Unmasking Deep Fakes: A Candid Exploration of Our Synthetic Society

INTERVIEW WITH: PROFESSOR HANY FARID

Dr. Hany Farid is a professor at the University of California, Berkeley's School of Information. Dr. Farid earned his Ph.D. in computer science from the University of Pennsylvania, and subsequently completed a postdoctoral fellowship in brain and cognitive sciences at MIT. He served on the faculty at Dartmouth College until 2019 when he arrived at Berkeley.

Dr. Farid's research spans forensic science, image analysis, and the detection of deep fakes and misinformation. He contributes to the Berkeley Artificial Intelligence Lab, the Center for Innovation in Vision and Optics, and the Vision Science program. Recognized for his expertise in image analysis and human perception, Dr. Farid acquired the prestigious Alfred P. Sloan Fellowship, John Simon Guggenheim Fellowship, and National Academy of Inventors Fellowship. He is the author of two books, *Fake Photos* and *Photo Forensics*, and consults for political agencies and news organizations.

BY: ERICA PAN, CALI BOND, ELLA KAUFMAN & ANDREW DELANEY

BSJ: You have a background working in image analysis, human perception, and cognitive science. How did this inspire and evolve into your current research interest in synthetic media?

HF: What I love about my introduction to synthetic media is how it demonstrates the randomness of science. In some ways, my background led me here, but it was not the defining factor.

One day, in the late 90's when I was a postdoc, I went to the library to get a book and I was waiting in the return line. There was a book called The Federal Rules of Evidence sitting on the return cart, which has absolutely nothing to do with me since I am not a lawyer or a legal scholar. I open it to a random page and the title reads: "Introducing photographs into a court of law." I was studying imaging, so I was interested in photographs. It described this new format called "Digital" that is emerging, and from the perspective of the federal courts, a digital image will be treated the same as a physical film. I thought that seemed strange because digital is inherently malleable. Even in the late 90s, it was malleable, let alone today. I thought, "This is going to be a problem someday." I went back to my office, but something just kept nagging at me. What are we going to do when digital comes?

Fast forward some years later, I am now an assistant professor at Dartmouth College. I had a colleague who I played tennis with every weekend. As a joke I photoshopped my friend's face into a photo of a tennis pro. As I did that, I thought, "I just introduced this interesting artifact in the image that I should be able to detect." I started writing some code and I was able to detect the manipulation that I had made—that was the beginning. Very little of that story has anything to do with my training other than that I read that page in the book because I was interested in images. But, this is sort of how science works. It is serendipitous and it is random.

Now around 1999, the internet was picking up and digital was coming and you could begin to see that there were going to be some problems. This motivated the development of our image analysis tools. A lot of the techniques we developed leveraged understanding how images are formed and how to analyze images. I have also been trained as a cognitive neuroscientist, so we would do perceptual studies where we would show people images and ask, "Do you notice these things that are right or wrong?" Then we would use that to develop an algorithm that improved over time. But how I got there was completely and utterly random and bizarre. If it were not for that book, I probably would have been doing something else.

There are all these great stories about the history of how science is so random like this, but I like telling the story to students because it really is. So, I always try and tell students, "Do not try to micromanage your career." You cannot imagine what the world is going to look like 10 years, let alone five years from now. So, my best advice is to learn as much as you can and be curious."

BSJ: What change, good or bad, do you think will be normalized in the entertainment industry due to deep fake technology in the next decade?

HF: I think deep fake technology is going to shake up the entertainment industry in a couple of ways. The writers' strike last summer was significant for many writers and performers. I think existing performers need to figure out how to protect their material and avoid being cloned. Future performers may be out of a job. Online influencers are already AI-based, meaning the entertainment world is going to fundamentally change.

I also think that the studios who have been advocating for the use of AI should be careful. I think that in the next decade not only performers but also studios will be out of a job. Pretty soon, we will have easy access to text to video. I will be able to say, "Give me a three-minute clip of two people walking down the street, having this conversation." I will be able to cobble together video clips to become a movie maker in my own home. Not only will I be able to make feature-style movies, but I will be able to distribute them across social media. So, what do you need a multi-million dollar studio for? I think we will see the democratization of access to a
Hollywood-like studios. The next Steven Spielberg is going to be people like you. This will be a big disruption.

We are already seeing it in your generation. TikTok is your TV. I think the way we consume entertainment is very much lined up with the ability to create shorts. I believe that we will see these advancements in your lifetime.

BSJ: How will media companies cope with large industry disruptions due to synthetic media? Will industries like entertainment be able to harness detection techniques to their advantage?

HF: There are two categories for detection, proactive and reactive. We should first acknowledge that not all generative AI is bad. There are nasty things, like nonconsensual sexual imagery, child abuse imagery, fraud, disinformation scams, election interference, and more. The proactive techniques include authenticating media upon recording. At the point of recording, a camera records the location, date, and time of recording, and authenticates the content of the recording, all of which is cryptographically signed. This information can then be stored on a blockchain. This technology is highly reliable and will eventually be widely available, but this will probably take a decade because hardware cycles are very slow. On the AI side, there are a number of efforts to bring companies together to apply front-end verification like watermarks and fingerprints so that the content is more easily identifiable downstream. These techniques are great because they are engaged at the point of creation. But, it is an adversarial system, so the bad guy is going to figure out how to attack these systems.

The reactive techniques are my bread and butter. We wait for something to go viral online, it is sent to us, we analyze it, and then we can tell you if it is real or fake. The problem with this approach is that the half-life of a social media post is measured in minutes—half of all views of a piece of content happen in the minutes after upload, so setting the record straight post-hoc is difficult. Will proactive and reactive techniques improve? Yes. Will the ability to create convincing deep fakes get better? Certainly. But, we are equipped with a model and so an adversarial game continues where each side builds better defense and offense, as we have done with spam and malware.
Along the way, we will be able to stop creators from producing material like Biden's robocalls in New Hampshire. However, a sophisticated actor wanting to attack our democracy can still do it, because they have a lot of resources. Russia wanting to attack our democracy is a different problem because we need to deal with that on the national security level. We first need to stop the average content creator from producing harmful material, and then continually raise the bar.

BSJ: You mentioned in your research that there is a difference between audio detection, and then image and video detection. How do you think that gap is going to be bridged?

HF: Video probably will be the easiest for a while, because you have an audio track, and you have about 24 to 30 frames for every second video that is analyzed. So, give me a three-minute video and I have a lot of data. But, if you upload some low-quality image, maybe 100 by 100 pixels, detecting AI generation becomes harder because you have less data. So, it is easier to hide your traces of manipulation. The video will always be the easiest, and audio will be something in between. This is just because it has to do with the sheer amount of data you have to produce.

Consider images of UFOs, Bigfoot, and aliens. They are always really low-resolution, blurry images. Why can nobody get a crisp

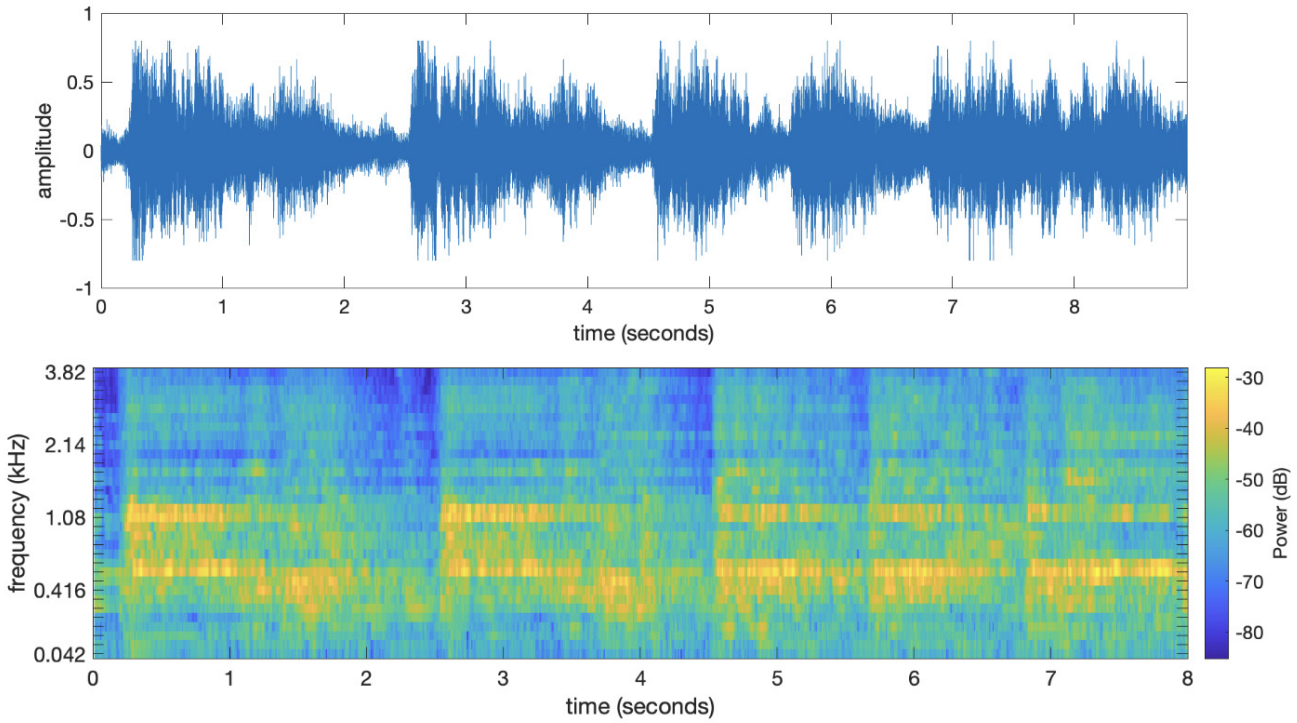

Figure 1. General-Purpose Voice Synthesis Reaches New Pitches: WaveNet pioneered the direct synthesis of raw audio signal waveforms (top). Raw waveforms result from a sampling density of symbolic representations of speech characteristics per second, necessitating conversion. WaveNet uses an autoregressive neural network to generate speech from previous samples (bottom).The results are networks conditioned on the characteristics of the speaker and highly realistic sounds.

image of Bigfoot, this makes no sense at all in this modern age. It is because when you create a fake image, you blur it and convert it to grayscale, add noise and conceal the evidence. So, first of all, if you are claiming you have a picture of Bigfoot I know you are lying. Second of all, low image quality is often indicative of manipulated images.

BSJ: How do you predict education is going to react to the increased accessibility of AI tools like Chat GPT?

HF: I am not sure, but a change is coming. We are seeing the shift already. People like to equate the use of AI to calculators. You know, when I was in school we had to learn arithmetic and long division. You probably had to also. At the time, it was somewhat unclear, like, "Why am I doing this? I have a calculator." However, I think it turns out, you actually do need it. I think knowing how computation works is important—whether you are going to do it or not I think is less important. I think something is coming in education, I think there are two things that I can see. One, the academic integrity side is going to be very messy. But that has been messy for a while, honestly, this is just another version of that. There has been an untold number of ways of cheating with technology. So that part does not worry me too much. I think what is more interesting is the idea of imagining a little AI bot that sits on your computer. As you are doing your homework, it analyzes what you do and says, "Oh, you are having trouble with this concept. Let us go watch this video." Or, "I see that you have learned this quickly, let us skip over this part and do this advanced topic." This would create a very customized way of learning, a kind of interactive, highly personalized tutor. I think that is very interesting.

How will faculty incorporate this technology into the classroom? I do not know. I am comfortable predicting that they will be very, very slow to do it. Because we typically are slow to incorporate things, and some of that is by design. When technology is moving this quickly, you cannot just start injecting it into the classroom today, because we do not know where it is going. We have to wait for the dust to settle, we have to figure out "What are the core principles we have to teach?" Then we can adapt.

In my world with computer science, we are used to this, we have been adapting for years. But we used to measure that adaptation in five-year cycles. So for example, as long as I have been on the faculty, we have had five-year reviews of our curriculum, which seems like a really long time. But you know, it is about right. In five years, you are like, "Alright, what is working, what is not working, and what has changed?" And then you make modifications. In the age of AI, it seems like every two weeks, we see something new. So, I think we are going to have to wait a few years to see where the dust settles.

What does this all mean? Ultimately, I think AI is going to shake up education. Frankly, I think AI is going to shake up everything. Every corner of the economy is going to end up being affected by this technology. Well, except plumbers.

I was talking to a bunch of students the other day and they asked "What jobs are safe?" I said, "Plumbers, electricians…mostly trades." I mean, right now we do not have robots and we will not for a long time. This surprised everybody. Now, I do not say this in a way to mock blue-collar jobs, in fact, quite the opposite. I think what startled so many about AI is that it seems to threaten intellectual and creative jobs, something we did not think computers could do, but we were all wrong.

BSJ: As far as detection techniques, which ones do you see improving the most in the next decade?

HF: I have no idea. I have been doing this for a long time, and I cannot even tell you what will improve a year from now, let alone a decade. I think everything will improve incrementally. I think that text will be the hardest to detect because it is a very nuanced thing. I think image, audio, and video detection will all steadily improve, I do not think one will be faster or slower than the other. I think they will all improve but it is impossible to know what that will look like ten years from now.

You will notice this in 10 years, but when I think back to when I started my career, it is unimaginable what is accessible and possible today. Have you heard of how if you put a frog on a stove and you incrementally turn up the heat it will not move because it can only measure differential heat? Well, something similar to this is happening right now. Every year we might think "Oh, this new thing is cool." Then, you look back 20 years later and suddenly think, "How did we get here?"

I think that is what is going to happen 10 years from now. We are all going to have implants in our brains and we will not even realize how far we have come. We will not carry phones, our means of computing will be in our head. I can tell you that every generation resists technology like this. The baby boomers hate smartphones, but

Figure 2. Fake? Faces: A generative adversarial network (GAN), StyleGAN, was used to generate synthetic faces of the real individuals (top) represented in the bottom row. GAN is considered to be the most common computational image synthesis technique. The system uses neural networks composed of generator and discriminator machinery to yield iteratively higher quality faces until the generator produces an image that is indistinguishable from real faces.

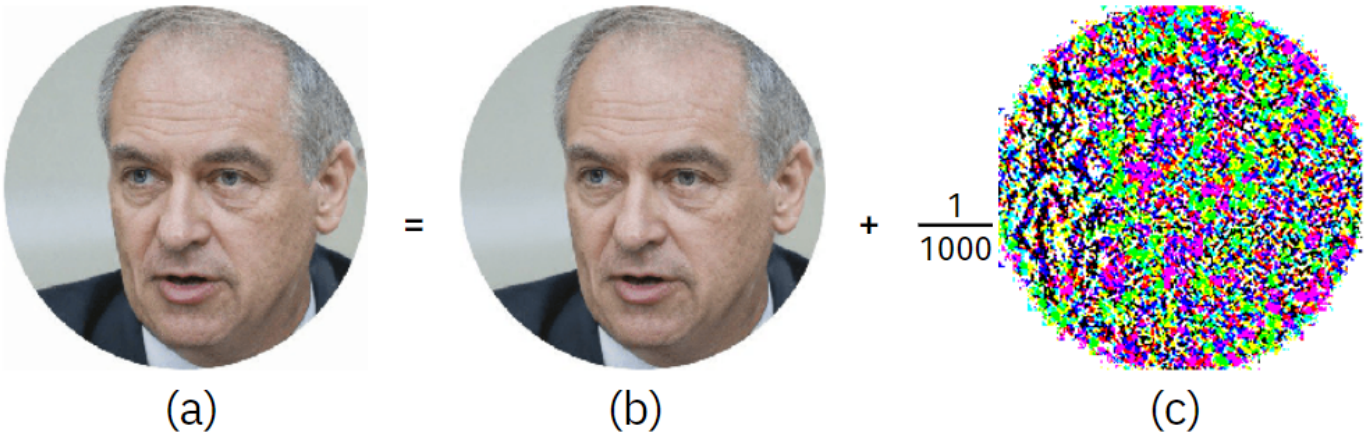

Figure 3. Learning to Unlearn: Learning-based methods of computationally detecting synthesized images can detect minor discriminatory visual artifacts, but suffer from vulnerability to attacks and slow adaptability to new synthesis techniques. Often, a learning network will inherit an image artifact that if removed, makes the classifier unfit for detection. Image (b) is the result of StyleGAN2's generative network. Image (a) was generated by a 17-year-old student with a novel image artifact (c), highlighting the scope of public accessibility to deep fake tools.

my generation adopted them fine, and your generation loves them. So, I think you will resist neural implants, but your kids will say, "Sure, wire it up."

I think, by the way, it is all getting pretty weird. I think we might all be like the frogs on the stovetop, and we are not really paying attention to how technology has infiltrated our lives.

BSJ: How do you anticipate your research regarding synthetic media evolving over the next few years? Have there been any recent developments that you wish to highlight?

HF: In one year, I do not think there will be much change. We will keep doing what we are doing. There is a great line in technology that we tend to overestimate what is going to happen in one year and underestimate what is going to happen in ten years.

Five years from now, I am going to retire, I am exhausted. I honestly do not know what it is going to look like. If you had asked me this question five years ago, I could have answered it. But right now, I feel like we are at this very steep part of a very weird time. And I cannot see out five years. We are going to just keep plugging along and trying to stop the world from collapsing because of generative AI.

BSJ: Within what time frame do you anticipate that deep fake technology will advance to a level where distinguishing between genuine and fabricated content becomes virtually impossible?

HF: So, are we going to get to the point where the average person living online will not be able to distinguish between real and fake content? In fact, in many cases, we are already there. We have already crossed the "uncanny valley," as we call it. Voice cloning audio is already advanced, and video will reach that level soon. Will we get to a point where AI can create what we call pixel-perfect fakes? I hope not, but I do not know.

What is the goal of generative AI? It is to make images that fool your eye, not my computational algorithms. An AI company is generally not incentivized to fool my detection system. Generative AI is designed with a different objective, to create media that looks visually compelling, even if it has measurable characteristics. I hope that the big AI companies are not interested in fooling me. They may not be interested in helping me, but they are not necessarily interested in actively getting in my way. I think we have a ways to go, years before that is true if it ever is. For instance, a video has an enormous amount of data, which makes it difficult for AI to recreate every single point with precision. So, I would be surprised if we have a pixel-perfect face that is impossible to detect with any of our computational tools in five years.

BSJ: You state that academic researchers should consider "not just how to do something, but if something should be done" when producing new technology. Can you discuss what the most critical safeguards are for creating novel technology, and where in the process of implementation should they come into play to ensure that ethical technology is being produced?

HF: There's a great line in the first Jurassic Park where Jeff Goldblum says, "Your scientists were so preoccupied with whether or not they could, they didn't stop to think if they should." This is similar to Oppenheimer where the physicists came out of it thinking, we spent so much time trying to figure out how to do this, we did not fully appreciate the consequences of it. I am not necessarily equating AI with the atomic bomb, although some days it may not be that far off.

Somebody asked me the other day: "Are you scared of AI?" I said, "No, I am scared of capitalism." The problem is not the technology per se, but that people can make trillions of dollars off of the technology. When capitalism takes hold of something, there is no stopping that train. You are pouring billions and billions of dollars into this, and you are trying to make trillions of dollars. That is what worries me.

When capitalism is unregulated, companies will do incredibly irresponsible things for profits. The oil and gas industry, for decades, has said climate change is not an issue because it would have interfered with their profits. The fact that we are burning the planet to the ground is irrelevant to them. The tobacco industry is the same way. For decades, they told everyone that smoking tobacco does not cause cancer.

When industries are left to their own devices, we often end up with problems. If we do not have really, really strong regulatory safeguards, both here in the US and everywhere else in the world,

if we do not train the next generation of young engineers to think ethically about what they are doing, we are going to have trouble. If we continue to have sociopaths running the tech companies, we are going to continue to have this problem. If we do not create real penalties for the harms that come from technology, we are going to continue to repeat the mistakes of the past. There is a path here where we could do this safely and thoughtfully that moves the technology forward but implements guardrails. But there is also a path without guardrails where the technology and the companies creating it become entirely impossible to control.

BSJ: Much of the public is scared by synthetic media, and we have spent a large portion of this interview discussing the risks that come with the production of synthetic media. However, what do you feel is the greatest benefit that synthetic media will bring to humanity over the next few decades?

HF: I am glad you asked the question. Of course, it is important to talk about the problematic things and tackle them head-on, but we should not pretend that there is nothing good. AI and technology's impact on medicine is amazing; its ability to read CT scans, X-rays, and MRIs, and diagnose breast cancer early on is incredible. I also love an idea I have seen on an advisory board for the American Bar Association. We have been having conversations about how Chat GPT can bring world-class legal thinking to people who cannot afford the $2,000-an-hour lawyers. Imagine you are arrested, you cannot afford an attorney, and you are getting some public defender who has 2,000 clients he/she must juggle representing. No problem, you can just ask ChatGPT for advice. Democratized access to high-quality legal defense is wonderful.

From a creator's point of view, putting aside the disruption to the economy, which is not insignificant, if I am a young creative person without musical talent, I can still create beautiful music. I can create beautiful movies and photos. There is no barrier to entry anymore. I do not need a $7,000 camera. I do not need a microphone. I do not need expensive equipment. We are creating this world where anybody can create beautiful things with the power of technology. Some things are happening that are interesting, but we have to think about disruptions to the economy and consider that all of these systems were trained on actual human beings' content, and most of them probably have not been compensated for that. There is a lot to be excited about here and there is a lot to be concerned about.

I take some comfort in the fact that we are having conversations now at the highest levels of government: the White House, the EU, the UK, Singapore, and Australia. Serious people are having serious conversations about what this all means. How do we encourage innovation but put some guardrails? I do not think anybody knows what to do yet, but that is okay because at least we are having the conversations.

Here is the thing about the internet: it does not know anything about borders. So, this is not really a nation-state governance. This needs a different way of thinking that we are not particularly prepared or equipped to do as human beings or nation-states.

We tend to look at the problems of the world—from climate change to social justice to the weaponization of technology—and think that it is overwhelming and there is nothing that can be done. I absolutely understand that feeling. But, you must understand two things. Number one, you cannot underestimate the power of a small number of people to change the world. If you look through history, almost every significant change begins with a small number of people. Number two, change comes in excruciatingly small steps. I know, for example, when you vote you may feel that it is useless. It is not, things like that are how change is affected. Change can happen as a consequence of very, very small steps.

References:

1. Bohacek, M., & Farid, H. (2023, December 27). The making of an AI news anchor—and its implications. PNAS.org. https://www.pnas.org/doi/10.1073/pnas.2315678121
2. Farid, H. (2022). Creating, using, misusing, and detecting deep fakes. Journal of Online Trust and Safety, 1(4). https://doi.org/10.54501/jots.v1i4.56
3. Figure 1: Farid, H. (2022). Creating, using, misusing, and detecting deep fakes. Journal of Online Trust and Safety, 1(4). https://doi.org/10.54501/jots.v1i4.56
4. Figure 2: Farid, H. (2022). Creating, using, misusing, and detecting deep fakes. Journal of Online Trust and Safety, 1(4). https://doi.org/10.54501/jots.v1i4.56
5. Figure 3: Farid, H. (2022). Creating, using, misusing, and detecting deep fakes. Journal of Online Trust and Safety, 1(4). https://doi.org/10.54501/jots.v1i4.56