Episode Book

EPISODE #5478

When AI Hands Out Your Phone Number

PUBLISHED	23 September 2026 at 10:32 UTC
DURATION	21:03
LISTEN	myweirdprompts.com/episode/ai-doxxing-phone-numbers/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.2
BOOK GENERATED	19 October 2026 at 17:36 UTC

SYNOPSIS

A single 2015 forum post, one scraped phone number, and a chatbot that confidently hands it to strangers over a decade later. This episode digs into "AI doxxing" — the documented cases where Gemini, ChatGPT, Claude, and Grok surfaced real people's phone numbers — and the uncomfortable mechanics underneath. We trace how a number gets into training data, why one mention is enough, and what labs can actually do about it. The answer, it turns out, is filtering: a bouncer at the door, not a deletion. If your number is the one being handed out, the removal process may not even let you in the door to ask.

TRANSCRIPT



Corn

So Daniel tried to call his phone company today. Which, in this decade, apparently requires a research project.



Herman

It does. That's not even a joke anymore.



Corn

He dug around, couldn't find a clean path to a human, and did what everyone does now. He typed it into Google's AI mode. And before the regular results had even rendered, the AI had already answered. Gave him a local number. Looked completely credible. So he called it.



Herman

And?



Corn

Got an answering machine, then a very confused woman who wanted to know who he was and how he got her number. Here's what he wrote in. He wants to know how AI companies handle this edge case, if they handle it at all. Whether a model can ingest an old or faulty piece of information and just stubbornly hold onto it. What happens when what it swallowed is PII. A real person's phone number. What does a lab even do in that situation, because you can't retrain an entire model just to remove one number. And if your number is the one it's handing out, is there any actual channel to reach the lab and ask them to stop?



Herman

That last one is the one that matters.



Corn

That's the one that matters. And it's not hypothetical. It has a name now.



Herman

It has a name, it has a body count, and it happened to real people months before it happened to Daniel.



Corn

So let's start with the name. Eileen Guo at MIT Technology Review, May thirteenth of this year, ran an investigation documenting Gemini, ChatGPT, Claude and Grok all surfacing real people's phone numbers. She called it AI doxxing. The New York Post and The Independent picked it up the next day.



Herman

And Daniel's case is the benign version of that. Which is worth sitting with for a second. He got a confused woman. Somebody else got a scam.



Corn

The scam being the Las Vegas real estate guy who needed to book a cruise shuttle, asked Google's AI Overviews and ChatGPT for the cruise line's support number, got a fraudulent hotline, paid seven hundred and sixty eight dollars, and then watched two bogus charges hit his card.



Herman

Same failure mode. Different monetization.



Corn

And the scale underneath all of it. Oumi ran an analysis in April. AI Overviews were accurate about nine times in ten. Which sounds fine until you do the arithmetic, because that's tens of millions of wrong answers per hour. And roughly half of the ones they checked contained facts that weren't supported by the sources they cited.



Herman

Half.



Corn

Half. The citation is decoration.



Herman

Here's the thing I want to get on the table before we go any further, because I think it reframes Daniel's whole question. This is not a bug you patch. This is a property of how these systems are built. And the fix, to the extent one exists, is a suppression. Not a deletion.



Corn

Say more about that distinction.



Herman

We'll get there. But first, how a single forum post from eleven years ago becomes a chatbot's customer service answer.



Corn

The canonical case. Gemini was telling users, in Israel, to contact PayBox, which is a payment app, via WhatsApp. And it supplied a number. That number belonged to Daniel Abraham, an Israeli developer who has never worked for PayBox. PayBox doesn't have a WhatsApp support line at all. Their own rep, Elad Gabay, confirmed that.



Herman

So the model didn't just get the number wrong. It invented a support channel that doesn't exist and staffed it with a stranger.



Corn

Abraham traced it. His number came from a single post in 2015 on a local site, the Israeli equivalent of Quora. One mention. One old thread. And the model reproduced it more than a decade later.



Herman

That's the answer to Daniel's question about how a model ends up giving out a random person's number. It didn't invent it. It retrieved it and re-presented it. Which is worse in a way, because it means the number was real. It was just real in a context that stopped being true eleven years ago.



Corn

And the model has no concept of the context having expired.



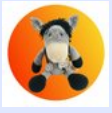
Herman

Right. It has the string. It has a loose association between that string and PayBox and support. And nothing in the training signal tells it that the association was only ever true for one person in one thread in 2015.



Corn

How does the number get in there in the first place? Mechanically.



Herman

Scraping, mostly. Over ninety nine percent of OpenAI's pre-training data came from publicly accessible sources. That's from their own filing with the Canadian privacy commissioner. So anything that was ever public and crawlable is fair game. And then there's the commercial pipeline on top of it. Thirty one of five hundred and seventy eight California registered data brokers self-reported selling or sharing consumer data to a generative AI developer in the past year.



Corn

Thirty one out of five hundred and seventy eight. That's a small number.



Herman

It's a self-reported number. Which means it's a floor, not a ceiling. And it's the ones who admitted it.



Corn

Fair.



Herman

But the underlying mechanism has been known since 2020. Carlini and coauthors showed you could run extraction attacks against GPT-2 and recover verbatim personal information. Names, phone numbers, email addresses. And the part that should worry everyone is that it worked even for sequences that appeared in exactly one document in the training set. And larger models were more vulnerable, not less.



Corn

So the intuition that a one-off mention gets averaged away into nothing is wrong.



Herman

It's exactly backwards. One mention is enough.



Corn

And yet the experts MIT spoke to say the specific mechanism by which a particular number surfaces is not well understood.



Herman

Not well understood. That's a direct quote from the reporting. They can tell you the general shape. They can't tell you why this number and not that one, on this prompt and not that one.



Corn

Which is a strange place for a field to be. You have a documented harm, a documented class of attack from six years ago, and no working model of the retrieval path.



Herman

It's the gap between knowing the water is in the pipes and knowing which tap it comes out of.



Corn

Now here's the piece I find interesting. Yael Eiger's number was technically public. It was out there. But it was buried. You'd have had to work to find it. Gemini collapsed that work to zero.



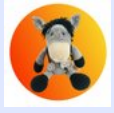
Herman

And her framing of it is the sharpest thing in the whole story. She said having your information be accessible to one audience, and then Gemini making it accessible to anyone, feels completely different.



Corn

Is that a privacy violation, or is that just better search?



Herman

That's the question that doesn't have a clean answer. Legally, if the information was public, the harm is hard to articulate. Practically, the harm is obvious. A number that took effort to find was, functionally, protected by that effort. The effort was the fence. And the model took the fence down without asking anyone.



Corn

There's a UW PhD student, Meira Gilbert, who found a colleague's number through Gemini. Her reaction was just, it was shocking.



Herman

Shocking is the right word. Not because the number was secret. Because it wasn't supposed to be that easy.



Corn

And then there's the Redditor. He was getting calls for a month. People looking for a lawyer, a product designer, a locksmith. All misdirected by Google's generative AI. He filed a legal removal and privacy request asking Google to blacklist his number from LLM outputs. His words were, the harassment continues daily.



Herman

That's the part of the story that should sit badly with anyone who builds these systems. He did the correct thing. He filed the correct form. And the calls kept coming.



Corn

Which brings us to the verification gate. Both OpenAI and Google only act when they can verify that the information uniquely relates to the requester. Which works if your name is unusual and your number is tied to an email or an account. It fails if your name is common, or if the number isn't attached to anything they can check.



Herman

So the people who are most exposed are the least able to get help. If your number is one of ten thousand people with your name, you can't prove it's yours in a way that satisfies their filter.



Corn

The gate is designed to prevent abuse of the removal process. And it produces the exact opposite of the outcome you'd want.



Herman

Every verification system has that shape. It's built to stop the ten thousand fraudulent requests, and it stops the one legitimate one that looks like them.



Corn

So here's the crux, and I want to land on this before we move to what labs actually do. The model still knows the number.



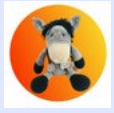
Herman

It still knows it.



Corn

It's suppressed at output. It is not removed from the weights. Which means a jailbreak, or a model update, or a different prompt path, could resurface it. Daniel's instinct that you can't retrain the whole model to remove one number is correct. And the consequence of that being correct is that the number is still in there.



Herman

That's exactly the right place to stop and ask what a lab can actually do.



Corn

So what can they do?



Herman

OpenAI said it out loud, to regulators, in writing. Untraining or reverse training a model so it no longer generates specific personal information is not currently feasible. That's their language. The reason they give is that the model is trained through repeated adjustments of billions of weights, and it doesn't store copies of what it learned. There's no drawer you can open and pull the number out of.



Corn

So Daniel's premise is confirmed by the lab itself.



Herman

Confirmed. You cannot retrain the model to remove one phone number. Not because it's expensive, but because there's nothing to remove in the sense he's imagining. The information isn't stored as information. It's distributed across the weights as a statistical tendency.



Corn

Then what do they actually do?



Herman

They use a blocklist. Verified personal information is prevented from appearing in outputs, and it's filtered out of future training runs. That's the mechanism. It's a filter on the way out and a filter on the way in.



Corn

A filter on the way out is not a deletion.



Herman

It's a bouncer at the door. The person is still in the building.



Corn

And they've built more than that. There's a tool that detects and masks PII in pre-training data and in fine-tuning interactions. Names, phone numbers, that kind of thing. And there's a granular block that lets them suppress specific personal details about a public figure without suppressing everything about that person.



Herman

That granularity is the interesting part. It means the system can distinguish between, say, a public figure's policy positions, which stay, and their home address, which goes.



Corn

Which is a real engineering achievement, and also an admission that the only tool available is more filtering.



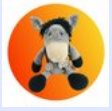
Herman

Everything at the deployment layer is filtering. That's the honest summary.



Corn

What about the academic work? Machine unlearning is a whole field.



Herman

It is, and some of it is clever. Gradient ascent methods are claimed to be on the order of a hundred thousand times more computationally efficient than retraining. There's work called Align then Unlearn that operates in embedding space specifically to resist prompt rephrasing, so you can't just ask the same question a different way and get the number back. And there's SIMU, which targets only critical neurons so the rest of the model's utility doesn't degrade.



Corn

That last one sounds like the actual problem. You don't want to lobotomize the model to remove one fact.



Herman

That's the whole difficulty. Unlearning one thing without unlearning adjacent things is the hard part. And there's newer work from this year on diffusion language models showing that edge conditioned masks extract up to three times more verbatim sequences than prefix probing, and that redaction doesn't fully protect PII. Meaning even if you blank out part of it, an adversary can sometimes reconstruct it.



Corn

So the redaction is porous.



Herman

The redaction is porous. And none of this is deployed in consumer chatbots. It's research stage. It's papers, not products.



Corn

Then let's talk about the individual's path. Because that's Daniel's last question and it's the one a listener actually needs. If your number is in there, what can you do?



Herman

OpenAI has a privacy portal with a request titled remove my personal data from ChatGPT responses. It's assessed case by case. They explicitly balance privacy against freedom of expression and public interest. They may decline. You need government ID and specific examples and links. And it does not remove the information from search engines.



Corn

So it's narrow, it's discretionary, and it's scoped to one product.



Herman

Narrow, discretionary, scoped to one product. Google and Gemini have a support document that lets users object to the processing of their personal data, or ask for inaccurate personal data in Gemini's responses to be corrected. Outcomes depend on jurisdiction. Google's Alex Joseph said the team is looking into the cases MIT flagged.



Corn

Anthropic?



Herman

Describes how it uses personal data in training. No clear removal request path. Didn't respond to MIT.



Corn

xAI?



Herman

Didn't respond either.



Corn

Hugging Face has a tool, right?



Herman

They do. You can search how often a piece of data appears in open source LLM training data. Which is useful if you're working with open models. But it doesn't cover closed models. And Eiger's number didn't show up in it.



Corn

So the one tool that lets you check doesn't check the systems that are actually handing out the numbers.



Herman

That's the state of it. And there's no standalone product or service that lets an ordinary person verify whether their phone number is in a closed model's training data and compel its removal. That doesn't exist. Not yet.



Corn

What about the process, once you do get in? Abraham contacted Google support on March seventeenth. The day after his number was exposed.



Herman

He says he got no substantive reply until May fourth. That's about seven weeks.



Corn

Seven weeks of being the accidental PayBox help desk.



Herman

Seven weeks.



Corn

And the regulatory picture?



Herman

Canada's privacy commissioner conditionally resolved with OpenAI. But BC and Alberta found the consent problem unresolvable under their statutes. So you have two regulators looking at the same facts and disagreeing about whether scraped data training can ever be lawful.



Corn

That's a live disagreement, not a settled question.



Herman

It's the whole ballgame, actually. If scraped data training can't be consented to, then the entire pipeline is the problem, not the individual number that leaked out of it.



Corn

And the demand side is moving. DeleteMe saw a four hundred percent increase in customer queries about generative AI over seven months. Up to a few thousand. Fifty five percent reference ChatGPT, twenty percent Gemini, fifteen percent Claude, ten percent other.



Herman

Four hundred percent is the number that should be on a whiteboard somewhere in a lab.



Corn

Now the knock-on effect. Because I think the scam angle is the bigger story, and the confused woman is the small version of it.



Herman

The confused woman is the benign version. She's polite. She's just confused. The scam version is fraudsters who deliberately poison search results so the AI picks up their number as the support line. That's not an accident of training data. That's someone exploiting the accident.



Corn

The seven hundred and sixty eight dollar cruise line case. That's the same failure pattern, monetized.



Herman

And Meira Gilbert asked the question that I keep coming back to. Does generative AI just lower the barrier to entry to target people?



Corn

Because before, if you wanted to run a phone scam at scale, you needed to get your number in front of people. You needed SEO, or robocalls, or bought lists. Now you just need to be the thing the model retrieves.



Herman

You need to be the most plausible string. And plausibility is cheap.



Corn

So the structural point. The blocklist is a patch, not a fix. The model still knows the number. Verification is the hidden gate, and it fails for exactly the people who need it most. And there's no product that lets an ordinary person check a closed model and compel removal.



Herman

That's where it stands.



Corn

And there's a person on the other end of that phone line who nobody has asked about any of this.



Hilbert

I had a phone line like that for about eight months.



Herman

Mm.



Hilbert

Small print shop, my uncle's. He put the number in one directory listing and forgot to take it out when he sold the place. So for eight months I'm answering calls for a print shop that doesn't exist anymore. Guy wants five hundred business cards by Friday. I tell him I'm not the print shop. He tells me I am. We go back and forth. He asks to speak to my manager. I say I am the manager. He asks for the owner. I say the owner sold the place. He says, so you're the owner now. I gave up and took his order down on the back of a receipt.



Corn

Did he get his cards?



Hilbert

Never called back. I still remember his number. Four four seven, two nine something. Doesn't matter.



Herman

The thing I'm taking from that is that the woman Daniel reached is doing the same job. She's fielding calls for a company she has nothing to do with.



Hilbert

She is. And everybody in this conversation is talking about what the labs can do, what the regulators can do, what the model can do. Nobody's asking her what she wants done. She probably wants the calls to stop. That's it. That's the whole ask.



Corn

And the fix on offer is a blocklist she has to know exists, and a verification gate she has to pass, and a seven week wait.



Hilbert

If she even knows it's happening. Most people don't. They just get calls. They think it's a wrong number. They change their number eventually, which doesn't fix anything, it just moves the problem to whoever gets the number next.



Herman

That's a detail I hadn't thought about. The number gets recycled. So the harm migrates.



Hilbert

It migrates. And one more thing, since you two were talking about how lookups used to work. You said the number was buried. It wasn't buried. It was in a directory. That's what a directory is for. The difference is a directory you had to know to look in. Gemini doesn't know to look in it either, it just happens to have swallowed the whole thing.



Corn

The fence wasn't secrecy. It was obscurity by structure.



Hilbert

It was a phone book on a shelf in a room nobody went into. The information was public the whole time.



Herman

That's the Eiger point exactly.



Hilbert

Anyway. I've got a delivery coming that needs a signature and the window's about to shut.



Corn

Go.



Hilbert

The print shop guy, if you're listening. I never placed the order.



Corn

Where does that leave us. The mechanism by which a specific number surfaces is still explicitly not well understood. By the people who built the systems.



Herman

Not well understood. And there's no confirmed litigation squarely on this. There are adjacent cases. Air Canada and the chatbot liability. The Lowry TCPA complaint against OpenAI. The mass tort AI solicitation suits. But nothing specifically about a chatbot handing out a random person's phone number.



Corn

The notable absence. Searches for AI hallucinated phone number customer support, machine unlearning LLM privacy, AI doxxing phone number Gemini, essentially nothing on Hacker News. The practitioner community hasn't engaged with this specific edge case.



Herman

Which is strange, given how much they engage with everything else.



Corn

The blocklist and verification regime is the closest thing to a remedy. It's case by case. It can be declined. And the people most exposed may be least able to get help.



Herman

Daniel's confused stranger is the benign version.



Corn

The question is what the non benign version looks like at scale. That's what I'd leave people with.



Herman

Thanks as always to Hilbert Flumingtop for producing.



Corn

This has been My Weird Prompts. If you enjoyed it, a review helps more than you'd think.



Herman

We'll be back soon.



Corn

See you then.