OPEN FORUM

Deepfakes enable sexual violence

By Theodore Kim

Last week, former Hollywood producer Harvey Weinstein's second criminal trial over alleged sexual assaults began — five years after the first stories exposing his illicit actions against women came to light, energizing and expanding the #MeToo movement. While the revelations brought much overdue attention to the complicity of the film industry in widespread sexual violence, Hollywood is still complicit in a form of intimate violence that hardly anyone is talking about: deepfake revenge porn.

Deepfake technologies allow users to digitally insert anybody's face into any video clip. In 2017, for example, the actress Gal Gadot had her face credibly swapped into a pornographic video.

A generation ago, nascent versions of this technique were used to painstakingly insert Tom Hanks as a digital Forrest Gump alongside John F. Kennedy. Today, Bruce Willis' face shows up in commercials for Russian telecom with neither his consent nor participation. These techniques are already being used to sow geopolitical chaos, as we saw last March when a fake video appeared of Ukrainian President Volodymyr Zelenskyy ordering his country's citizens to surrender. While that particular video was widely ridiculed as an obvious fabrication, it illustrates where these technologies are headed.

Despite these dangers, the visual effects studio Weta FX recently disclosed its latest version of this technology used in the Marvel film "Shang-Chi and the Legend of the Ten Rings." Weta FX claimed that openly discussing these techniques combats misinformation and argued that the more widely these algorithms are known, the easier it becomes to detect fakes.

This self-exonerating argument fundamentally misjudges the dynamics of misinformation and ignores the most widespread and clandestine application of the technology: nonconsensual pornography.

Since the video of Gadot appeared in 2017, open-source software projects have emerged with detailed instructions on how to produce similar results with anybody's face. The 2017 video was generated by referencing a large corpus of publicly available images, but current efforts aim to enable face-swapping with a single photo. While it is impossible to know how many people have had their likenesses unwillingly inserted into these videos, a 2019 study of nearly 15,000
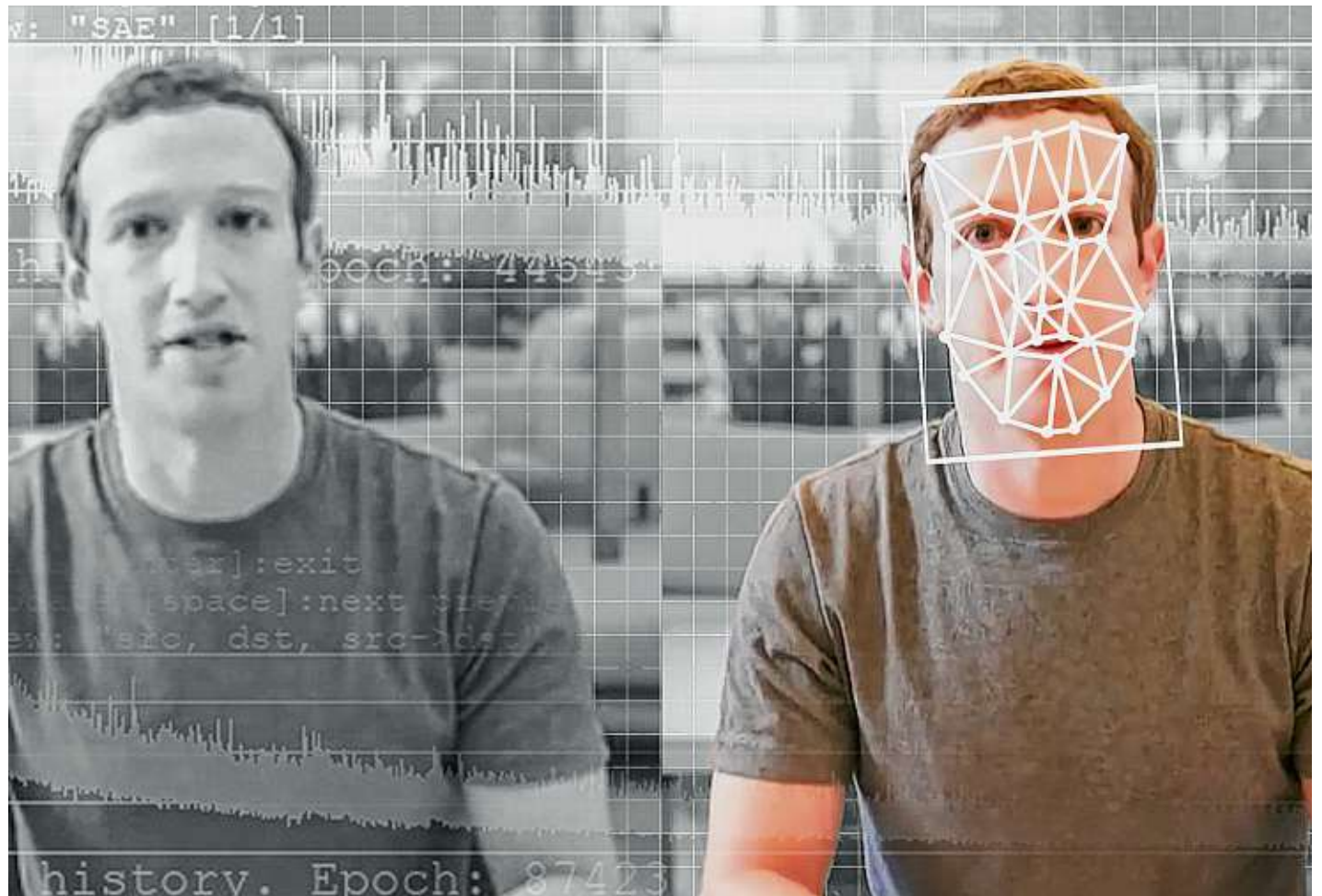
Elyse Samuels/The Washington Post via Getty Images

An original and deepfake of Meta CEO Mark Zuckerberg. The technology has been used for nonconsensual pornography.

deepfakes found that 96% were pornographic.

The Weta FX argument tries to borrow principles from the open-source software movement, where increased public scrutiny improves a system's robustness. While this approach may work when developing software, it is totally misguided when we know the technology is overwhelmingly used for harm.

Some will say that Pandora's box has already opened, so "if we don't do this, somebody else will." This view, that technology evolves along inevitable paths and individuals are powerless to influence it, is known as techno determinism.

But techno determinism is an ideology, not an ironclad law of nature. To the contrary, science is a collective activity, and when social legitimacy and robust community are removed, progress grinds to a halt. Take for example the history of another form of mass violence: the development of nuclear weapons.

The top-secret U.S. Manhattan Project successfully developed the atom bomb because it had both legitimacy and community. Initiated by top Jewish physicists who had just fled Nazi Europe, the project had a clear social goal: Get the bomb before Hitler does. It was a massive, nationwide collaboration by the entire community of U.S. physicists, resulting in the successful Trinity test in 1945.

Lesser known is Hitler's atom bomb project. Werner Heisenberg, the project's leader, was reluctant to build the bomb for the Nazis: The project lacked social legitimacy. Cut off from the rest of the scientific community, his isolated team made basic mistakes. Even upon hearing of the U.S. bombing of Hiroshima, he insisted the device was not atomic, because his (wrong) calculations showed it needed an impossible 10 tons of purified uranium. Scientific breakthroughs happen in socially favorable environments.

Scientists who work to refine deepfake methods — most of whom are men — including researchers at Weta FX, USC and Disney, can no longer pretend that their primary application will be anything other than antisocial weaponizations used against women.

As for the studios backing the research, they have decided shaving a few dollars off the budget of their next film outweighs their sponsoring the genesis of a new form of sexual violence. Deepfakes in Hollywood are purely a cost-saving measure. Facial replacement can already be done by skilled visual effects artists. Automating the process using machine learning just makes films cheaper, not better.

The film industry has the power to erode the social legitimacy of deepfake research and withdraw some of its best minds from its technical community. If other researchers can no longer point to Marvel movies as possible deployments of their techniques, one of their biggest claims to neutrality will be gone. Instead of being seen as a scientifically interesting puzzle with a variety of applications, deepfakes will be seen for what they are: a violently misogynistic line of inquiry.

Theodore Kim is an associate professor of computer science at Yale University and a former senior research scientist for Pixar.