# Facebook’s problems moderating deepfakes will only get worse in 2020

Last summer, a video of Mark Zuckerberg circulated on Instagram in which the Facebook CEO appeared to claim he had “total control of billions of people’s stolen data, all their secrets, their lives, their futures.” It turned out to be an art project rather than a deliberate attempt at misinformation, but Facebook allowed it to stay on the platform. According to the company, it didn’t violate any of its policies.

For some, this showed how big tech companies aren’t prepared to deal with the onslaught of AI-generated fake media known as deepfakes. But it isn’t necessarily Facebook’s fault. Deepfakes are incredibly hard to moderate, not because they’re difficult to spot (though they can be), but because the category is so broad that any attempt to “clamp down” on AI-edited photos and videos would end up affecting a whole swath of harmless content.

Banning deepfakes altogether would mean removing popular jokes like gender-swapped Snapchat selfies and artificially aged faces. Banning politically misleading deepfakes just leads back to the same political moderation problems tech companies have faced for years. And given there’s no simple algorithm that can automatically spot AI-edited content, whatever ban they do enact would mean creating even more work for beleaguered human moderators. For companies like Facebook, there’s just no easy option.

*“it applies to such a huge category of thing that it’s unclear if it means anything at all”*
“If you take ‘deepfake’ to mean any video or image that’s

edited by machine learning then it applies to such a huge category of thing that it's unclear if it means anything at all," Tim Hwang, director of the Harvard-MIT Ethics and Governance of AI Initiative, tells *The Verge*. "If I had my druthers, which I'm not sure if I do, I would say that the way we should think about deepfakes is as a matter of intent. What are you trying to accomplish at the sort of media that you're creating?"

Notably, this seems to be the direction that big platforms are actually taking. Facebook and Reddit both announced moderation policies that covered deepfakes last week, and rather than trying to stamp out the format altogether, they took a narrower focus.

Facebook said it will remove "manipulated misleading media" which has been "edited or synthesized" using AI or machine learning "in ways that aren't apparent to an average person and would likely mislead someone into thinking that a subject of the video said words that they did not actually say." But the company noted that this does not cover "parody or satire" *or* misleading edits made using traditional means, like last year's viral video of House Speaker Nancy Pelosi supposedly slurring her words.

New apps like Doublicat (pictured) will make deepfakes easier and more fun.
*Credit: Doublicat*

Reddit, meanwhile, didn't mention AI at all, but instead said it will remove media that "impersonates individuals or entities in a misleading or deceptive manner." It's also created an exemption for "satire and parody," and added that it will "always take into account the context of any particular content" — a broad caveat that gives its mods a lot of leeway.

As many have pointed out, these policies are full of loopholes. Writing at *OneZero*, Will Oremus notes that Facebook's only covers edited media which includes speech, for example. This means that "a deepfake video that makes it look like a politician burned the

American flag, participated in a white nationalist rally, or shook hands with a terrorist" would not be prohibited — something Facebook confirmed to Oremus.

These are glaring omissions, but they highlight the difficulty separating deepfakes from the underlying problems of platform moderation. Although many reports in recent years have treated "deepfake" as synonymous with "political misinformation," the actual definition is far more broad. And the problem will only get worse in 2020.

*Apps will make deepfake generation easier, and much more fun*
While earlier versions of deepfake software took some patience and technical skill to use, the next generation will make creating deepfakes as easy as posting. Apps that use AI to edit video (the standard definition of a deepfake) will become commonplace, and as they spread — used for in-jokes, brand tweets, bullying, harassment, and everything in between — the idea of the deepfake as a unique threat to truth online will fade away.

Just this week, an app named Doublicat launched on iOS and Android that uses machine learning to paste users' faces onto popular reaction GIFs. Right now it only works with preselected GIFs, but the company's CEO told *The Verge* it'll allow users to insert faces into any content they like in future, powered by a type of machine learning method known as a GAN.

Does all this make Doublicat a deepfake app? Yep. And will it undermine democracy? Probably not.

Just look at the quality of its output, as demonstrated by the

GIF below, which shows my face pasted onto Chris Pratt's in *Parks and Recreation.* Technologically, it's impressive. The app made the GIF in a few seconds from just a single photo. But it's never going to be mistaken for the real thing. Meanwhile, the creator of TikTok, ByteDance, has been experimenting with deepfake features (though it says they won't be incorporated into its wildly popular app) and Snapchat recently introduced its own face-swapping tools.

*The author as Chris Pratt*.

Hwang argues that the dilution of the term deepfakes could actually have benefits in the long run. "I think the great irony of people saying that all of these consumer features are also deepfakes, is that it in some ways commoditizes what deepfake means," says Hwang. If deepfakes become commonplace and unremarkable, then people will "get comfortable with the notion of what this technology can do," he says. Then, hopefully, we can understand it better and focus on the underlying problems of misinformation and political propaganda.

It's possible to argue that the problem of moderating deepfakes on social media has been mostly a distraction from the start. AI-edited political propaganda has failed to materialize in a meaningful way, and studies show that the vast majority of deepfakes are nonconsensual porn (making up 96 percent of online deepfake videos).

Social media platforms have happily engaged in the debate over deepfake moderation, but as Facebook and Reddit's recent announcements show, these efforts are mostly a sideshow. The

core issues have not changed: who gets to lie on the internet, and who decides if they’re lying? Once deepfakes cease to be believable as an existential threat to truth, we’ll be left with the same, unchanging questions, more pressing than ever before.

https://www.theverge.com/2020/1/15/21067220/deepfake-moderation-apps-tools-2020-facebook-reddit-social-media