MY WEIRD PROMPTS

Podcast Transcript

EPISODE #303

The Death of Seeing is Believing: Deepfakes in 2026

Published January 26, 2026 • Runtime: 20:23

https://myweirdprompts.com/episode/deepfakes-authenticity-digital-truth/

EPISODE SYNOPSIS

In this episode, Herman and Corn dive into the escalating crisis of deepfakes and the erosion of digital trust as we head into 2026. They respond to a listener's skepticism about the quality of AI-generated content by highlighting the "survivorship bias" of deepfakes—noting that the most effective deceptions are the ones we never realize are fake. The discussion covers the devastating real-world impacts of this technology, from $25 million corporate heists to the psychological toll of non-consensual imagery and the "liar's dividend," where the mere existence of AI allows bad actors to dismiss genuine evidence as fabrications. The hosts also break down the emerging technical solutions, such as Google's SynthID invisible watermarking and the C2PA standards being integrated directly into professional camera hardware. They argue that we are entering a paradigm shift where the burden of proof is moving from "detecting fakes" to "proving reality." However, this shift brings its own set of problems, including a potential "credibility gap" for those without access to high-end, verified hardware. Tune in to learn how to upgrade your "internal software" and navigate an era of epistemic nihilism where the very concept of shared evidence is under siege.

DANIEL'S PROMPT

Daniel

How big of a problem are deepfakes already in terms of deceiving people or causing reputational damage? Is invisible watermarking the future of proving authenticity? If platforms can already detect AI-generated content through tools like SynthID, what is the point of asking for voluntary disclosures, and why do we need both?

TRANSCRIPT

Corn

You know Herman, I was looking at some old family photos the other day, and I had this momentary flash of panic. I found myself squinting at a picture of us from ten years ago, wondering if the lighting looked a bit too perfect or if your smile was just a fraction of a second off. It is wild how quickly the baseline for trust has shifted.

Herman

Herman Poppleberry here, and Corn, that is the perfect way to kick things off. It is that creeping feeling that reality itself is becoming a bit... elective? Our housemate Daniel actually sent us a voice note about this earlier today. He was asking about the state of deepfakes right now, in early twenty twenty-six, and whether all these tools like invisible watermarking and voluntary disclosures are actually doing anything, or if we are just rearranging deck chairs on the Titanic.

Corn

It is a great prompt from Daniel because it hits on that skepticism a lot of people feel. He mentioned in his note that he hasn't seen many deepfakes that actually fooled him yet. He thinks they still look a bit mechanical or uncanny. But I think we need to address that head-on, because the gap between what people think they can detect and what is actually happening in the wild is getting dangerously wide.

Herman

Exactly. There is this massive survivorship bias at play. You only notice the bad deepfakes. The ones that are actually deceiving people, by definition, go unnoticed. And to Daniel's question about how big of a problem this already is... it is massive. We are well past the point of this just being a parlor trick or a funny meme. Reports indicate generative AI scams have increased significantly in recent years.

Corn

Right, and I want to dig into the reputational damage aspect he mentioned. It feels like we have seen two distinct waves of this. There is the high-profile celebrity stuff, which is awful, but then there is this much more insidious corporate and personal level of attack. Remember that case from a couple of years ago? The multinational firm in Hong Kong where an employee was tricked by a deepfake CFO voice into authorizing a transfer?

Herman

Oh, the twenty-five million dollar heist. That is the gold standard for deepfake consequences. For those who don't recall, the employee was tricked by a cloned voice of the CFO over the phone, leading to a $25 million transfer. Deepfake scams like this show the risks. That was back in 2024. Imagine where the tech is now, two years later, when you can buy scamming software on the dark web for twenty dollars that goes live in under two minutes.

Corn

And that is the thing. Daniel says they look mechanical to him, but if you are in a high-pressure situation, or you are looking at a low-resolution video on a phone screen, or you are just busy at work... that uncanny valley disappears pretty quickly. But Herman, let us talk about the reputational side. It is not just about stealing money. It is about the destruction of character.

Herman

This is where it gets really dark. We have seen a massive spike in what people call non-consensual deepfake imagery. It is being used for harassment in schools, for political character assassination, and for extortion. The problem is that even if you can prove it is fake later, the initial blow is often fatal to a person's career or mental health. We talked a bit about the emotional toll of digital life back in episode two ninety-six when we were discussing the rental crisis, and this feels like an extension of that. Research shows that victims experience significant psychological distress from deepfake harassment.

Corn

And it is not just the presence of fakes that is the problem. It is what researchers call the liar's dividend. This is something I have been thinking about a lot. If everyone knows that deepfakes exist, then anyone caught doing something real can just claim the footage is a deepfake. It provides a universal get-out-of-jail-free card for anyone in power.

Herman

That is such a crucial point, Corn. We are seeing it in courtrooms already. It is being called the deepfake defense. Attorneys are starting to challenge video evidence by claiming it could have been AI-generated, even when it is perfectly legitimate. It creates this fog of epistemic nihilism where nothing can be proven and nothing is certain. So, to Daniel's question: yes, it is a huge problem. It is an existential threat to the concept of shared evidence.

Corn

So let us move to the second part of his prompt. Is invisible watermarking the future of proving authenticity? He specifically mentioned SynthID, which is Google's tool. For those who aren't deep in the weeds like you are, Herman, can you explain what makes invisible watermarking different from just a little logo in the corner of a video?

Herman

This is where the tech gets really clever. Invisible watermarking, like SynthID or the standards being pushed by the Coalition for Content Provenance and Authenticity... which we usually just call C two P A... it doesn't live on the surface of the image. SynthID actually embeds the watermark into the frequency domain of the image data. It is like a digital fingerprint that is mathematically woven into the pixels themselves.

Corn

Robust in what sense? Like, if I take a screenshot or crop the image, does it stay there?

Herman

That is the goal. Because it is in the frequency domain, it is designed to survive being compressed, cropped, or having the colors adjusted. It is mathematically detectable even if the file format changes. It is like a DNA sequence that is woven into the fabric of the file.

Corn

Okay, but Daniel's question was: if platforms can already detect this stuff using these tools, why are we asking for voluntary disclosures? It seems redundant. If Instagram or YouTube can see the SynthID tag, why do they need the creator to check a box saying this was made with AI?

Herman

That is a very incisive question. And the answer is that efforts like California's proposed AI transparency measures and industry standards are pushing for both hidden watermarks and clear disclosures from large AI providers.

Corn

So the law is finally catching up. But why both? Why the belt and the suspenders?

Herman

Think of it like a layered security system. The disclosure is about establishing a social norm and a legal trail. If a creator lies and says something is real when it is fake, and the platform detects the watermark, they can penalize that creator for a policy violation. It is about intent. Plus, not all AI content has a watermark. We are in a cat-and-mouse game. There are plenty of open-source models that don't embed SynthID or any other signature. In those cases, the disclosure is the only line of defense.

Corn

That makes sense. But it also feels a bit like we are relying on the honor system with people who are, by definition, trying to deceive us. If I am a state-sponsored actor trying to influence an election, I am probably not going to check the made with AI box on my deepfake video.

Herman

You are absolutely right, and that is why we need the technical side too. But here is the thing that people often miss: the future of authenticity might not be about detecting fakes at all. It might be about proving what is real.

Corn

Wait, explain that. That sounds like a complete flip of the script.

Herman

Well, think about how we handled the mainframe security issues we talked about way back in episode one sixty-three. You don't just try to block every bad actor; you create a secure, verified path for the good ones. We are moving toward a world where cameras themselves... like the new models from Nikon, Canon, and Sony... have C two P A hardware built directly into the sensor. When you take a photo, the camera signs it with a digital certificate that says, this photo was taken at this time, at these coordinates, on this device, and it has not been altered.

Corn

So instead of looking for a watermark that says this is AI, we will be looking for a digital signature that says this is a raw capture from a physical lens?

Herman

Precisely. In a few years, if a video doesn't have a provenance trail that traces back to a physical camera, people will just assume it is AI-generated by default. The burden of proof is shifting. It is no longer innocent until proven fake; it is fake until proven authentic.

Corn

That is a massive shift in how we consume information. It reminds me of the AI preference problem we discussed in episode two fifty-four. We are essentially training ourselves to prefer verified, cryptographically signed reality because the unverified stuff is just too noisy and dangerous. But I wonder, Herman, does this create a new kind of digital divide? If you need a five-thousand-dollar Leica to prove your photo is real, what happens to the average person with a budget smartphone?

Herman

That is a huge concern. We are already seeing the early stages of a credibility gap. If you are a citizen journalist in a conflict zone and you are using a cheap phone without these hardware-level signing features, your footage might be dismissed as a deepfake by people who don't want to believe it. The tools meant to protect the truth could end up being used to suppress the voices of people who can't afford the tech to prove they are telling the truth.

Corn

It is the ultimate irony. We build the tools to fight deception, and the tools themselves become a new way to gatekeep reality. I want to go back to Daniel's point about the uncanny valley, though. He says he can still tell. I think he is being a bit optimistic, but there is some truth to the idea that our brains are very good at spotting weirdness in human movement. But what about audio? We just listened to Daniel's audio prompt. Herman, if I played you a clip of Daniel saying something completely different, do you think you could tell if it was him or a clone?

Herman

Honestly? Probably not. Audio deepfakes are much, much further along than video. The latency is almost zero now, and the emotional inflection is incredible. We saw those robocalls in the twenty twenty-four primaries that sounded exactly like Joe Biden telling people not to vote. That was two years ago. Today, you can clone a voice with about three seconds of high-quality sample audio. Daniel has been on this podcast enough that there is plenty of data out there to make a perfect Herman or Corn or Daniel clone.

Corn

That is a bit unsettling. It makes me think about our housemate dynamic. We could be sitting in different rooms in Jerusalem, messaging each other or leaving voice notes, and we could be talking to a ghost. It really highlights why that physical proximity matters. But let us get into the practical takeaways for our listeners. Because this can all feel very overwhelming and a bit hopeless. If the tech is getting better, and the fakes are everywhere, what do we actually do?

Herman

First, we have to upgrade our internal software. We need to move away from the idea that seeing is believing. If you see a video of a politician saying something completely out of character, or a celebrity endorsing a weird investment scheme, or your boss asking for an urgent wire transfer... your first instinct shouldn't be to react. It should be to verify through a second, independent channel.

Corn

Like, call them back on a known number?

Herman

Exactly. If your boss Slacks you a video saying pay this invoice, you pick up the phone and call them. Or better yet, talk to them in person. We need to revert to out-of-band verification. And for the general public, we need to start looking for those provenance markers. In twenty twenty-six, most major news organizations are using the C two P A standards. If you are on a news site, look for that little eye icon or the c r symbol in the corner of the image. That is the Content Credentials mark, and it will show you the history of the file.

Corn

And what about the platforms? Daniel was asking why we need both detection and disclosure. I think as listeners, we should be demanding more transparency from the platforms about how they are using these tools. If a platform detects a deepfake but doesn't label it because the creator didn't check the box, that is a failure on the platform's part.

Herman

I agree. We are seeing a lot of pushback against platforms that are being too lax. With ongoing AI transparency efforts, we are seeing those AI labels become more standardized for content generated by major models. If the system detects a SynthID watermark, the label is applied automatically. No questions asked.

Corn

It is a bit like the nutrition labels on food. You don't ask the manufacturer if they want to tell you how much sugar is in there; you make it a requirement for being on the shelf. But here is a thought experiment for you, Herman. What happens when the AI is good enough to watermark its own fakes with a fake authenticity certificate?

Herman

That is the nightmare scenario, Corn. It is the cryptographic equivalent of a forged passport. But that is why the hardware-level stuff is so important. It is much harder to forge a physical chip's private key than it is to forge a bit of software code. It is an arms race, and right now, the defense is finally starting to deploy some heavy artillery.

Corn

It feels like we are living through the end of the amateur era of the internet. For twenty years, we could just post whatever and assume people would take it at face value. Now, everything has to be credentialed. It is a much more formal, much more rigid way of interacting.

Herman

It is. It is the death of digital innocence. But maybe that is not entirely a bad thing. We have been pretty reckless with information for a long time. This forces us to be more intentional, more critical, and more connected to the physical world.

Corn

I like that. A return to the physical. Maybe that is why Daniel's prompt feels so urgent... we are all feeling that pull back to things we can actually touch and verify. Before we wrap up this section, I want to pivot back to something you mentioned earlier. You said you've been reading some new research on the psychological impact of deepfakes on the people who are targeted. Is it just the reputational damage, or is there more to it?

Herman

It is actually a form of digital trauma. Studies have shown that victims of deepfake harassment experience significant psychological distress, even when they know the content is fake. The brain has a hard time distinguishing between a visual violation and a physical one. Seeing a version of yourself doing things you never did creates a profound sense of dissociation. It is a violation of the self.

Corn

That is a heavy realization. It makes the stakes so much higher than just twenty-five million dollars or a lost election. It is about the integrity of our own lived experience. Herman, I think we have covered a lot of ground here, but I want to make sure we answer Daniel's question about the future. Is invisible watermarking the future?

Herman

It is a part of the future, but it is not a silver bullet. The future is a multi-layered ecosystem of trust. It is hardware-level signatures, it is invisible watermarks like SynthID, it is platform-level detection, and most importantly, it is a more skeptical and digitally literate public. We can't rely on any one of those things alone.

Corn

It is a team effort. Much like this podcast. And speaking of the podcast, we have been doing this for nearly three hundred episodes now. It is wild to think about how much the world has changed since we started. If you have been listening for a while, or even if you are new here, we really appreciate you spending your time with us.

Herman

We really do. And hey, if you are finding these deep dives helpful, it would mean a lot to us if you could leave a review on your podcast app or on Spotify. It genuinely helps other curious people find the show. We are trying to build a community of people who aren't afraid to ask the weird questions.

Corn

Exactly. And you can always find us at our website, myweirdprompts dot com. There is a contact form there if you want to send us a prompt like Daniel did, or you can just browse the archives and see where we have been. We are on Spotify as well, of course.

Herman

So, to wrap it up for Daniel... yes, the problem is real, it is already here, and it is causing damage. The tools we have are good, but they are just the beginning of a long struggle to reclaim reality. But as long as we keep asking these questions and looking under the hood, I think we have a chance.

Corn

I think so too. Herman, thanks for diving into the technical weeds with me today. I always feel a bit more grounded after we talk these things through, even when the topic is as slippery as deepfakes.

Herman

Any time, Corn. It is what we do.

Corn

Alright everyone, thanks for listening to My Weird Prompts. We will be back next week with another deep dive into whatever is on your minds.

Herman

Stay curious, and maybe keep a little healthy skepticism in your back pocket. Until next time!