Deepfake detection tools must work with dark skin tones, experts warn

Publication Date: 2023-08-17

Author: Hibaq Farah

Section: Technology

Tags: Artificial intelligence (AI), Race, Computing, Deepfake, news

Article URL: https://www.theguardian.com/technology/2023/aug/17/deepfake-detection-tools-must-work-with-dark-skin-tones-experts-warn

Detection tools being developed to combat the growing threat of deepfakes – realistic-looking false content – must use training datasets that are inclusive of darker skin tones to avoid bias, experts have warned. Most deepfake detectors are based on a learning strategy that depends largely on the dataset that is used for its training. It then uses AI to detect signs that may not be clear to the human eye. This can include monitoring blood flow and heart rate. However, these detection methods do not always work on people with darker skin tones, and if training sets do not contain all ethnicities, accents, genders, ages and skin-tone, they are open to bias, experts warned. Over the last couple of years, concerns have been raised by AI and deepfake detection experts who say bias is being built in these systems. Rijul Gupta, synthetic media expert and co-founder and CEO of DeepMedia, which uses AI and machine learning to assess visual and audio cues for underlying signs of synthetic manipulation said: “Datasets are always heavily skewed towards white middle-aged men, and this type of technology always negatively affects marginalised communities.” “At DeepMedia, instead of being race blind, our detectors and our technology actually look for a person’s age, race, gender. So when our detectors are looking to see if the video has been manipulated or not, it has already seen a large amount of samples from various ages and races.” Gupta added that deepfake detection tools that use visual cues, such as blood-flow and heart-rate detection, can have “underlying biases towards people with lighter skin tones, because darker skin tones in a video stream are much harder to extract a heart rate out of”. The “inherent bias” in these tools means that they will perform worse on minorities. “We will see an end result of an increase of deepfake scams, fraud and misinformation caused by AI that will be highly targeted and focused on marginalised communities”, Gupta says. Mutale Nkonde, AI policy adviser and the CEO and founder of AI for the People, said the concerns tap into larger exclusions minorities face. “If we’re gonna have a technology that is maintaining the security of some people it really should maintain the security of all and, unfortunately, the technology isn’t quite there yet,” Nkonde said. “We are well educated around the issues that facial recognition has in recognising dark skin, but the general public don’t realise that just because the technology has a new name, function or use doesn’t mean that the engineering has advanced. “It also doesn’t mean that there is no new thinking in the field. And because there is no regulation anywhere in the world that says: ‘You can’t sell a technology that doesn’t work,’ the underlying bias continues and is reproduced in new technologies.” Ellis Monk, professor of sociology at Harvard University and visiting faculty researcher at Google, developed the Monk Skin Tone Scale. It is an alternative scale that is more inclusive than the tech-industry standard and will provide broader spectrum of skin tones than can be used for datasets and machine learning models. Monk said: “Darker skinned people have

been excluded from how these different forms of technology have been developed from the very beginning. “There needs to be new datasets constructed that have more coverage, more representativeness in terms of skin tone and this means you need some kind of a measure that is standardised, consistent and more representative than prior scales.”