|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | 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  20/02/2021 | page  1 | our reference  Manuscript Submission | |
|  |  |  |  |

Dr. Berenbaum,

Please find enclosed our manuscript, “Deepfaked Online Content is Highly Effective in Manipulating People’s Attitudes and Intentions” for your consideration as a Research Report in the *Proceedings of the National Academy of Arts and Sciences*.

In recent times, disinformation has spread rapidly through social media and news sites, biasing our (moral) judgements of other people and groups. “Deepfakes”, a new type of AI-generated media, represents a powerful new tool for spreading disinformation online. Although Deepfaked images, videos, and audio may appear genuine, they are actually hyper-realistic fabrications that enable one to digitally control what another person says or does (see <https://youtu.be/cQ54GDm1eL0>).

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

Recognizing these dangers, politicians have called for legislation to regulate Deepfaking while industry leaders (Facebook, Google, and Microsoft) are developing algorithms to detect and eliminate it from their platforms. But legislation and technology alone won’t be enough. We need to start studying the *Psychology of Deepfakes* and their ability to manipulate our thoughts, feelings, and actions.

Towards this end, we carried out seven pre-registered studies (*N* = 2558) which were the first of their kind. Our goal was to examine how Deepfaked online content psychologically impacts viewers. In our studies, we extracted a person’s face (video) and voice (audio) from genuine recordings, and used an artificial intelligence technique known as a Generative Adversarial Network (GAN), to create a Deepfake of that individual. We then had this Deepfake confess to either virtuous or horrible past actions. These fabricated confessions were inserted into videos and uploaded to YouTube where participants watched them. Their implicit and explicit attitudes as well as intentions towards the target were then assessed.

We consistently found that Deepfaked online content had a strong psychological impact on the viewer, and allowed the content creator to take control the public perceptions of a given target. Results also indicated that Deepfakes are just as effective in biasing (implicit) attitudes and intentions as genuine content; that most people are unaware that Deepfaking even exists; find it difficult to detect when they are being exposed to it; and perhaps most importantly, neither awareness nor detection served to protect them from its influence.

Although politicians, journalists, academics, and think-tanks have all warned of the dangers that Deepfakes pose, our paper is the first to offer systematic empirical support for such claims.

We believe our empirical and conceptual contributions, as well as the wider societal implications of Deepfaking, make this paper of broad interest to those in the fields of journalism, psychology, neuroscience, political science and law.

We thank you for your consideration and look forward to hearing from you soon.

Sincerely,

Sean Hughes Ohad Fried Melissa Ferguson Ciaran Hughes Rian Hughes Xinwei Yao Ian Hussey