Department	School of Computing
Supervisors	Sean McKeown, Bill Buchanan
Project Title	Mitigating the Threat of Digital Media Produced or Modified by Artificial Intelligence (AI): A Digital Forensics Perspective

PROJECT DESCRIPTION

We live in a multimedia-centric society, where images and video are increasingly leveraged in social media, news, and entertainment. The ubiquity of multimedia files, and the accessibility of modern creation and editing tools, has created a very difficult environment in which to ascertain the veracity, or integrity, of such media. Contemporary election cycles in various countries have increased the prominence of fake news and edited media, such that the problems generated by poor image authentication tooling are becoming impactful and widespread.

Recent advances in Artificial Intelligence (AI) have compounded the problem, as advanced forgeries and edits can now be generated without the need for sophisticated skills or large quantities of time. With AI assisted technologies becoming much more accessible, the general public can now generate Deep Fake videos, access mobile phones with dynamic object/person removal, and even utilise text-to-image artificial image generators from Dall-E 2 and Stable Diffusion. All of these approaches can potentially generate harmful, abusive, or misleading media.

With the increasing potential for abuse, we require sufficient Digital Forensics research to allow for image verifiability and authentication in this new AI frontier. Without such techniques, there is the potential that innocent people are prosecuted based on falsified media, or images are manipulated in such a way that criminals can evade justice. There is also the possibility of needing to detect AI produced images in order to prevent targeted harassment campaign and the generation and distribution of images conveying illegal content.

This problem space is complex and multi-faceted, and it is not expected that it can be solved, or completely addressed, within a single PhD project. However, there is room to build on existing work in image forensics, source verification/identification, manipulation detection, and machine learning techniques, in order to begin to combat the problem. The digital forensics context is particularly important for this project, with the goal of not just producing a theoretical solution or framework, but to produce actionable results or practical tooling that would allow practitioners to combat manipulated media in real world investigations.

Perspective applicants are encouraged to contact the Supervisor before submitting their applications. Applications should make it clear the project you are applying for and the name of the supervisors.

Academic qualifications
A first degree (at least a 2.1) ideally in a Computing related discipline with a good fundamental knowledge of Digital Forensics, AI, or Multimedia Formats.

English language requirement
IELTS score must be at least 6.5 (with not less than 6.0 in each of the four components). Other, equivalent qualifications will be accepted. Full details of the University's policy are available online.

Essential attributes:

- Experience of fundamental Computing and Digital Forensics

- Competent in Basic Programming
- Knowledge of Cybersecurity and Digital Media
- Good written and oral communication skills
- Strong motivation, with evidence of independent research skills relevant to the project
- Good time management

Desirable attributes:
Digital Forensics, Multimedia Formats and Encoding, Aritificial Intelligence, Software Development (e.g. Python), Some Statistical Knowledge

Indicative Bibliography	Z. Sha, Z. Li, N. Yu, and Y. Zhang, "DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Diffusion Models." arXiv, Oct. 13, 2022. Accessed: Nov. 18, 2022. [Online]. Available: http://arxiv.org/abs/2210.06998 J. Rando, D. Paleka, D. Lindner, L. Heim, and F. Tramèr, "Red-Teaming the Stable Diffusion Safety Filter." arXiv, Nov. 10, 2022. Accessed: Nov. 18, 2022. [Online]. Available: http://arxiv.org/abs/2210.04610 R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models." arXiv, Nov. 01, 2022. Accessed: Nov. 18, 2022. [Online]. Available: http://arxiv.org/abs/2211.00680 M.-H. Maras and A. Alexandrou, "Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos," The International Journal of Evidence & Proof, vol. 23, no. 3, pp. 255–262, Jul. 2019, doi: 10.1177/1365712718807226.
Enquiries	For informal enquiries about this PhD project, please contact s.mckeown@napier.ac.uk
Web page	https://www.napier.ac.uk/research-and-innovation/research-degrees/application-process