigo publicado na revista *Jus Navigandi*, em 20.06.21, sob o título: "A proposta regulatória da União Europeia para a inteligência artificial (3ª parte): sistemas de alto risco". Disponível em: <https://jus.com.br/artigos/91363/a-proposta-regulatoria-da-uniao-europeia-para-a-inteligencia-artificial-3-parte-sistemas-de-alto-risco>

[20] Ver, a propósito, notícia publicada no site português Politico, em 03.03.23, acessível em: <https://www.politico.eu/article/eu-plan-regulate-chatgpt-openai-artificial-intelligence-act/#:~:text=Europe's%20original%20plan%20to%20bring,technology's%20new%2C%20shiny%20chatbot>

.

[21] Ver notícia publicada no site Euractiv, em 14.03.23, acessível em: <https://www.euractiv.com/section/artificial-intelligence/news/leading-eu-lawmakers-propose-obligations-for-general-purpose-ai/>

[22] Um porta-voz da *OpenAI* confirmou o defeito e acrescentou que o *bug* não expunha transcrições completas das conversas, mas apenas os títulos. Ver notícia publicada no site *Olhar Digital*, em 21.03.23, acessível em: <https://olhardigital.com.br/2023/03/21/seguranca/chatgpt-bug-expoe-historico-de-chat-de-usuarios/>

[23] É o próprio Sam Altman, o CEO da OpenAI, quem advertiu para essa possibilidade, segundo notícia do *The Guardian*, publicada em 17.03.23, acessível em: <https://www.theguardian.com/technology/2023/mar/17/openai-sam-altman-artificial-intelligence-warning-gpt4>

Date Created

03/04/2023