BIZCOMMUNITY

# Discriminatory AI and disinformation powered by deep fakes

By Anna Collard

10 Jan 2022

Remember Skynet, the artificial intelligence that wanted to wipe out humanity in the Terminator movies? Now that is an example of AI gone wrong. Luckily, this will not be the case for us in 2022. AI today is by far not as advanced yet. But the movie does raise a couple of interesting questions. For example, how do we define ethics when it comes to developing and applying AI?

**Anna Collard**

Here are some of the concerns where AI might go wrong and that I believe need more awareness in 2022:

## Gender and racial bias in AI

According to a Unesco report, only 12% of the artificial intelligence researchers and 6% of the software developers are women. Women of colour are even less represented. The field is predominantly White, Asian and male. These White middle-class men simply cannot be aware of the needs of all of humanity. And the tech that they develop is inevitably biased towards White middle-class men.

Because of the way machine-learning works, when you feed it biased data, it gets better and better—at being biased. This means that if there is any sexism or racism, based on conscious or unconscious bias, embedded within the data, the algorithm will pick up that pattern. We have already seen examples of self-driving (i.e. AI-driven) cars, disregarding certain ethnicities when deciding how to avoid collisions. Does this make a car racist? Well, not on purpose. The developers simply omitted to provide enough qualified training models for the AI in the car to learn from. This essentially created a bias that affected its decision-making process negatively.

In her book *Invisible Woman*, Caroline Criado-Perez, explains the impacts of data gaps where these algorithms, unless addressed, will result in far-reaching consequences exacerbating gender (and racial) inequality. This highlights the need to

increase general awareness within society regarding the negative and positive implications of AI for girls, women and gender non-binary people as well as the need for more African representation in the AI field and global policy decision making around AI ethics and rules.

## Deep fakes

Deep fakes — a hybrid of the terms “deep learning” and “fake” are realistic video and audio recordings that use artificial intelligence and "deep" learning to create "fake" content. The technology can replace faces and speech to make it appear as if someone said or did something that never happened.

Deep fakes, when done right, can be used to create content with an extremely high potential to deceive. All you need is a powerful computer and enough existing data about the person you want to replace the original media content with to ‘train’ the AI (deep learning algorithms) and to create new realities.

Deep fakes have a great future in the film industry, for instance in reproducing a shot without the actual actor having to be flown in every time. Or in the medical field, recreating someone’s voice if they lost it.

But deep fakes can also be used for more nefarious goals. In 2020, the FBI already warned about a combination of deep fakes as an addition to the highly successful social engineering attack form called Business Email Compromise (Bec).

Effectively leveraging AI to add credibility to an attack by creating a deep fake audio message impersonating a legitimate requestor, such as the CEO of a company authorising a fraudulent money transfer (common practice in a Bec attack).

## Disinformation

According to the MIT paper *The Spread of True and False News Online*, it takes true stories about six times as long to reach 1,500 people as it does for false stories to reach the same number of people. This is due to the emotive nature of misinformation, triggering readers with surprise and disgust, and making them more likely to share.

Due to the pandemic, most political meetings are now being held virtually – which opens up the opportunity to leak voice and video recordings. These recordings can be highly misleading because they lack the crucial context in which a specific comment was made.

Imagine combining these leaked recordings with deep fake technology that change the meaning of what was said and potentially trigger emotional responses in anyone listening. These clips may then have a powerful and damaging impact when making the rounds on WhatsApp, Telegram or other chat apps. These platforms lend themselves to spread disinformation because they are not easily monitored and people are used to trusting voice notes from their groups.

This means that the political views of potentially millions of voters could be negatively influenced. South Africa’s government stepped in to try to stop the spread of misinformation by introducing legislation that made spreading fake information a

prosecutable offence, but how they will enforce this remains to be seen.

One organisation that attempts to curb the spread of fake news and misinformation is the Real411 group, which provides a platform for the public to report digital harms, including disinformation. Special attention is given to topics such as Covid-19 and during election periods to complaints about elections.

While the positive application of new technology might be plentiful, there is always scope for abuse. This is certainly the case with AI and AI-related technology such as deep fakes. It requires new innovative approaches (perhaps the use of NFT to ensure the validity of video content?), forward-thinking policies as well as more awareness to effectively prepare our societies for this new reality.

## ABOUT ANNA COLLARD


Anna Collard is the senior vice president of content strategy and tech evangelist at KnowBe4 Africa

- #BizTrends2022: Discriminatory AI and disinformation powered by deep fakes - 10 Jan 2022
- 5 important lessons to learn from the REvil ransomware attack - 13 Jul 2021
- Here's how hackers break into the business environment and how it can be avoided - 11 Jun 2021
- PoPI Act readiness: 6 things to do - 12 Apr 2021
- Top IT security threats in 2021 - 20 Jan 2021

View my profile and articles...