METAL

№ 43, SPRING/SUMMER 2020

Arca, Andrew Thomas Huang, Stelarc
Lucy Mcrae, Wolfgang Beltracchi, Henry Ajder
Charlie Parsons, Doug Aitken
Juliana Huxtable

This might be a bit unconventional for a magazine, but please allow me to take you back to a Subreddit in 2017. Because it was then and there an unsettling phenomenon emerged. Visitors of the Subreddit could enjoy pornographic videos featuring their favourite female celebrities, but the celebrities in question had never acted in such a video – or knew of its existence for that matter. Created through artificial intelligence that face-swapped celebrities' faces with those of porn actors, the videos were eerily realistic. The originator was a Reddit user going by the name of Deepfakes.

HENRY AJDER

Goodbye to Innocence

WORDS BY MARJOLIJN OOSTERMEIJER – PHOTOGRAPHY BY LAURA MARTINOVA

Flash forward to nearly three years later, when the term 'deepfakes' has long outgrown its origin in a Subreddit. Today, the term describes AI generated synthetic media that – when done well – could pass for reality. This contemporary form of technology assists in the creation of compelling art and entertainment, but can also be weaponised, aiding in deception, fraud or sexual humiliation. Deepfakes haven't only allowed us to alter reality, they've made us question it altogether. Because, how can we spot reality in a digital landscape filled with potential landmines of synthetic media?

Seeking an answer to that question is Deeptracelabs, a tech studio dedicated to detecting deepfakes and guiding people through the increasingly tricky digital media landscape. To understand this space, we reached out to their head of threat intelligence, Henry Ajder. After conducting extensive research, Henry has obtained an aerial perspective on the deepfake landscape, allowing him to disentangle myth from fact. We discussed the power of both the technology and the term as we met, perhaps somewhat ironically, over a digital video call.

Marjolijn Oostermeijer: Henry, one can approach the issue of deepfakes from many perspectives; technological, political, philosophical and so on. What's your perspective?

Henry Ajder: What I'm trying to do is map the deepfake landscape as it emerges. My research in Deeptracelabs is aimed at understanding how deepfakes are manifesting in different forms, which might be comedic, pornographic, malicious, positive… Anything. My job is to help people understand what's going on without imparting my value judgements, or speculative anticipations of what might happen.

MO: And what about Deeptracelabs, how do they approach the issue of deepfakes?

HA: We're trying to be as broad as possible. We don't want to put a spin on the issue but instead we try to build technological solutions for certain problems deepfakes might create. Our focus, however, is on countering the malicious uses of the technology.

MO: In preparation for this interview, I attempted to create a deepfake.

HA: Really? How did that go?

MO: I used online technology which resulted in a mediocre face swap, but it did illustrate how accessible the technology is for those with more time on their hands.

HA: There are many tools out there which aid in the accessibility and commodity of deepfakes. These tools do some of the processing and complicated work for you but, as you've probably noticed, the results aren't fantastic. Some people believe anyone can make a deepfake, which is true, but the quality is going to be terrible. There are few good deepfakes out there and many of them have some form of post-production processing. So, it's important to distinguish between creating a deepfake, and creating a good deepfake.

MO: That's a good point. I was also surprised by how endless the opportunities are. What are some of the forms in which you've seen deepfakes appear?

HA: In September last year, we released a report mapping the deepfake landscape. This defined that, by far, the majority of deepfakes are pornographic. 96% at the time of that finding, which hasn't changed much today. However, as the technology is maturing, we're seeing an increase of people engaging with it for other purposes; entertainment, comedy, art and more. Some creatives who are making 'safe for work' deepfakes are doing very notable work, partnering with special effects studios, advertising agencies or TV shows in the process. However, it might be interesting to further define the term 'deepfake'. It's a nuanced yet important piece of language which runs the risk of losing its meaning when used carelessly. Deepfakes emerged around November 2017 on a Subreddit where a user named deepfakes started to work with an open-source library and an autocoder. In its initial understanding, deepfake exclusively meant face-swapping people's faces into pornography. Today, the term has expanded to include voice synthesis, facial reenactment, body pose transfer and so much more. You can find highly realistic images of people – or even Airbnb's – who don't exist, you have GPG2 and text generating modules which are arguably also deepfakes. But they differ from realistic photoshops or special effects because they're created by deep learning. In scientific terms, it means AI-generated synthetic media, which encompasses text, video, audio and image. It's still up for debate if the term should only define malicious uses of the technology, or all of them. Although deepfake already has quite a negative connotation.

MO: Malicious deepfakes can be used for political, business and personal attacks – amongst other things. Which group does Deeptracelabs mainly aim to protect and why?

HA: All of them, to be honest. A lot of the damage deepfakes cause can be seen as analogies of computer viruses. They infect the human mind, causing malfunctions, harm or extortions. If you'd receive an email from an unknown address containing a suspicious-looking file, you'd use your antivirus scanner, or not open it at all. Why in the age where synthetic media is becoming increasingly realistic and vast, would you trust images and videos on social media? Our approach is placing a safety layer into platforms people are already using. In terms of how it works, we're developing what's called an AGI, which protects platforms and their users against deceptive synthetic information by providing a safety layer in their existing information pipeline. Let's say you're a social media platform moderating videos. The system would detect deepfake videos, after which it's up to the platform to do something with them.

MO: When I got my first computer with internet access, my parents were very adamant on teaching me potential deceptiveness of online information, a lesson practically every Internet-using person has been taught. So why is it that we still blindly seem to trust digital video, image and even words?

HA: I think when deepfakes confirm certain cognitive biases or are on platforms we trust, even if we shouldn't, the content, context and semantic influence of that video plays a bigger role in making us believe it than the video itself. We're less rational than we think, not to mention we live in an age of information overload, making us greatly trust media organisations. It speaks to the nature of this age that we're often without a guide, in uncontrolled digital environments that make us default to what we know and how we feel safest. But let's not forget the realism of synthetic media is also getting better. A good illustration is a website called Which Face Is Real. It's a game where players have to decide between two faces. One is real and one is generated. They have a 58% success rate, only 8% better than chance, even though they know one of the faces is fake.

MO: Speaking of the age of information overload, nowadays that landscape is so saturated yet filled with echo chambers. From online newspapers to Subreddits, whatever our opinion, we'll find – and believe – aligning information. What role do deepfakes play in reinforcing and potentially radicalising our viewpoints?

HA: I think these spaces, echo chambers or information silos, exist to an extent. I've seen it in certain political groups and spaces on Reddit; online spaces where you can insulate yourself against any voices differentiating from your quite rigid one. Deepfakes play the role you'd expect. Imagine a video of someone doing something they haven't, yet posted in a group who thinks they have. This group will be much more susceptible to believing it because those doubting it just aren't in that space. This comes back to an interesting question: How much more visceral and believable are videos than text or a crudely photoshopped photo, for that matter? I believe there's some difference but, in a way, the framework is the same; deceptive content that looks believable conforming to what people want to see. Of course, you can also question the impact of echo chambers, but deepfakes certainly play a role in enhancing them, yet rarely force things to escalate.

MO: Oppositely, can people claim truthful information is a deepfake to escape accountability for their actions?

HA: Yeah, absolutely. One of the key findings of our report was that good deepfakes – at least for the moment – are fairly contained. Yet, they've muddied the water, not only allowing people to say fake things but also plausibly denying real things are real. There've been a few cases, particularly in politics, where people have used the concept of deepfakes either to smear someone or potentially deny the realness of a video. The concept of deepfakes alone is enough to destabilise processes and, for the moment, it'll be weaponised more than the phenomenon itself. Again, when talking about the context of a deepfake, if we saw a shocking video 10 years ago, our first thought wouldn't be "Oh, that must be AI-generated synthetic media". Once a doubt is planted in our minds, the innocence with which we view audiovisual media disappears. But to be honest, that should have disappeared a long time ago.

MO: Despite a fear of political or economic implications, you mentioned that the majority of deepfakes are pornographic. Why is it that this form of technology is so frequently used to sexually degrade women?

HA: I suppose pornography has always been a space where tech innovation is applied. If you think about it, people have been writing erotica or drawing erotic images for a long time, and fantasising about celebrities too. So now that the technology is available, it's a natural extension of the pre-existing objectification of celebrities, especially women, as you've said. What concerns me is that the creators of this type of pornography don't see what they're doing as wrong, or they simply don't care. They see it as fun, like, "If this were happening to me I wouldn't care." I don't think that's true.

MO: There's a double standard as well. If a woman is featured in pornography, she's scrutinised much more than a man.

HA: Right, and functionally all of the people featured in the pornography were women. We found a few (less than 20) men in gay pornography, which can be potentially distressing and dangerous if you're coming from a country where homosexuality is a crime. I think it's because female celebrities, in particular, are viewed as public assets. They've stopped being people and have become a commodity to be shared and traded. What concerns me even more is the private individuals featured in deepfake pornography, meant as revenge porn or a form of bullying. Unlike celebrities, these women lack the resources to have videos featuring them taken down, which we've also seen with real revenge porn. I've seen a particular kind of deepfake being used extensively on private individuals who are very likely unaware that their image is being used in this way and shared publicly. This could become a prevalent form of digital sexual violence against women.

MO: When talking about the context of a video, I can imagine if we saw a pornographic video featuring a celebrity, we could probably guess it's not real. Whereas a similar video featuring a private individual wouldn't strike us as questionable.

HA: That's a great point. Additionally, if you saw a compromising video of a private individual you did know you might wonder, why would anyone want to target them? And, therefore, assume the video is real. It speaks to the humble way in which we view the private individual as opposed to the less humble way in which we view the celebrity. Something else I wanted to add is that these videos emerged on pre-existing toxic online communities such as 4chan and Reddit. They're now moving towards independent platforms, but there's a reason they emerged on these less regulated spaces.

MO: It's interesting you bring this up because I feel like these particular unregulated and unreal digital environments are very toxic, very objectifying towards women. Do you think these spaces and the deepfakes created there influence or amplify the collective perception of women in real life?

HA: I wonder, is it a symptom or the cause? Is the creation of pornographic deepfakes a result of the way our patriarchal society already views women – as objects for amusement and pleasure? Or, is deepfake pornography causing more people to behave that way? I feel like deepfake pornography is a symptom of a deeper issue in society. This is illustrated in the spaces where deepfake pornography originates from, they're almost entirely male-dominated toxic subcultures that don't see what they're creating as wrong. So, I don't think this will necessarily change society's perception of women but, if it becomes a common phenomenon, it'll add another toolkit for people attacking women based on their gender. Another distinction between pornographic deepfakes and various malicious deepfakes is the realism of the videos. Because when it comes to pornography, does realism matter much? If the likeness is close enough, it'll bring the desired satisfaction to the viewer. So, deepfake pornography doesn't have to fool anyone that it's real for it to still do a lot of damage to the victim.

This also results in more difficult and controversial questions. There's a paper called the Pervert's Dilemma which came out as a response to deepfake pornography. It explores the question, what is distinct, ethically, about the process of fantasising and creating a private deepfake porn which is not publicly released? Of course, this has no impact on public deepfakes, but questions like these make us interrogate the way we look at other practices. Deepfakes make us wonder, why did we ever think that was okay? Or, oppositely, we accept this, so maybe there are some positive uses.

MO: Coming back the realism, the Deeptracelabs report mentioned quite a significant viewership on websites dedicated to deepfake pornography, meaning the viewers know what they're watching is not real. What does it say about the way in which we view pornographic media?

HA: I think a lot of the people who've created pornographic deepfakes have fantasised about the video they're making. They know it might not be real, but it's close enough to be mapped onto a possible world. These platforms have quite a passionate community, requesting videos they want to see next, voting and even paying for their favourites. This illustrates that realism, or being fooled, isn't the most important factor in pornography. The video itself isn't doing all the work, a lot of it is happening inside the mind of the viewer. Likewise, there are sex workers who make their career by looking like a celebrity. People know it's not them but it's close enough to let their mind fill in the gaps.

MO: Of course, not all deepfakes are as malicious as these. Yet, I can imagine it's hard to find a balance between art and cyber weapon. How does Deeptracelabs determine if a deepfake is positive or malicious?

HA: We've thought about it a lot and, as a company, we're not here to make normative value judgements. Ultimately, our job isn't to tell you what to do with the deepfakes. We just detect them and allow you to make an informed decision based on your company's policies, national and international laws and so on.

Yet, the question of how one judges the nature of a deepfake is an interesting one. It's still being decided on by society at large. There are some very intuitive notions: if it's non-consensual, purposely deceptive or fraudulent it's probably malicious. But, of course, it's not always as simple as that because, what about the recklessly released ones? Can you even see the intention behind a deepfake? Probably sometimes, but not always.

MO: Now that we've faced the scary (un)reality of deepfakes, what can we as individuals do to protect ourselves against maliciously fabricated information?

HA: Simply put, it's not fair to assume the individual media consumers have the responsibility to verify everything they see. Just think about the massive number of images you see every day, and the amount of work it'd cost to verify them all. On that basis, we find it important to engage with platforms and people hosting content, so those consuming it are seeing what's already been checked. This can also provide people with more trust in what they're seeing.

In terms of what individuals can do, a healthy sense of skepticism is crucial. Be critical about the media you consume, don't idealise a certain source too much and back up everything you read with at least one other source. In terms of visuals, you can use a reversed image search to see where something has been posted before. It's dangerous to give more detailed guidelines on how to detect a deepfake, as the technology is changing all the time. In the long term, it's unsustainable to put the responsibility on the individual and we should rather focus on educating people to approach media with an investigative mindset. That being said, there's no way to discern every fake from reality.