|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | 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  27/07/2021 | page  1 | our reference  Manuscript Submission | |
|  |  |  |  |

Dear Editor,

Please find enclosed our manuscript, “Deepfaked Online Content is Highly Effective in Manipulating People’s Attitudes and Intentions” for your consideration as an Article in *Nature Communications*.

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 thought, feeling, and action.

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 can be used to control (automatic) attitudes and behavioral intentions towards others. In our studies, we extracted a target individual’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 same individual. At its core, this Deepfake constituted a digital copy of the target that provided us with total control over his appearance and actions. We used this Deepfake to make him confess to either virtuous or horrible behaviors. These fabricated confessions were then inserted into videos and uploaded to YouTube where participants watched them. Their perceptions (automatic and self-reported attitudes as well as behavioral intentions) of the target were then assessed.

Across studies we consistently found that Deepfakes can quickly and powerfully impact viewers, equipping their creators with a ready means of controlling how others are perceived. This is true for different types of Deepfaked content (video and audio) and different Deepfake creation methods (‘cut and paste’ vs. ‘fabricate from scratch’). Results showed that many people are unaware that Deepfaking is even possible; find it difficult to detect when they are being exposed to it; and most importantly, neither awareness nor detection serves 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 evidence in support of 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 *Nature Communications’* readership, especially those working in the fields of human behavior and social science (e.g., 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