|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | Sean Hughes  E Sean.Hughes@UGent.be  T 0032 (0)9 264 86 49  Faculty of Psychology and Educational Sciences  Henri Dunantlaan 2  B-9000 Ghent  Belgium |
|  | |  |
|  |  |  |
| date  30/01/2021 | page  1 | our reference  Manuscript Submission | |
|  |  |  |  |

Dear Editor,

Please find attached a manuscript titled “Using Deepfakes to Hack the Human Mind” to be considered for publication in *Nature Human Behavior* by Sean Hughes, Ohad Fried, Melissa Ferguson, Ciaran Hughes, Rian Hughes, Xinwei Yao, and Ian Hussey. Our manuscript best fits the ‘Article’ rubric of the journal.

Conventional wisdom dictates that seeing is believing. However, thanks to recent advances in artificial intelligence, this may no longer be the case. A branch of machine learning known as ‘deep learning’ has made it increasingly easy to take a person’s likeness (whether their face, voice, or writing style), feed that data to a computer algorithm, and have it generate a synthetic copy or “Deepfake”.

Deepfaking has quickly become a tool of harassment against activists, and a growing concern for those in the business, entertainment, and political sectors. Concern grows that this new technology may be used to spread disinformation, fuel social tensions, and undermine election outcomes.

Recognizing these dangers, politicians have called for legislation to regulate this new technology while industry leaders (Facebook, Google, and Microsoft) seek algorithms to detect and excise it from their platforms. However, no law or algorithm can guarantee that the public will be completely insulated from Deepfakes.

What is needed then, alongside legislative and technological solutions, is a greater focus on the *Psychology of Deepfakes*: how this new technology influences our thoughts, feelings, and actions. For instance, can Deepfakes be used to manipulate our (implicit) attitudes and intentions? How effective are they in doing so, especially when compared to genuine content? Are people aware of this new technology, and more importantly, can they detect when they are being exposed to it? Finally, does awareness of Deepfaking and the ability to detect when it is present immunize people from its influence?

We systematically explored these questions in seven pre-registered experiments (N = 2558). By synthetically cloning a person’s appearance (video) and voice (audio), and then using this Deepfake to manipulate what he ‘said’, we were able to control (implicit) attitudes and intentions towards him. Results also indicated that Deepfakes psychologically impacted viewers just as effectively as genuine content did. Our results show that most people are unaware of this new technology, find it difficult to detect when they are being exposed to it, and perhaps most importantly that neither awareness nor detection serves to protect them from its influence.

Deepfakes have the potential to be weaponized to manipulate not only the public’s attitudes and intentions, but to alter what they remember, who they trust, and what they believe. Instead of asking if a specific image, video, or audio clip is authentic, this new technology may cause us to question *everything* we see and hear, thereby accelerating a growing trend towards ‘epistemic breakdown’ (an inability or reduced motivation to distinguish fact from fiction).

We conclude our paper by considering these issues, and highlight the need for psychological interventions that inoculate the public against Deepfakes, and together with technology and legislation, create a shared informational immune system that safeguards our individual and collective belief in truth.

Our paper will be of interest to NHB’s readership given the multi-disciplinary nature of the phenomenon under investigation (Deepfaking) and its societal implications. All authors have approved the current version of the manuscript and it is not under review elsewhere nor has it been previously published or accepted for publication.

Kind Regards, Sean Hughes