RÖHE, Anderson; SANTAELLA, Lucia. Confusões e dilemas da antropomorfização das inteligências artificiais. *TECCOGS* – Revista Digital de Tecnologias Cognitivas, n. 28, 2023, p. 67-75.

Recebido em: 15 jan. 2024
Aprovado em: 20 fev. 2024

dx.doi.org/
10.23925/1984-3585.2023i28p67-75

Confusões e dilemas da antropomorfização das inteligências artificiais

Anderson Röhe[1]

Lucia Santaella[2]

Resumo: Uma das principais razões de confusões, que rondam a compreensão daquilo que a Inteligência Artificial Generativa (IAGs) é ou não capaz de realizar, consiste em tomá-las à imagem e semelhança dos humanos, como se fossem, inclusive capazes de sentimentos ou como se elas próprias fossem racistas. Diante disso, este artigo visa debater quais os limites das IAGs e qual o papel que os humanos desempenham no uso que fazem delas. O resultado esperado é desmistificar a eficácia das IAs para a detecção de emoções ou para comportamentos éticos, sobretudo em relação aos cuidados a serem tomados em ambientes vulneráveis como a sala de aula.

Palavras-chave: Inteligência Artificial Generativa; emoções; racismo; uso humano

1 Advogado na Comissão Especial de Privacidade, Proteção de Dados e Inteligência Artificial da OAB-SP. Fellow no Think Tank ABES (GT de Inteligência Artificial). Pós-graduado em Direito Digital pela Universidade do Estado do Rio de Janeiro (UERJ) e Instituto de Tecnologia e Sociedade de Rio (ITS Rio). Mestre em Políticas Internacionais pela PUC-Rio. Doutorando em Tecnologias da Inteligência e Design Digital da PUC-SP. Orcid: https://orcid.org/0000-0002-3104-6365.

2 É pesquisadora 1A do CNPq, professora titular na pós-graduação em Comunicação e Semiótica e em Tecnologias da Inteligência e Design Digital (PUC-SP). Doutora em Teoria Literária pela PUC-SP e Livre-docente em Ciências da Comunicação pela USP. Publicou 56 livros e organizou 33, além da publicação de quase 500 artigos no Brasil e no exterior. Recebeu os prêmios Jabuti (2002, 2009, 2011, 2014), o prêmio Sergio Motta (2005) e o prêmio Luiz Beltrão (2010). Orcid: https://orcid.org/0000-0002-0681-6073.

Confusions and dilemmas with the anthropomorphization of artificial intelligences

Abstract: One of the main reasons for confusion, which surrounds the understanding of what Generative Artificial Intelligence (IAGs) is or is not capable of achieving, consists of taking them in the image and likeness of humans, as if they were capable of feelings or as if they themselves were racist. This article aims to debate the limits of IAGs and the role that humans play in their use of them. The expected result is to demystify the effectiveness of AIs for detecting emotions or ethical behaviors, especially in relation to the precautions to be taken in vulnerable environments such as the classroom.

Keywords: Generative Artificial Intelligence; emotions; racism; human use

Introdução

A dificuldade, que o ser humano encontra em reconhecer que há outras formas de inteligência não humanas, constitui-se em uma das possíveis explicações para as tendências de antropomorfizar as Inteligências Artificiais (IAs) até o ponto de emprestar-lhes ou exigir-lhes uma consciência moral, entre outros presupostos, como supor que os sistemas de IA sentem, mentem ou sejam racistas. Essas tendências acentuaram-se agudamente, depois do advento da IA generativa (IAG) na capacidade que ela revela de responder a demandas humanas, por verbo, imagem ou por ambos, com prontidão e até gentileza.

Há muita contradição aí: não sendo capaz de reconhecer uma forma de inteligência distinta da sua, mas não podendo deixar de perceber que os resultados das operações da IA apresentam sinais de inteligência, o ser humano se livra do problema seguindo por um entre dois caminhos: ou declara que a IAG é burra, um mero papagaio estocástico, que não entende nada do que escreve e produz ou, então, esperam-se dela reações cognitivas tipicamente humanas. Mesmo que seja verdade que a IAG não entende o significado daquilo que está escrito ou das imagens que produz, pois não tem consciência nem de si, nem consciência geral, não é isso que importa, pois ela responde e age como se tivesse. Eis a questão.

Essas contradições têm atravancado muito a compreensão da natureza e das consequências pragmáticas, ou seja, dos efeitos sensíveis que a IA tem provocado e deverá provocar, prejudicando o encontro de ações mais éticas e eficazes de lidar com ela. Por isso, no que se segue colocaremos em discussão os três fatores-chave das distorções que a antropomorfização da IA tem provocado.

As IAGs sentem?

Inicialmente definia-se IA da perspectiva estritamente racional e só mais recentemente a pesquisa expandiu-se para o campo dos sentimentos e das emoções. Até então a questão do sentir não fazia parte do repertório, pois esbarra no "duro problema da consciência", no sentido de ser muito complexo explicar o que é consciência (Chalmers, 2018, p. 6).

Para Tenenbaum (2018, p. 482-483), o termo "consciência" possui significados diferentes para campos do conhecimento diferentes, não havendo unanimidade, por exemplo, entre cientistas cognitivos, filósofos e neurocientistas. Tudo depende de como a conceito de consciência é to-

mado, segundo Bostron (2018, p. 11), não havendo ainda clareza do que é realmente necessário e quais as condições para o seu desenvolvimento. Outros vão mais longe, como Marcus (2018, p. 127), ao afirmar que consciência nem sequer é um pré-requisito para o desenvolvimento de uma IA. Já Koller (2018, p. 395) dispensa consciência até mesmo para se alcançar uma IA Geral ou superinteligência (um estado hipotético em que a IA suplantaria a humana). Logo, se entre pessoas já é difícil encontrar explicações para definir consciência, quanto mais em não humanos (Santaella, 2022, p. 56-66).

De todo modo, é um campo de pesquisa que se acelera a cada dia. Se nos anos de 1990 os computadores começavam a adquirir a capacidade de expressar e reconhecer afetos para, em seguida, vir a adquirir a capacidade de ter emoções (Picard, 1995), já na década seguinte, Rosalind Picard, em Computação Afetiva (2000), afirmava que as emoções desempenham papel mais relevante do que se imaginava; não só na tomada de decisões, mas também na percepção e no aprendizado, a ponto de influenciar os próprios mecanismos do pensamento racional. Portanto, deveria ser dado também aos computadores a capacidade de reconhecer, compreender e até mesmo de ter e expressar emoções com o objetivo de interagirem melhor conosco e aprimorarem a própria inteligência humana (Picard,1995, p. 1).

A computação afetiva despontava, então, como nova área de pesquisa com resultados promissores, sobretudo no reconhecimento das expressões faciais e síntese da inflexão da voz (*ibid.*, p. 24), uma vez que as máquinas são melhores que os humanos para reconhecer padrões e micro sutilezas que não são tão rapidamente perceptíveis por humanos, assim como são hábeis em reconhecer quando essas inflexões se relacionam com sentimentos, tais como estresse, alegria ou raiva (Somers, 2019). As máquinas são tão eficazes nessas tarefas que isto seria apenas o início da descoberta de um estado afetivo oculto, não sendo nem mesmo necessário que esses estados sejam universais em sua expressão para que um computador os possa reconhecer.

Em *As Microexpressões Faciais: uma investigação semiótica do poder de comunicação da face humana*, Alessandra S. Cejkinski (2022, p. 1) percebe o reconhecimento das expressões faciais como etapa do próprio processo de reconhecimento das emoções. A autora o faz mostrando a divergência entre o cientista Charles Darwin, para quem as emoções são inatas e universais, e a antropóloga Margaret Mead para a qual as emoções podem sofrer interferências da cultura, do ambiente e até mesmo dos processos de aprendizagem.

A "IA emocional" ou computação afetiva foi, então, dividida em três grandes áreas de estudo por Eva Hudlicka (2008), especialista em computação afetiva aplicada aos *games*: (a) o reconhecimento de emoções; (b) modelos computacionais de emoções e (c) expressão de emoções em agentes artificiais e robôs. A primeira, de reconhecimento de emoções, é a que busca traduzir emoções humanas através de uma interação mais natural entre humanos e máquinas para depois transformá-las em dados computacionais (ver Trevisan; Braga, 2022). Quanto ao item (b), relativo aos modelos computacionais de emoções, eles são introjetados para que o resultado soe como expressão de emoções, ou seja, o item (c).

A IAG mente?

As IAGs têm sido acusadas de apresentarem condutas de tipos similares a mentir. Diz-se que elas iludem, desinformam e até mesmo trapaceiam para conseguir completar uma tarefa, pois não há nada mais próximo do comportamento humano do que mentir (Martins, 2023). São atribuições que geram, então, um temor ético e de alto risco (OpenAI, 2023, p. 94).

Conforme foi amplamente discutido por Santaella (2021, p. 35-54), mentir apresenta uma semântica muito peculiar. A mentira é distinta do erro, do lapso, da hipocrisia e do cinismo. Enquanto o erro e o lapso não são deliberados, na hipocrisia e no cinismo há sempre uma certa dose de deliberação, mesmo que vaga. Mas só há mentira quando houver uma intencionalidade para mentir, cujo alvo é enganar, ou seja, fazer crer em algo que trai a realidade dos fatos.

Aqui é importante diferenciar o propósito da intencionalidade. Propósito é uma ação dirigida para um fim, agenciada pelos meios que lhe são necessários para chegar à sua finalidade. Disso são capazes animais superiores ou inferiores e, certamente, a IA, alimentada que é por estatísticas sofisticadas. Intencionalidade, entretanto, refere-se estritamente à versão psicológica do propósito. Portanto, simples assim: a IAG tem propósito, mas não tem intencionalidade, pois a intencionalidade depende da existência da consciência e autoconsciência, algo que a IA está longe de ter. Por isso, na escolha dos meios para atingir seus fins, o ser humano revela-se ético ou trapaceiro em todas as suas variedades fortes ou fracas. De resto, muitas vezes, a falta de ética pode provir de uma ingenuidade movida a ignorância ou mesmo por narcisismo que oblitera suas próprias fraquezas, ou seja, por falta de autoconsciência dos limites de si e do

contexto. Esses são comportamentos complicados que a matemática de é feita a IA não lhe permite, portanto de que a IA não poderia ser capaz. Entretanto, os bancos de dados, o treinamento e a estatística dos algoritmos podem não funcionar muito bem e a IAG produzir um resultado que fica perto, mas não chega a ser algo que alcance a psicose humana, ou seja, o ChatGPT alucina. Vejamos.

O grande diferencial do ChatGPT está na democratização de seu chatbot, fazendo com que seu sistema conversacional gere uma imediata conexão simuladamente afetiva entre usuário e máquina (Lerobitcast, 2023). Mas isso não deveria enganar o usuário, pois não dota o ChatGPT de intencionalidade e consciência humana. Consequentemente, afetividade temos que buscar em outras searas estritamente humanas, mesmo que não sejamos bem-sucedidos em nossas buscas. Tanto é que nem sempre somos bem-sucedidos que já existem sistemas de IA (Replika) para relacionamentos amorosos, basta inscrever-se, pagar uma taxa e saborear essa experiência.

Mesmo que saiba fingir, digamos assim, a IA está equipada para gerar respostas com muita prontidão, ainda que incompletas, imprecisas ou inteiramente falsas, mas evidentemente sem ter a menor noção do que é a verdade, aliás um tema que tem ocupado os filósofos há séculos. Conclusão, o propósito da IA e não sua intencionalidade é o de gerar uma resposta que, além de padronizada, muitas vezes pouco especializada, pode também passar longe da correção. Isso costuma ser chamado de alucinações (Malar, 2023), uma terminologia que deve ser vista, quando se trata da IA, dentro do seu próprio contexto, pois no caso de humanos, alucinação é caso de psicose, facilmente detectável pelos especialistas.

Todo discurso dispõe de profundidade, ou seja, sua capacidade conotativa, e de extensão, seu potencial de aplicação a situações e contextos. Ora, a IAG pode até ser capaz de ensaiar, com algum êxito, linguagens poéticas e imagens metafóricas. Mas é desprovida de senso comum, da sabedoria baseada na experiência vivida, do conhecimento tácito e outras tantas faculdades cognitivas que concedem aos humanos seus encantos e desencantos. Portanto, é justamente a carência de todas as faculdades, de certa forma parentes do senso comum, e especialmente a ausência de discernimento em saber se a resposta gerada está certa ou incorreta que esses sistemas alucinam sem nenhum apreço pelas suas fontes de referência.

Em suma: a IA não está programada para dizer a verdade, e sim "gerar conteúdos coerentes para os humanos a partir da base de dados

usada no seu treinamento" (Malar, 2023, p. 2). Portanto, dizer que a IAG não é responsável pelo que apresenta não é a mesma coisa que negar que ela possa ser usada pelos humanos para mentir, enganar, tripudiar e todos os outros tipos de malidicências que hoje encontram nas redes da web um fluxo promissor para disseminar. A IAG, portanto, não dispõe da intencionalidade para mentir, mas pode ser usada para isso. Isso também não significa negar que os bancos de dados podem estar viciados, contaminando o que se seguirá. Também não significa negar que os desenvolvedores podem ser negligentes em seus cuidados com a ética. Por fim, não significa negar que os resultados que a IA passa à frente podem servir para atender interesses escusos das mais diversas espécies.

As IAs são racistas?

Já na metade da década de 1990, Bayat Friedman e Helen Nissenbaum (*apud* Kaufman; Junquilho; Reis, 2023, p. 44), pioneiras nos estudos sobre sistemas de computação tendenciosos, alertavam para suas ameaças à sociedade e seus potenciais impactos para os direitos humanos e fundamentais. Para tanto, identificaram três categorias de viés desses sistemas: a preexistente, cujas origens estão nas "instituições, em práticas e atitudes sociais" (Kaufman; Junquilho; Reis, 2023, p. 44), mas não necessariamente na tecnologia *per se*; a técnica (que surge de restrições operacionais dos próprios sistemas); e a emergente (a partir de determinadas situações e contextos de uso da nova tecnologia).

Portanto, um dos problemas mais sérios da IA encontra-se nos vieses, considerando-se que muitos deles precedem as IAs e podem estar presentes nos dados. Ora, dados não são neutros, e, portanto, não confiáveis. Vieses já se encontram dentro da base de dados que é utilizada para o treinamento dos algoritmos, que, em algumas situações ou contextos específicos, se revelam tendenciosos, já que alimentados predominantemente por um público masculino, de pele clara (Kaufman; Junquilho; Reis, 2023, p. 44-45) sem a devida representatividade e diversidade, seja socioeconômica, racial, étnica, regional ou de gênero. Uma "fórmula" que já se mostrou problemática consiste no algoritmo que foi considerado racista por rotular um determinado grupo de fotos de pessoas negras como sendo a de gorilas. E, quando "corrigido", o sistema passou a não identificar nem mesmo os gorilas (*ibid.*, p. 45). Assim, se os dados não forem higienizados, ao passar pelos algoritmos e chegar aos resultados, a IA os intensifica, como bem exemplifica o racismo estrutural no Brasil.

Os problemas se acentuaram com as IAGs que automatizam a produção de textos e imagens cada vez mais realistas produzindo sérias dissociações entre fantasia e realidade. Além disso, como a maioria dos dados é produzido fora do Brasil, baseados em outras culturas, tais como a dos EUA e da União Europeia, isto é, por regiões mais ricas e poderosas do que a nossa, países representantes do Sul Global (com exceção da China, pois é um caso à parte) têm poucas chances, recursos humanos e infraestrutura computacional para tentar competir com aquelas superpotências (Santos e Soares, 2023). E, na falta de um *framework*/modelo autóctone, ao tentar replicar tais modelos, acabam não só importando problemas estrangeiros, mas também reforçando (pré)conceitos e estereótipos que já existem por aqui.

Referências

BOSTROM, Nick. Entrevista concedida a Martin Ford. *In:* FORD, Martin (ed.). *Architects of intelligence*: The truth about AI from the people building it. Birmingham: Packt, 2018, p. 97-116.

CEJKINSKI, Alessandra Sciammarella. *As microexpressões faciais*: uma investigação semiótica do poder de comunicação da face humana. Dissertação de Mestrado em Comunicação e Semiótica. São Paulo, PUC-SP, 2022.

CHALMERS, David J. The meta-problem of consciousness. *Journal of Consciousness Studies*, v. 25, n. 9–10, p. 6–61, 2018. Disponível em: https://philpapers.org/archive/CHATMO-32.pdf. Acesso em: 10 jan. 2024.

HUDLICKA, Eva. Affective computing for game design. *In: Proceedings of the 4th North American Conference on Intelligent Games and Simulation*. Montreal: McGill University, p. 5-12, 2008.

KAUFMAN, Dora; JUNQUILHO, Tainá; REIS, Priscila. Externalidades negativas da inteligência artificial: conflitos entre limites da técnica e dos direitos humanos. *Revista de Direitos e Garantias Fundamentais,* Vitória, v. 24, n. 3, set./dez. 2023 p. 44. Disponível em: https://sisbib.emnuvens.com.br/direitosegarantias/article/view/2198. Acesso em: 21 dez., 2023.

KOLLER, Daphne. Entrevista concedida a Martin Ford. *In:* FORD, Martin (ed.). *Architects of intelligence*. The truth about AI from the people building it. Birmingham: Packt, 2018, p. 387-403.

LEROBITCAST. Antropomorfização da IA. *Medium*, 7 maio 2023. Disponível em: https://medium.com/@lerobitcast/antropomorfiza%C3%A7%C3%A30-da-ia-caccc5d13a80. Acesso em: 2 jul. 2023.

MALAR, João Pedro. O ChatGPT mente? Entenda as "alucinações" de inteligências artificiais. *Exame*, 19 maio, 2023. Disponível em: https://exame-com.cdn.ampproject.org/c/s/exame.com/future-of-money/o-chatgpt-mente-entenda-as-alucinacoes-de-inteligencias-artificiais/amp/. Acesso em: 3 jul. 2023.

MARCUS, Gary. Entrevista concedida a Martin Ford. *In:* FORD, Martin (ed.). *Architects of intelligence*. The truth about AI from the people building it. Birmingham: Packt, 2018, p. 305-330.

MARTINS, Flávia. Chat GPT-4: inteligência artificial mente para completar tarefa e gera preocupação. *CNN Brasil*, 24 mar. 2023. Disponível em: https://www.cnnbrasil.com.br/tecnologia/chat-gpt-4-inteligencia-artificial-mente-para-completar-tarefa-e-gera-preocupacao/. Acesso em: 22 dez. 2023.

OPENAI. *GPT-4 Technical Report*. Disponível em: https://cdn.openai.com/papers/gpt-4.pdf. Acesso em: 26 dez. 2023.

PICARD, Rosalind Wright. *Affective Computing*. MIT Media Laboratory Perceptual Computing Section Technical Report No. 321, 1995.

PICARD, Rosalind Wright. *Affective computing*. Cambridge, MA: MIT Press, 2000.

SANTAELLA, Lucia. *De onde vem o poder da mentira?* São Paulo: Estação das Letras e Cores, 2021.

SANTAELLA, Lucia. *Neo-humano*. A sétima revolução cognitiva do Sapiens. São Paulo: Paulus, 2022.

SANTOS, Nina; SOARES, Matheus. Potencial do Brasil para IA está na qualidade dos dados oficiais. *Desinformante*, 26 dez., 2023. Disponível em: https://desinformante.com.br/brasil-ia-dados/. Acesso em: 27 dez., 2023.

SOMERS, Meredith. Emotion AI, explained. *MIT Edu*, 2019. Disponível em: https://mitsloan.mit.edu/ideas-made-to-matter/emotion-ai-explained. Acesso em: 29 set., 2023.

TENENBAUM, Joshua. Entrevista concedida a Martin Ford. *In:* FORD, Martin (ed.). *Architects of intelligence*. The truth about AI from the people building it. Birmingham: Packt, 2018, p. 473-491a.

TREVISAN, Daniel; BRAGA, Alexandre. Inteligência artificial nos games. *TECCOGS – Revista Digital de Tecnologias Cognitivas*, n. 26, jul./dez., 2022, p. 89–101.