Naif Arab University for Security Sciences
Journal of Information Security and Cybercrimes Research
مجلة بحوث أمن المعلومات والجرائم السيبرانية
https://journals.nauss.edu.sa/index.php/JISCR

JISCR

AI in the Era of Fakes and Deepfakes: Risk of Fabricated Photographs and Identities in Academic Publishing

Jaime A. Teixeira da Silva*
Independent researcher, Ikenobe 3011-2, Kagawa-ken, 761-0799, Japan.

Received 22 Oct. 2023; Accepted 08 Nov 2023; Available Online 28 Nov. 2023.

Abstract

Academic publishing has entered an era of fake, including fake authors who are either real entities using fake credentials, or totally concocted personalities that give the impression of real humans. Both can be achieved via the use of artificial intelligence (AI) and software that is capable of completing such a task, and ultimately a deepfake is created. The creation of fictitious deepfakes, even more so when assisted or driven by AI, allows creators to not only establish a fake image or photo, but also embed it within a fake context (e.g., profile). For whatever reason, there are risk of deepfakes during manuscript submission and the publication process, as well as on academic social network sites, like ResearchGate, but are academics, journals and publishers sufficiently prepared to detect them?.

There are ample artificial intelligence (AI)-driven software tools to alter faces in photos and videos [1], abilities that should be a reason for concern if their purpose is neither artistic nor scholastic. Such misuse can also impact social media and social science networks, and in the digital world, deepfakes can appropriate identities, while the rights (specifically their images) of individuals can be hijacked. Deepfake imagery, which is profoundly unethical because its premise lies in deception [2], is an issue that may increasingly impact academia. The barriers between academia and society have eroded, and members of the public can easily access the same academic portals that scholars use regularly, including journals and academic social network sites (ASNSs) like ReseachGate where academic profiles are often accompanied by images or photos of individuals. The existence of fake "academic" profiles on ReseachGate has already been noted [3]. But what if the stated author is fake (i.e., a real identity posing as a pseudonymous entity, a completely fabricated entity, or a deepfake)? How then are privacy rights of individuals protected, even posthumously [4]?

Having access to AI-based software that is able to create a fake image, based on existing images or text-inputted projections, would allow a convincingly real set of fake images, such as stand-alone facials or in-context head-and-body shots, of "authors" to be produced, even though such individuals might not exist. In other words, as was achieved for several famous personalities

Keywords: Cybersecurity, Abuse, Deception, Artificial Intelligence, Ethics, Responsibility.

Production and hosting by NAUSS

* Corresponding Author: Jaime A. Teixeira da Silva
Email: jaimetex@yahoo.com
doi: 10.26735/KNJA7076

[5], AI could be used to create fake images of real scientists, i.e., deepfakes [6]. How effective are current methods to detect deepfakes [7]?

Fake authors already form part of the academic publishing landscape, increasingly encroaching upon legitimate scholarly territory. One prominent example are paper mills, which are for-profit services that provide fake data, images or text – even whole papers – on demand, including the possibility of inserting "authors" who have done nothing, i.e., authorship on sale schemes [8]. For journals requiring a photo for authors, would such dishonest authors who employ paper mills use their real photos, or might they revert to the use of deepfakes? Provided that there is a market demand, or a reason for using fake human images, the risk of deepfakes in academia will surely grow. While much distracting attention is being paid to ChatGPT as a disruptive AI-driven large language model, and its impact on research and academia, mainly from a text-based perspective [9], not enough attention is being paid to AI-driven software that can create fake imagery and deepfakes for innocuous or nefarious purposes.

What situations can be envisaged for using fake human images in academic publishing, and what could be the motivations of those who employ such techniques? As was noted above, the first possibility would be to illustrate or populate "academic" profiles on ASNSs, or as photos in academic papers. One nefarious use of fake entities is to "sting" journals by tricking them into publishing papers with junk or fake content, as a way to prove that they are either "predatory", or that they make false claims regarding peer review [10]. Such sting operations, devised by real individuals, might employ fake identities (or pseudonyms) – some with seemingly remarkable academic credentials – to trick (or sting) status quo legitimate as well as unscholarly journals entities that they wish to expose or shame [11]. Such actions amount to narcissistic attention-seeking behavior, in the notion that by exposing unscholarly practices of journals, using deceptive techniques, that they are somehow exercising an act of "good", and thus feel gratified by such actions, often motivating them to repeat such actions [12]. Of note, individuals who use such AI software to create deepfakes possess the financial means to order them (on-demand deepfakes) from a service akin to a paper mill, and/or have the skills, time and finances to create such deepfakes themselves. Another possible reason why some may create academic deepfakes could be to engage in the impersonation, for example of a rival, as an act of dark humor. Finally, some uses of deepfakes could actually have a legitimate reason, such as to protect individuals' privacy [13].

Academia and the publishing industry should develop sufficiently sensitive techniques that are able to detect deepfakes on platforms such as ResearchGate, to protect the image and reputation of that platform, and its users.

Funding

This article did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The author declares that he has no conflicts of interest.

References

[1] C. Otto, J. Naruniec, L. Helminger, T. Etterlin, G. Mignone, P. Chandran, G. Zoss, C. Schroers, M. Gross, P. Gotardo, D. Bradley, and R. Weber, "Learning dynamic 3D geometry and texture for video face swapping," Comp. Graphics Forum, vol. 41, no. 7, pp. 611–622, 2022, doi: 10.1111/cgf.14705

[2] A. de Ruiter, "The distinct wrong of deepfakes," Phil. Technol., vol. 34, no. 4, pp. 1311–1332, 2021, doi: 10.1007/s13347-021-00459-2

[3] N. C. Eva, and T. A. Wiebe, "Whose research is it anyway? Academic social networks versus institutional repositories," J. Libr. Schol. Commun., vol. 7, article no. eP2243, 2019, doi: 10.7710/2162-3309.2243

[4] A. C. Heugas, "Protecting image rights in the face of digitalization: A United States and European analysis," J. World Intellect. Prop., vol. 24, no. 5-6, pp. 344–367, 2021, doi: 10.1111/jwip.12194

[5] B-S. Cho, B. M. Le, J-W. Kim, S. Woo, S. Tariq, A. Abuadbba, and K. Moore, "Towards understanding of deepfake videos in the wild," In: CIKM '23: Proceedings of the 32nd ACM International Conference on Information and Knowledge

Management, October 2023, Association for Computing Machinery, New York, NY, USA, pp. 4530–4537, 2023, doi: 10.1145/3583780.3614729

[6] M. Mustak, J. Salminen, M. Mäntymäki, A. Rahman, and Y. K. Dwivedi, "Deepfakes: Deceptions, mitigations, and opportunities," J. Bus. Res., vol. 154, article no. 113368, 2023, doi: 10.1016/j.jbusres.2022.113368

[7] L. Stroebel, M. Llewellyn, T. Hartley, T. S. Ip, and M. Ahmed, "A systematic literature review on the effectiveness of deepfake detection techniques," J. Cyber Sec. Technol., vol. 7, no. 2, pp. 83–113, 2023, doi: 10.1080/23742917.2023.2192888

[8] J. A. Teixeira da Silva, "A dangerous triangularization of conflicting values in academic publishing: ORCID, fake authors, and the lack of criminalization of the creators of fake elements," Epistēmēs Metron Logos, vol. 7, pp. 1–10, 2022, doi: 10.12681/eml.27238

[9] E. A. M. van Dis, J. Bollen, W. Zuidema, R. van Rooij, and C. L. Bockting, "ChatGPT: Five priorities for research," Nature, vol. 614, no. 7947, pp. 224–226, 2023, doi: 10.1038/d41586-023-00288-7

[10] J. A. Teixeira da Silva, "An alert to COVID-19 literature in predatory publishing venues," J. Acad. Libr., vol. 46, no. 5, article no. 102187, 2020, doi: 10.1016/j.acalib.2020.102187

[11] J. A. Teixeira da Silva, "Assessing the ethics of stings, including from the prism of guidelines by ethics-promoting organizations (COPE, ICMJE, CSE)," Publ. Res. Quart., vol. 37, no. 1, pp. 90–98, 2021, doi: 10.1007/s12109-021-09784-y

[12] F. Edwards, "An investigation of attention-seeking behavior through social media post framing," Athens J. Mass Media Commun., vol. 3, no. 1, pp. 25–44, 2016, doi: 110.30958/ajmmc.3.1.2

[13] M. Khamis, H. Farzand, M. Mumm, and K. Marky, "DeepFakes for privacy: Investigating the effectiveness of state-of-the-art privacy-enhancing face obfuscation methods," In: AVI 2022: Proceedings of the 2022 International Conference on Advanced Visual Interfaces, June 2022, Association for Computing Machinery, New York, NY, USA, article no. 21, pp. 1–5, doi: 10.1145/3531073.3531125