15ª Conferência Lusófona de Ciência Aberta (ConfOA)
Ciência Aberta e outras expressões de conhecimento aberto
Modalidade: Pecha Kucha

Ciência aberta e inteligência artificial: desafios éticos e transparência em modelos generativos

Open science and artificial intelligence: ethical challenges and transparency in generative models

Moisés Rockembach
Universidade de Coimbra (UC)
Coimbra, Portugal
Lattes: 1304688580274983
Orcid: 0000-0001-9057-0602
moises.rockembach@gmail.com

RESUMO

O artigo aborda os desafios éticos e de transparência associados ao uso de Inteligência Artificial (IA) em alinhamento com os princípios da Ciência Aberta, com um foco especial nos modelos generativos, como o ChatGPT e Mistral. Utilizando uma metodologia que combina pesquisa documental e bibliográfica, o estudo examina tanto modelos de IA abertos quanto fechados, analisando como essas tecnologias podem distorcer a integridade científica. O objetivo central é fomentar um debate sobre as implicações dessas distorções e propor mecanismos para mitigar tais efeitos, reforçando a necessidade de modelos transparentes que assegurem a reprodutibilidade científica. Conclui-se que, quando aplicados com rigor ético, os modelos generativos têm o potencial de auxiliar significativamente a Ciência Aberta, promovendo um novo nível de colaboração e inovação no campo científico.

Palavras-chave: ciência aberta; inteligência artificial; ética; transparência.

INTRODUÇÃO

É notável que tecnologias como a Inteligência Artificial (IA) vem transformando as fronteiras no contexto de investigação científica teórica e aplicada nos últimos anos. Embora a IA tenha sido discutida em ambientes acadêmicos desde meados do século XX, seus desenvolvimentos mais recentes, particularmente os modelos de linguagem grande (LLM, *Large Language Model*) e modelos generativos como o *Generative Pretrained Transformer* (GPT), tem desencadeando mudanças significativas em diversas áreas e no trabalho científico não tem sido diferente. O fenómeno do uso de IA na ciência também traz problemas, nomeadamente, a produção de informação falsa ou "alucinações", viés de dados, dificuldade na reprodutibilidade, ou o uso indevido na produção textual, que acabam por refletir no nível de qualidade dos outputs científicos. Posto isto, problematiza aqui quais são os atuais desafios éticos e de transparência no uso da IA em consonância com a Ciência Aberta? O objetivo é propor um debate sobre como tais distorções impactam a integridade da ciência e sugerindo mecanismos para mitigar esses efeitos no contexto da Ciência Aberta.

PROCEDIMENTOS METODOLÓGICOS

Na pesquisa documental, o foco esteve na observação de alguns dos principais modelos generativos de IA, identificando os avanços e as implicações éticas dessas ferramentas. Esse procedimento permitiu identificar modelos como o ChatGPT e Claude (modelos proprietários) e Mistral (modelo aberto), destacando como utilizam os dados de treinamento e como tratam pesos e parâmetros nos respetivos modelos de IA. Por sua vez, a pesquisa bibliográfica foi conduzida através da base de dados Dimensions.ai, com a identificação dos artigos que abordam a relação entre IA e Ciência Aberta. Para a busca, utilizaram-se os termos em inglês "Artificial Intelligence" e "Open Science" em títulos ou resumos, restringindo-se a artigos publicados nos anos de 2022 a 2024, até 15 de abril de 2024. Essa delimitação temporal foi escolhida devido ao significativo aumento no uso de modelos generativos após o lançamento do ChatGPT em novembro de 2022, com 26 artigos encontrados em 2022 e um aumento para 44 em 2023. Em 2024, foram registrados 20 artigos até metade de abril, sugerindo um continou crescimento.

Vantagens e desafios da inteligência artificial na ciência aberta

Em 2016, muito antes do *boom* do uso de Inteligência Artificial generativa, Castelvecchi (2016) já destacava a complexidade dos sistemas de aprendizado de máquina e os desafios enfrentados para tornar a IA transparente, enfatizando a necessidade de complementar o aprendizado profundo com técnicas mais interpretáveis para avançar no conhecimento científico sem sacrificar a transparência. Lin (2023) destaca que a IA não apenas exibe habilidades que imitam a inteligência humana, como resolução de problemas e raciocínio,

mas também são notavelmente versáteis, capazes de gerar texto coerente em uma vasta gama de áreas do conhecimento e colaborar eficazmente em tarefas acadêmicas. Contudo, também apresentam limitações significativas, como a falta de verdadeira compreensão e a possibilidade de gerar informações incorretas ou alucinações. Além disso, Lin (2023) ressalta as implicações éticas, de igualdade e educacionais, sugerindo que a IA pode tanto ajudar a nivelar desigualdades quanto ampliá-las, e enfatiza a importância de integrar essas tecnologias ao currículo educacional para fomentar uma análise crítica e habilidades analíticas nos alunos. Desta forma, a formação e atuação de profissionais especializados, como os eticistas digitais (Rockembach; Geerts, 2024), torna-se fundamental nestes novos ambientes tecnológicos, sobretudo com a adoção maciça da IA na sociedade.

O fenômeno das "alucinações artificiais" é discutido por Alkaissi e McFarlane (2023), onde um experimento demonstrou que, enquanto ChatGPT pode produzir textos coerentes a partir de notas dispersas, também é suscetível a produzir informações enganosas ou incorretas. Baronchelli (2024) argumenta sobre a urgência na criação de novas normas sociais em resposta à crescente integração da IA em nossas vidas (como o *AI Act*), salientando a rápida evolução da IA em comparação com o tempo de formação de normas, o que representa um desafio sem precedentes para as sociedades. Bockting *et al*. (2023) argumentam sobre a necessidade de orientações dinâmicas para o uso de IA generativa, realçando a importância da supervisão científica para manter a confiança pública e a integridade científica diante dos riscos que as tecnologias de IA apresentam. Clyde (2022) propõe que a interdisciplinaridade, a democratização e a justificação coerente são forças motrizes essenciais no desenvolvimento de IA para a ciência, com o objetivo de acelerar a descoberta científica e abordar questões globais urgentes.

Em seu estudo sobre IA em radiologia e medicina nuclear, Kocak *et al*. (2023) destacam a escassa disponibilidade de dados e modelos, evidenciando como barreiras à replicação e validação científica. Van Dis *et al*. (2023) destacam a importância de abordar desafios críticos trazidos pela adoção de tecnologias de IA conversacional no campo científico. O estudo enfatiza a necessidade de estratégias para adaptar-se a estas inovações, que incluem uma etapa de verificação por especialistas humanos e a formulação de diretrizes que estabeleçam a responsabilidade no uso de ferramentas de IA. É preciso ressaltar que a adoção de conceitos de IA explicável e responsável são peças-chave na busca da transparência e usos éticos.

Por sua vez, Grossmann *et al*. (2023) destacam como os avanços em IA, especialmente os LLM, estão transformando a pesquisa em ciências sociais, oferecendo novas oportunidades para testar teorias e hipóteses sobre comportamento humano em grande escala e velocidade, enquanto também apresentam desafios para a adaptação das práticas de pesquisa. Neste sentido, não só investigadores, mas aqueles que trabalham no tratamento de fontes de informação, como bibliotecários e arquivistas, podem influenciar positivamente o desenvolvimento e uso de inteligências artificiais especializadas (Rockembach, 2021). Observamos que o debate entre IA generativa aberta, como o Mistral, e fechada, como o GPT

(OpenAI Platform, 2024, OpenAI, 2024) e Claude (Anthropic, 2024) é uma questão central na sua adoção, nomeadamente em campos onde a precisão e o conhecimento especializado são fundamentais. Modelos de IA aberta, como o Mistral (Mistral, c2024a), permitem uma ampla colaboração e transparência, oferecendo uma base para a inovação e o desenvolvimento comunitário. No entanto, algumas limitações estão relacionadas à qualidade e à especificidade dos dados, que podem ser apenas aqueles disponíveis publicamente. Modelos de IA como o GPT (OpenAI, 2024), utilizam: 1) dados públicos da internet, 2) licenciados de terceiros e 3) por formadores humanos e utilizadores, permitindo-lhes gerar mais informações relevantes que não estão disponíveis publicamente. O ajuste fino de modelos de IA usando técnicas como o RAG ou *Retriever-Augmented Generation* (Mistral, 2024c; Nguyen, *et al*., 2024) representa um método promissor para aprimorar sua aplicabilidade na ciência, permitindo que modelos preexistentes sejam especializados em novos domínios ou conjuntos de dados específicos, potencializando a precisão e a relevância dos resultados em investigações científicas também específicas.

No contexto da Ciência Aberta, a disponibilização de pesos e parâmetros de Inteligências Artificiais em modelos abertos (Mistral, 2024b) é fundamental para garantir a reprodutibilidade e o rigor científico, contribuindo para uma comunidade científica mais transparente e colaborativa. Além disso, a qualidade dos dados utilizados para treinar essas IAs é outra questão para o desenvolvimento de modelos de IA confiáveis. Portanto, as práticas de coleta, limpeza e compartilhamento de dados precisam ser transparentes e bem documentadas, seguindo padrões éticos e científicos rigorosos.

CONSIDERAÇÕES FINAIS

Por um lado, os modelos de IA são ferramentas promissoras que democratizam o acesso ao conhecimento científico, incentivando a participação global e fomentando a interdisciplinaridade, por outro, a ausência de normativas sólidas e mecanismos de supervisão pode comprometer a integridade da ciência. Portanto, recomenda-se que as prioridades para a pesquisa relacionada a modelos abertos e fechados devem se focar na reprodutibilidade científica, com sistemas controlados de revisão e validação por pares, equilibrando inovação e responsabilidade, garantindo que a IA sirva à ciência de forma ética e transparente.

É possível conceber uma ciência aberta sem também adotar modelos de Inteligência Artificial Aberta? Diante disso, ressalta-se a relevância de desenvolver modelos de IA confiáveis e recomenda-se promover a literacia em inteligência artificial nomeadamente para investigadores no âmbito da ciência. Tornar o conhecimento sobre como as ferramentas de IA funcionam acessível e compreensível é essencial para garantir uma utilização segura e ética. Nesse contexto, as práticas de Ciência Aberta são fundamentais, pois promovem uma cultura de compartilhamento e transparência que é vital para a construção de abordagens guiadas por IA (*AI-driven*) eficazes e confiáveis.

REFERÊNCIAS

ALKAISSI, H.; MCFARLANE, S. I. Artificial allucinations in ChatGPT: implications in scientific writing. **Cureus**, [*s. l.*], v. 15, n. 2, Feb. 19, 2023. Disponível em: https://www.cureus.com/articles/138667-artificial-hallucinations-in-chatgpt-implications-in-scientific-writing#!/. Acesso em: 26 ago. 2024. DOI: https://doi.org/10.7759/cureus.35179.

ANTHROPIC. Introduction to Claude. **Anthropic**, [*s. l.*], 2024. Disponível em: https://docs.anthropic.com/en/docs/intro-to-claude. Acesso em: 26 ago. 2024.

BARONCHELLI, A. Shaping new norms for AI. **Philosophical Transactions of the Royal Society B**, [*s. l.*], v. 379, p. 1-6, 22 Jan. 2024. Disponível em: https://royalsocietypublishing.org/doi/10.1098/rstb.2023.0028. Acesso em: 26 ago. 2024. DOI: https://doi.org/10.1098/rstb.2023.0028.

BOCKTING, C. L.; VAN DIS, E. A.; VAN ROOIJ, R.; ZUIDEMA, W.; BOLLEN, J. Living guidelines for generative AI—why scientists must oversee its use. **Nature**, [*s. l.*], v. 622, p. 693-696, 26 Oct. 2023. Disponível em: https://doi.org/10.1038/d41586-023-03266-1. Acesso em: 26 ago. 2024.

CASTELVECCHI, D. Can we open the black box of AI? **Nature**, [*s. l.*], v. 538, p. 21-23, 6 Oct. 2016. Disponível em: https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731. Acesso em: 26 ago. 2024. DOI: https://doi.org/10.1038/538020a.

CLYDE, A. AI for science and global citizens. **Patterns**, [*s. l.*], v. 3, n. 2, p. 1-3, 11 Feb. 2022. Disponível em: https://www.sciencedirect.com/science/article/pii/S2666389922000198?via%3Dihub. Acesso em: 26 ago. 2024. DOI: https://doi.org/10.1016/j.patter.2022.100446.

GROSSMANN, I.; FEINBERG, M.; PARKER, D. C.; CHRISTAKIS, N. A.; TETLOCK, P. E.; CUNNINGHAM, W. A. AI and the transformation of social science research. **Science**, [*s. l.*], v. 380, n. 6650, p. 1108-1109, 15 June 2023. Disponível em: https://www.science.org/doi/10.1126/science.adi1778. Acesso em: 26 ago. 2024. DOI: https://doi.org/10.1126/science.adi1778.

KOCAK, B.; YARDIMCI, A. H.; YUZKAN, S.; KELES, A.; ALTUN, O.; BULUT, E.; BAYRAK, O. N.; OKUMUS, A. A. Transparency in artificial intelligence research: a systematic review of availability items related to open science in radiology and nuclear medicine. **Academic Radiology**, [*s. l.*], v. 30, n. 10, p. 2254-2266, Oct. 2023. Disponível em: https://www.sciencedirect.com/science/article/pii/S1076633222006353?via%3Dihub. Acesso em: 26 ago. 2024. DOI: https://doi.org/10.1016/j.acra.2022.11.030.

LIN, Z. Why and how to embrace AI such as ChatGPT in your academic life. **Royal Society Open Science**, [*s. l.*], v. 10, n. 8, 23 Aug. 2023. Disponível em: https://royalsocietypublishing.org/doi/10.1098/rsos.230658. Acesso em: 26 ago. 2024. DOI: https://doi.org/10.1098/rsos.230658.

MISTRAL AI. Open weight models. **Mistral AI**, [*s. l.*], c2024a. Disponível em: https://docs.mistral.ai/getting-started/open_weight_models/. Acesso em: 26 ago. 2024.

MISTRAL AI. Fine-tuning. **Mistral AI**, [*s. l.*], c2024b. Disponível em: https://docs.mistral.ai/guides/finetuning/. Acesso em: 26 ago. 2024.

NGUYEN, Z.; ANNUNZIATA, A.; LUONG, V.; DINH, S.; LE, Q.; HA, A. H.; LE, C.; PHAN, H. A.; RAGHAVAN, S.; NGUYEN, C. Enhancing Q&A with domain-specific fine-tuning and iterative reasoning: a comparative study. **ArXiv**, [*s. l.*], Apr. 2024. Disponível em: https://arxiv.org/pdf/2404.11792. Acesso em: 26 ago. 2024. DOI: https://doi.org/10.48550/arXiv.2404.11792.

OPENAI. How ChatGPT and our language models are developed. **OpenAI**, [*s. l.*], 2024b. Disponível em: https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed. Acesso em: 26 ago. 2024.

OPENAI PLATFORM. OpenAI developer platform, **OpenAI Platform**, [*s. l.*], 2024a. Disponível em: https://platform.openai.com/docs/overview. Acesso em: 26 ago. 2024.

ROCKEMBACH, M. Ciência da informação e inteligência artificial: um caminho para arquivos e bibliotecas inteligentes. *In*: CONGRESSO ISKO ESPANHA-PORTUGAL, 5, 2021, Lisboa. **Anais** [...]. Lisboa: Universidade de Lisboa, 2021. p. 235-242. Disponível em: https://lume.ufrgs.br/handle/10183/233477. Acesso em: 26 ago. 2024.

ROCKEMBACH, M.; GEERTS, D. Eticista digital: uma função emergente no campo da informação. **Boletim do Arquivo da Universidade de Coimbra**, Coimbra, v. 37, n. 1, p. 75-93, jul. 2024. Disponível em: https://impactum-journals.uc.pt/boletimauc/article/view/14166. Acesso em: 26 ago. 2024. DOI: https://doi.org/10.14195/2182-7974_37_1_3.

VAN DIS, E. A. M.; BOLLEN, J.; ZUIDEMA, W.; VAN ROOIJ, R.; BOCKTING, C. L. ChatGPT: five priorities for research. **Nature**, [*s. l.*], v. 614, p. 224-226, 9 Feb. 2023. Disponível em: https://www.nature.com/articles/d41586-023-00288-7. Acesso em: 26 ago. 2024. DOI: https://doi.org/10.1038/d41586-023-00288-7.