REAL OR FAKE

Called "the father of digital forensics," **HANY FARID** walks us through Generative AI and the spread of maliciously manipulated content—and what business needs to worry about. Brunswick's **CHELSEA MAGNANT** reports.

TRAINED AS AN APPLIED MATHEMATICIAN and computer scientist, Hany Farid stumbled upon a book in the late '90s that would take him not just down an entirely new trail, but into the wilderness of an entirely new field: digital forensics. His story begins in a very old institution, the library. • "That's how you know it was a long time ago," Farid, now a professor at the University of California at Berkeley, tells us. "It seems so quaint now, that I would go and get a book at the library." • Having earned his Ph.D. in Computer Science at the University of Pennsylvania in 1997, he was doing post-doctoral work in brain research about human perception. Standing in line at the library to check out, he absently picked up a book laying nearby, *Federal Rules of Evidence*. • "I knew nothing about law, but I was bored," recalls Farid. "And it literally opened to a page that said, 'Introducing photographs into evidence in a court of law.' I read a little further. It was talking about what types of photographs courts should accept as authentic. And there was this new thing, digital

PHOTOGRAPH: COURTESY OF HANY FARID

photography: 'We will treat digital images as having the same authenticity as a 35mm negative.' And I thought, 'That seems like a bad idea. What happens when digital takes over—which we all knew would happen eventually—and images become easier to manipulate?'"

Two years later, in a break from teaching classes at Dartmouth where he had his first faculty post, he found himself making a goofy photoshopped image, superimposing a friend's face on a photo of Andre Agassi, the famous tennis pro.

"I had to make his head a little bigger to fit the body," Farid says. "And I realized, I've introduced a digital artifact into the image because I had to add some pixels."

Such an artifact, he realized, could be used as proof of manipulation. Together, those two insights propelled him into research later termed digital forensics—analyzing digital media to tell if and how it was manipulated. In the years since, manipulated images, audio and video have exploded into headline threats not only to individuals, business and corporations, but to public safety and fundamental assumptions of democratic society. Dealing with deepfakes in particular—where through video and digital tools, people's likenesses can be turned into realistic puppets—are currently a critical concern in security discussions.

But in the early 2000s, "nothing existed" in the way of tools or research, Farid says. "We just started writing papers and thinking about this. Nobody saw Generative AI coming at that point. Suddenly now, the ability to manipulate and synthesize media looks very different than it did 20 years ago. People say, 'oh you were so prescient!' But it started with me just screwing around in Photoshop."

Farid is now referred to as the "father of digital forensics" and regarded as the world's go-to expert on deepfakes and manipulated images. He regularly helps movie stars and other celebrities, as well as politicians, lawyers, law enforcement, journalists, and even the White House and the United Nations, in efforts to identify where deliberate digital alteration and fabrication have been used to create misinformation.

He is a Fellow of the National Academy of Inventors. *Enterprise* magazine said his first book, *Photo Forensics*, is "likely to become the bible of the field." Farid has also now founded GetReal Labs, a business to help organizations combat these threats.

Brunswick's Chelsea Magnant, a former student of Farid's at UC Berkeley, interviewed him for the *Brunswick Review*, where they discussed the fast-growing threat of misinformation and the arc of his professional life, a scientist ironically launched on his career path by a wink of serendipity—a stray book in the library.

"That's the way the world works," he says. "It's kind of beautiful and terrifying. A moment later, a moment earlier, I didn't look at that page and … I may have done something completely different."

"THE HALF-LIFE OF A SOCIAL MEDIA POST IS MEASURED IN ABOUT **60** SECONDS. … YOU DON'T HAVE HOURS OR DAYS. YOU HAVE MINUTES TO RESPOND."

Could you just describe your work on deepfakes and manipulated media?

The core of what we do, the technical part, is we build computational techniques that will ingest an image, audio or video, and try to determine if it has been manipulated, edited, fully AI synthesized. Our concern is, how do we authenticate media?

The applications for our work are courts of law, media outlets, Fortune 500 companies being attacked, regulatory bodies. In media, I don't think a single day has gone by in the last year where I have not had to talk to a reporter about something that is happening around the world or they're not sure if an image, audio or video has been manipulated or is AI-generated.

Can you talk a little bit about what you're seeing in the evolution of manipulated media?

Almost from the start of Generative AI and deepfakes, we saw the creation of nonconsensual sexual imagery. Taking the likenesses of mostly women, inserting them into explicit material, and then using that for extortion, weaponization, embarrassment, humiliation, whatever. It's not just celebrities; it's anybody with a single image of themselves online. You have an image on LinkedIn? I can take that image and now insert you into a video, using deepfake technology. The creation of child sexual abuse material—taking images of young children and putting them into sexually explicit material—that's also being done. Just awful, awful, awful.

Small-scale fraud, where people are starting to get phone calls from who they think is their loved one, like a phishing scam, but now it's a phone call in the voice of your son, daughter, granddaughter, mother, father. Large-scale fraud, institutions being separated from tens of millions of dollars because they are transferring money to an organization that is fraudulent.

In business hiring: People will interview for jobs on live video calls, only the applicant is not who they think it is, and then they've got a hacker inside their organization who's inserting malware—this has now happened many, many times.

Obviously, too, we are seeing disinformation and election interference on the rise with the Generative AI. I don't see any of these going away. I see them only getting worse.

But the big one is that when you live in a world where anything you read, see or hear can be fake, then nothing has to be real. Just last year, when Biden stepped away from the upcoming presidential election, there was a press conference with Vice President Harris. President Biden had been diagnosed with COVID, so he called in to the conference, spoke for four minutes, introducing the vice president. Some people said, "Oh, that was fake! He's actually dead." And a whole conspiracy emerged, including members of Congress calling for investigations—why? Because it doesn't have to be real.

Where are we, as a society, when you can't believe anything? This builds onto an erosion of trust already in the media, in the government, in scientific experts. It's a really dangerous world, a really weird world we're entering right now. Because of the larger infrastructure that we live in, it's not just that the bad guy can create fake information; they can carpet bomb the internet with it. All that content is now being amplified by the underlying social media algorithms.

Can we talk a little bit about best practices for businesses, keeping people safe in this environment?

If you're in the Fortune 1000, here are the things you need to worry about. This is going to happen: Somebody's going to create a video of your CEO kicking puppies down the street. And they're going to release it on Twitter [X] where it will get millions of views. It's going to be bad for you and really hard to combat. Once people see those videos, nobody unsees them. Or, somebody's going to create a fake earnings call of your CEO, saying your profits are down 5%, and the stock market is going to move billions of dollars before you can figure anything out. People in your organization are going to get phone calls from who they think is their CEO, or CTO, or CFO, saying, "Do this. Tell me this. Give me this information." We're seeing password reset attacks with voice cloning. Social engineering [manipulating people to make security mistakes] is real and deepfakes are going to supercharge those social engineering attacks on your organizations, both internally and externally.

So what do you do about it? Well, everything in this space is mitigation, not elimination strategies. If somebody wants to hurt you, they're going to hurt you. But you can mitigate that and you can minimize the damage.

If I'm the leader in an organization or government, every single piece of content that I release publicly should be digitally signed by me. My earnings calls, images, video interviews, so that when something comes out that purports to be an earnings call or a photo or a speech, if it's not digitally signed by me, it's not real. You immediately have reasonable proof this is probably fake. So that's number one.

Number two is, you need to do tabletop exercises. How are you going to respond? Who's your first call? Who's your second call? How do you get this stuff off Twitter and Facebook as fast as possible? How do you figure out who did it and hold them accountable? Because the one thing we know is that if they get away with it, it's going to happen again. Who in your organization is responsible? Is it the CISO (Chief Information Security Officer)? Is it public relations? As I talk to these organizations, often nobody really knows whose responsibility it is.

A press release has to go out. You've got literally minutes to respond to these things. The half-life of a social media post is measured in about 60 seconds. Half of all views happen in the first minute. You don't have hours or days. You have minutes to respond.

What's the best place for that function within an organization?

It's a great question. What we're seeing is the CISOs are owning this. And that seems about right to me. That's probably the right place for it. But it is not in their wheelhouse. This is not something they know about, most of them. So I spend a lot of time talking to CISOs, and CEOs for that matter.

You and I have discussed manipulated media in terms of the "uncanny valley." Can you talk about that?

The term came from robotics, building humanoid robots that physically interact with us. But it also is used for images, audio and video. Technology that looks like cartoons are funny and pleasurable to watch. But when it starts approaching human-like appearance, but not quite human, we become very uncomfortable with it. It feels weird, uncanny.

The quality is such that faked images of people have now passed through the uncanny valley. They are highly photorealistic and we don't find them weird or uncomfortable. People can't reliably say, when they look at images, whether it's a real person or not. Audio—just speech—is just about through

"THIS IS GOING TO HAPPEN: SOMEBODY'S GOING TO CREATE A VIDEO OF YOUR CEO KICKING PUPPIES DOWN THE STREET. AND THEY'RE GOING TO RELEASE IT ON TWITTER WHERE IT WILL GET MILLIONS OF VIEWS. ... OR, SOMEBODY'S GOING TO CREATE A FAKE EARNINGS CALL OF YOUR CEO, SAYING YOUR PROFITS ARE DOWN 5 PERCENT AND THE STOCK MARKET IS GOING TO MOVE BILLIONS OF DOLLARS BEFORE YOU CAN FIGURE ANYTHING OUT."

the uncanny valley. Chance yields 50% correct answers in judging which clips are fake and which are not. The results now are around 65%, so still slightly better than chance.

Video is a bit of a mixed bag. Think about the Will Smith eating spaghetti videos from a year ago that were hysterically bizarre—although these have improved quite a bit, these videos are not quite there, yet. Deepfake videos, however, where one person's face is superimposed onto a video character—those are very good, but not yet perfect. But six months, 12 months, 18 months—these things are all going to get better and better, cheaper, more accessible, and they're going to be used more.

Generative AI—and this is within a very short window, within a couple of years—has moved from, "This is terrible," to, "Holy crap. I can't tell the difference." And I think video is going to follow the same suit. It's just going to take, maybe, another year, year and a half.

What does that mean for all of the problems we just talked about?

It's going to get worse. The only hope is that the interventions start to keep up. Spam gets worse, virus gets worse, but the interventions get better. Everything escalates together, that's the hope.

But if you look at how much money is being poured from the venture capitalists here in Silicon Valley into the Generative AI side versus the defense side, the interventions side, it's orders of magnitude different. Generative AI has billions of dollars being poured into it. Companies like mine, millions of dollars are being poured into it.

Defense is hard. And it's less lucrative. So we are a little outgunned, in terms of the VCs, the talent, the academic literature. That's worrisome. I keep waiting for a rebalancing, but it's not rebalancing. So I think things are going to get harder.

But the hope is a combination of awareness, conversations like you and I are having right now, some regulatory pressure, and some good tech will start to mitigate some of the risks.

Where are we with the regulatory end?

In October of last year, President Biden released an executive order on all issues of AI, from Generative AI to predictive AI. There is now something being housed under the National Institute of Standards of Technology called the AI Safety Institute that is being tasked with these issues.

I spend a lot of time talking to folks on the Hill, and there is not a single branch of government that is not thinking about this: FTC, FCC, DOJ, NSA, CIA, FBI, the executive branch, the legislative branch. It's a little incoherent and inefficient, things need to start getting consolidated, but everybody is thinking about the sort of new world that we are entering, both AI and Generative AI.

"WHERE ARE WE, AS A SOCIETY, WHEN YOU CAN'T BELIEVE ANYTHING? ... IT'S A REALLY DANGEROUS WORLD, A REALLY WEIRD WORLD WE'RE ENTERING RIGHT NOW."

The White House has been working internationally, bringing in our allies from Australia, Canada, UK and the EU, to think about this holistically. Some 95% of the problem is outside the US. Tech itself tends to be very US-centric. There are entire parts of the world where nobody on content moderation teams speaks the native language, for instance—the very parts of the world that you need content moderation.

If you ask me, "What country is doing this best," it's Australia, across the board: AI regulation, social media regulation, monopolistic regulation. Julie Inman Grant is their eSafety commissioner, and she is a force of nature. I tell everybody, "Go, look and see what Australia's doing." The EU has been maybe a little overly aggressive on the regulatory side, but I like where they're going. The UK has an online safety bill. President Biden's executive order, which is of course non-binding, I think is all good.

Those are the four that I'm looking at right now as leading. And then, the state of California is doing a pretty good job. If any of the individual states is going to regulate, it should be California, because most of the tech and most of the VC money is here. I do like some of the language that is coming out of there, but it is being met with fierce, fierce opposition from the VC community.

It's scary to hear you talk about it. I'll admit that I can be a bit of a tech optimist.

Yes. I'm out a little bit on the other side of that. There are days where I think, "This was an interesting experiment, the internet. Let's shut it down." But I wouldn't be doing this work if I didn't have some hope. What we do is necessary, but it's not sufficient. We need people to care, people downstream from us, the tech companies—and upstream from us, which is the regulators.

Thank you. This has been a wonderful conversation.

Great seeing you, as always, Chelsea. ◆

CHELSEA MAGNANT is a Director in Brunswick's Washington, DC, office and leads the firm's AI Client Impact Unit. She previously worked with Google on tech policy strategy. She began her career with the CIA helping US senior policymakers navigate complex geopolitical issues.