AiA news-service

Outlook

The Dark Art Of Deepfake: How This Height Of Manipulation Can Even Make Mona Lisa Frown

The science of making morphed videos using artificial intelligence is the latest rage on social media. Porn and politics are seeing Deepfake's immediate impact for obvious reasons

SIDDHARTHA MISHRA12 SEPTEMBER 2019

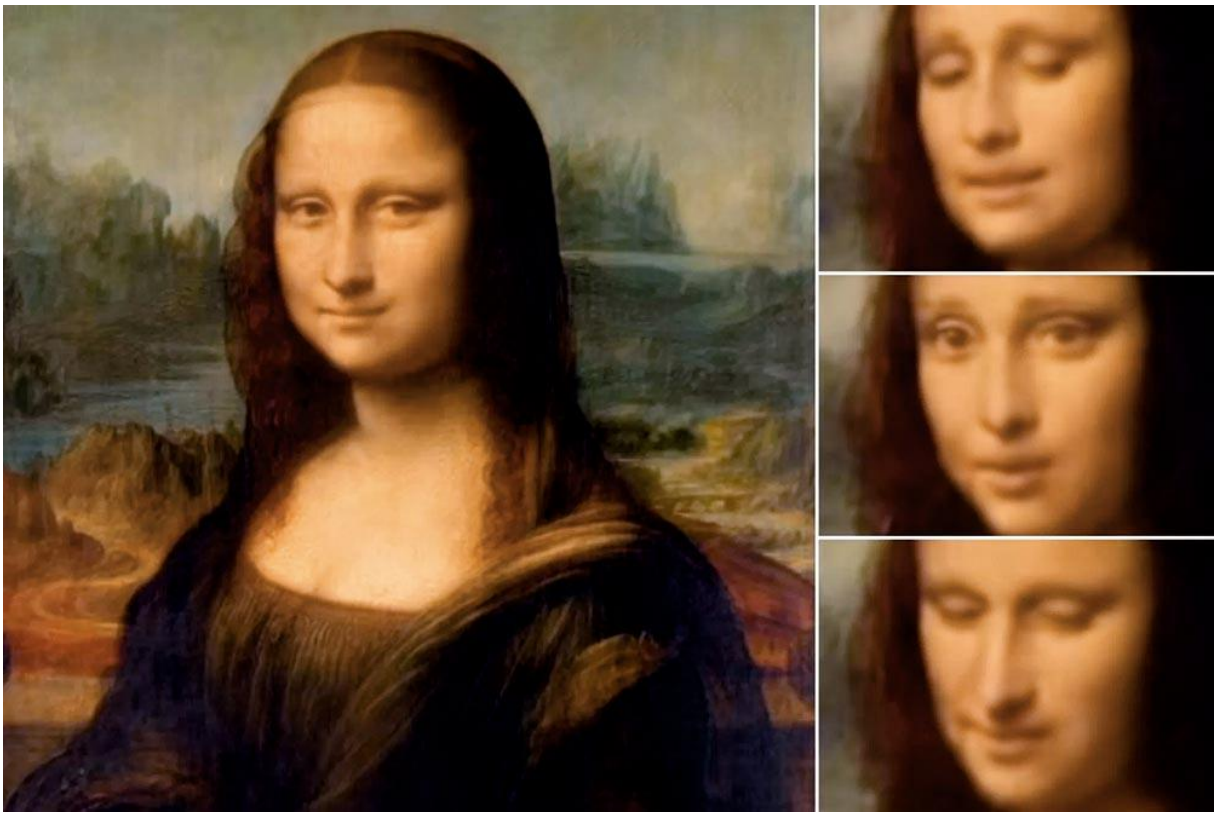
PHOTOGRAPH BY JITENDER GUPTA, IMAGING: DEEPAK SHARMA

You might have seen him already. He was all over TV news spots and social media clips just last week. That puckish, boyish face, unmistakably Chinese, perhaps that of a young brat from one of its glittering coastal cities—but eerily, very eerily, equipped with that equally unmistakable Leonardo DiCaprio mop of blond hair falling over his eyes. Even more unsettling, he was Dicaprio—or at least the characters he played, in those very frames. You may have also heard a name being uttered with awe...Zao! A dreaded Han general? A new virus? No, it's just a new Chinese app that allows its users to do something extremely alarming. The boy in that video turned out to be a harmless, 30-year-old games developer based in Auckland called Allan Xia, who could access the app because of his Chinese number. His use of Zao too was innocuous: "Every media story came embedded with a clip of him strutting around in a Hawaiian shirt in *Romeo+Juliet*, and basking in the golden sunset on deck in *Titanic*," wrote the *South China Morning Post*. But what if he had done something more sinister?

An early warning had come with a Buzzfeed video from 2018, where former US President Barack Obama is saying something...but wait, he *isn't* saying it! It's somebody else who pulls off the digital mask half-way through. But not just you or I, even Obama could have been fooled. It simply looks that *real*. Which means, AI-enabled computer technology has got to the point where you can make anyone's

digital alter-ego say anything! Now imagine this: a video purportedly of Prime Minister Narendra Modi or Rahul Gandhi with things they didn't say. Or of Imran Khan, or a Pakistani general! Or, for that matter, P. Chidambaram or D.K. Shivakumar, or an approver. Or just the girl who lives down the lane who had recently spurned a boy from the neighbourhood...we are perhaps sitting on a time-bomb here. Yes, advanced technology can detect the fake, but what good would that be if war or a riot has already broken out, or someone has already killed herself?

There's always that riddle: with reality itself so dark and bizarre, what use do we have for fakery? This question used to be once asked of art, but those were more innocent days. The building blocks of reality had a certain rock-solid aspect—everyone agreed, at least for form's sake, that the 'truth' was paramount. (Philosophers disagreed only on what the damn thing was!) But now we are stepping into a dystopia whose exact shape we can't trace with our hands or our eyes, ears or minds—our sense-making apparatus. So form-shifting and malleable it is. It's almost as if 'fake' is now a part of 'real'—a bit like night was part of Faiz's dawn. Its job is to unsettle us so thoroughly from our everyday empirical truths that we don't even know what's what anymore. And technology is keeping pace with this new, diabolical need. The latest in the arsenal has an unsettling name too: deepfakes. The Chinese didn't invent it, but only created an app that makes creating deepfake videos ridiculously simple. And make no mistake, it's coming home to you.



Samsung demonstrated how Mona Lisa can be a realistic, talking character in a deepfake video.

The demos of it how it works runs the gamut of human affairs—from politics to pornography, further reducing the distance between those two. The presence of technological magic in politics is nothing new. Just this March, after a few years of everyone playing coy, social media giants operating in India agreed to a 'Voluntary Code of Ethics for the General Elections 2019', after meetings with the Election Commission. The likes of Google, Facebook and Twitter made honourable mention of how they were committed to "increase confidence in the electoral process"—a tacit admission, if any, of manipulation via misinformation. That code was valid during the period of Lok Sabha elections; an updated one may be necessary with elections never far away, because the beast is evolving.

In a globalised world, parallels are easy because technology—and technological victimhood—connects us all. With a President in place who makes real appear fake and vice-versa, the US goes to the polls next year with a clear memory of the - previous one. And how all the talk then, bolstered later, was about pervasive Russian meddling through news/social media manipulation. The fears are back; the modus operandi is expected to be more sophisticated though. No wonder, in mid-July, the US House of Representatives intelligence committee chairman Adam

Schiff wrote to social media companies about a new threat in poll season: Deepfake.

So what exactly is this new mutant beast? Well, the earlier editions of fake news and misinformation as such were in textual form, taking up space as text-forwards on WhatsApp and websites dedicated to its amplification on social media. Images, on the other hand, had that other allure: lulling us with that old adage, ‘seeing is believing’. But the wave soon washed over that realm, with morphed photos, evolving into memes and the kind along the way. Videos were still seen to be inviolable, though. How do you simulate an entire moving ensemble of interconnected images? So, for a long time now, “I’ve seen the video” was the same as saying “I know it happened”. Well, no more. If ‘fake news’ was the Collins Word of the Year for 2017, one could do worse than bet on ‘deepfake’ winning it in the near future.

How did the word come about? You probably remember Deep Thought and Deep Blue, the IBM chess machines that beat grandmasters (the latter even Kasparov). A phrase of 1980s vintage, ‘deep learning’ came from that same edgy world of AI. But now, as that phrase forms a portmanteau with ‘fake’, we’re really being pushed off the deep end. The deepfake phenomenon is already laying siege to the West and slowly creeping into our internet spaces—it’s a simple Google search away. Porn is often the pioneering realm. Mrdeepfakes.com is a website that (self-admittedly) dedicates itself to videos with the faces of celebrities superimposed on - pornographic actors. Emma Watson, the most popular celebrity on the website, has over a hundred fake porn videos to her name there. However, out of the top 10 most-viewed videos on the website, five are of Indian actresses like Aishwarya Rai, Kareena Kapoor, Tamannah Bhatia and Sonakshi Sinha.

A bit of snooping reveals the website is hosted in Amsterdam and that there are over 150 videos of Indian actresses already on the platform. It’s forgery, plain and simple, and a clear violation of an individual’s privacy (besides partaking of the essential misogyny of a lot of porn). Yes, with a name like Mrdeepfakes.com, a URL that hardly suggests they’re real, users will know what they’re in for if they wash up on its shores. But what happens when you get a video forward of the kind on WhatsApp? There’s also the question of convention (yes, even in the world of fakes). While those forging clips abroad cite the original work in some cases, such exacting standards are quite clearly lacking closer home. What’s also interesting is the quality of the videos. The ones made about a year ago were badly made, the

ones being churned out now can easily fool the untrained, unsuspecting eye. In short, soon you must suspend belief, as a general rule, when it comes to videos. But think again of Indian social media, and the millions of newly recruited users…mass gullibility was never more frighteningly at hand.

PHOTOGRAPH BY GETTY IMAGES

It's almost as if 'fake' is a part of 'real' now.... And make no mistake, it's coming home to you.

So what is it and how do they do it? The first instance of a deepfake video was uncovered in late 2017 when Motherboard reported on a Reddit user called "deepfakes" who was since banned from the social media platform. The person picked up images and video footage of, say, a celebrity from all over the Web, and then, using tools from a free, open-source machine-learning software library called TensorFlow, created a kind of digitally manipulable clay which could recreate that face in any desired fashion. These were then inserted, frame-by-frame, into videos that were already out there. What videos? Porn, of course. His unwitting patrons were soon awash with what looked like 'celeb porn'—content that actually a -manipulated celebrity face grafted onto the original faces. Manipulated so skilfully by the software that the celeb would seem to behave exactly the way the original

actress in the video did, right down to replicating expressions, to a scarily unerring degree!

As for the deepfakes of Indian actresses, not all will pass a basic smell test even for the untrained eye. All it takes though is one breakthrough, a piece of code that can be replicated and built upon—something science has invariably delivered. And we are close, believes Anand Sahay, co-founder and CEO of Xebia Global, a company that deals in digital transformation technology.

“Here’s my view. I think it’s already here…I don’t think we have to wait. It was already happening, but it’s just not in the news because those artificial intelligence models were not available,” he says. A recent research paper by David Guera and Edward J. Delp of Purdue University, which put forward ways to detect deepfakes, pins the danger down to how easy it has become: “In recent months, a machine learning-based free software has made it easy to create believable face swaps in videos that leave few traces of manipulation…. The democratisation of modern tools such as Tensorflow or Keras, coupled with the open accessibility of the recent technical literature and cheap access to compute infrastructure, have propelled this paradigm shift…tampering images and videos, which used to be reserved to highly-trained professionals, have become a broadly accessible operation within reach of almost any individual with a computer.”

Think very recent. Wasn’t it just the other day that perhaps even you partook of the #FaceAppChallenge? For those living under a rock, this was all over the news with people on social media and even celebs (who probably stand the most to lose, more on that later) posting images of how they’d look like when they grow old. Well, FaceApp walks essentially the same technological path as deepfakes. We’re talking about deep learning sorcerer’s tools here. One of them, autoencoders—where the machine learns how you look and can then map it out afresh—has been around since the 1980s, being used inter alia for image compression. The other, with the formidable name Generative Adversarial Networks (GAN), is an innovation of 2014 vintage that was called “the most interesting idea in the last 10 years in machine learning”. As the name suggests, it’s a model that can generate or produce content. FaceApp merely took an image you supplied and used these tools to give you a transformation based on parameters like age, gender and hairstyles. If you didn’t hear of privacy fears then, think only of another app here that sounds, and behaves, similarly: FakeApp, a desktop application used to create deepfakes.

PHOTOGRAPH BY GETTY IMAGES

There was a time when every second of animated video cost thousands of dollars. Not anymore. Deepfakes cost under Rs 1 lakh a video, and coming down. “Once the model is trained, it will be open-sourced by some technologist. Other than that, there’s no cost. It’s becoming so scary…with FakeApp, you can do it in minutes with an absolutely trained model, which is available,” says Sahay. He’s not being unduly alarmist. A cursory search on the open-source platform Github reveals over a hundred source codes to create deepfakes. FaceApp “is a Russian company so it’s in the news, and it’s perfected the art,” says Sahay. “But that model is now available. If somebody has created a good one, it will be replicated soon.”

Porn and politics will see the immediate impact, for obvious reasons, Sahay confirms. The former embraces technology faster than any industry. AI has been in porn much before journalists woke up to an AI anchor on Chinese television. “For a deepfake to happen, you have to remember you need enough snapshots of the person in social media or out in the public domain. You can’t run the model without having multiple images of the person in question,” Sahay explains. Celebs, therefore, make the cut easily. The average wo/man will be less affected because the algorithm doesn’t have enough data in the form of available images to learn from and reproduce. “But once it reaches a certain level, it won’t be easy to figure

out: it will almost be seamless. Obviously, technology will be able to develop a forensic understanding of whether a video is fake, but for normal people, whatever damage has to happen will happen by the time forensic comes," he says.
In the Buzzfeed 'Obama video' from 2018, the former US president apparently says, presciently: "We're entering an era where our enemies can make it look like anyone is saying anything at any point in time, even if they would never say those things. So, for instance, they could have me say things like...how about this? Simply, President Trump is a total and complete dipshit." Forty seconds in, the video is split-screened, with comedian Jordan Peele saying the exact same lines—a cute caution. Less than a year earlier, scientists at the University of Washington also worked on an Obama video and managed "turning audio clips into a realistic, lip-synced video of the person speaking those words."

A deepfake of Trump and Putin as Mini Me and Dr Evil from Austin Powers

Sahay says if you record for about 40 minutes, "I have all your syllables recorded. Now I simply have to give a desired input, and because I have your voice sample, I can create it with the same tonality; that same line will be spoken." Frankly, the hassle of putting a gun to someone's temple and asking them to repeat what you want is passe. Audio, video: technology has it all covered.
In politics, it's beginning to happen. Last year, the President of Gabon, Ali Bongo, travelled abroad for treatment. People close to him assured the media that he would make a New Year's address. Duly, a video of Bongo addressing the western African nation was released. The opposition was having none of it. The US magazine *Mother Jones* reported that cries of 'deepfake' by Bongo's primary rival

Bruno Ben Moubamba led to Gabon's first attempted military coup in 55 years. It failed, but there was enough chaos. Experts in the West are still unsure if deepfakes were really a cause for a near-regime change. One could argue, though, that's precisely the purpose: enough confusion for the disbeliever and ammunition for the believer.

"We access the internet via dodgy websites and WhatsApp groups. The problem will be acute in smaller towns and cities where if, say, a politician wants to target a particular community by putting out a video and instigate something. You believe the video because it suits your bias," says Sagar Kaul, CEO of Metafact, a fact-checker. A deepfake won't be broadcast as a sermon; that makes it easier to spot the fake.

Experts see developing economies more at risk. In India, a spate of lynchings has happened due to rumours spread on WhatsApp. While junior home minister G. Kishan Reddy denied any "common pattern" to them while speaking in the Rajya Sabha on July 25, a hint lies in the fact that WhatsApp has assured of "prompt action" on the traceablity of its messages. The same day, NITI Aayog CEO Amitabh Kant said WhatsApp has close to 400 million users in India, about 80 per cent of the entire smartphone population. A fake video on an end-to-end encrypted platform is a scary thought.

Especially if its disguise is getting better. In December, Metafact's Kaul had told *Outlook* that their AI tools could debunk deepfakes. He had mentioned that a deepfake blinked a lot lesser than the average human, a clear sign. However, the technology has since improved. "We had some success initially, but the second generation took the eye blinks away," Kaul now confirms. "With the second generation, the tool can create artificial movements. We initially saw this only for big people because there's a lot of data on them out there and you could train the AI," he says. For detecting a deepfake, however, one has to go frame-by-frame—"an intensive job, though the creation itself is cheap."

Frames are where the game is. Several split-seconds of video spliced with the intended audio can trigger most folks at a time when vines and memes are a thing. In the US, the deepfake furore caught steam primarily because Trump has sent out campaign feelers for 2020. On May 24, he tweeted a video with the caption: "PELOSI STAMMERS THROUGH NEWS CONFERENCE". The video, heavily edited and put on air by a channel from the Fox stable, sought to establish that the House of Representatives speaker was inebriated. *The Daily Beast* tracked down its origin: a Facebook video posted by 34-year-old Trump fan David Brooks with the caption: "Is Pelosi drunk?" The video was found to be doctored by simply slowing down the audio "without lowering the pitch of her voice." Yes, not quite a deepfake, just carefully doctored. Experts called it a 'cheapfake'.

The dangers are the same, says Tarun Wadhwa, entrepreneur and a visiting instructor at Carengie Mellon University. He also mentions "voice mimicry technologies" that could be used to confuse, extort and scam people. Wadhwa says he's less concerned about a deepfake starting a war—a few hits apart, he trusts the media to take over and do its job. "However, on a personal level, I'm very worried forgery technologies are going to ruin relationships and reputations. I think about the lynchings in India and elsewhere on WhatsApp, where a rumour about a migrant in town would spread with some grainy footage, and that false information would lead to mob justice. I'm worried about what we as individuals will do to each other when we have this at our disposal," he says. "My point is that in this sort of context, where it's one-to-one, or one-to-many, who's checking to see if the video is real?"

AltNews has had its hands full with a myriad forgeries since its inception short of two years ago. Its co-founder Pratik Sinha tells *Outlook* he hasn't spotted a credible

deepfake as yet, so is not “worried” at the moment. Sinha agrees the tech will become problematic in the future, but cites a story the fact-checking website debunked on July 22. A video posted on Facebook seemed to suggest the Muslims of Agra were up in arms, sloganeering against PM Modi, the Bajrang Dal and the Shiv Sena over the mob-lynching of Tabrez Ansari in Jharkhand. The video was from a Moharram procession in Bihar in 2014, and the superimposed audio from a 2017 protest in Udaipur, in response to the infamous murder-on-video by Shambu Lal Regar. Sinha says such videos are capable of far more damage. And forged audio and obfuscation by confusion have been around for a bit. Who said ‘Bharat tere tukde tukde honge?’ Was it even said?

The AltNews experience shows India may still be embarking on the deep learning curve, but cyber law specialist Pavan Duggal says he’s already had clients asking for recourse after being subject to deepfakes. “No specific IPC provisions exist now, but Sections 468 and 469 (relating to forgery) could be applied,” he adds. “There’s a catch, though. At the end of the day, there’s really no falsity vis-a-vis data inputs, no fakeness of that sort. A deepfake takes a couple or more sets of data inputs and merges them.” The law should evolve, adds Duggal, because the existing sections don’t cut the mustard and the “propensity of the general user to be persuaded by a deepfake is higher”.

Apar Gupta, executive director, Internet Freedom Foundation, believes different IPC provisions can come in. “The video content itself will attract different legal provisions. It could be used for satire and other purposes, so we need to go case by case,” he says. “But deepfakes can act as a persuasive tool even for relatively educated people. We’ve not been trained to spot if a video is fake.”

As with most technology, researchers are constantly trying different ways to get the cat back into the bag. Another defeatist argument is that “people believe what they want to believe anyway” and that deepfakes are just another form of content. “That’s a dangerous slope,” says Wadhwa. “While everyone is free to have their opinion, when we no longer have a shared basis of information we agree upon, it becomes eventually impossible to have a discourse.” On the other hand, isn’t that true already?