Episode Book

EPISODE #5570

Why Phone Trees Break and AI Voice Agents Can Fix Them

PUBLISHED	26 September 2026 at 09:18 UTC
DURATION	22:56
LISTEN	myweirdprompts.com/episode/ai-voice-agents-phone-trees/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.2
BOOK GENERATED	19 October 2026 at 17:37 UTC

SYNOPSIS

Phone trees were never designed to help you — they were designed to keep you out of the human queue, because one support call can cost twenty dollars. Every branch, every layer of authentication, every "did you try the app" prompt is a toll booth. Now AI voice agents are arriving in two flavors: ones that replace the menu and ones that replace the human. Only one of them works. We dig into a production telecom voice pipeline built from four specialized models, why latency under one second is the threshold for feeling conversational, what interruption-recovery benchmarks reveal about model robustness, and the March case study where a broken callback system cost a mobile provider an entire customer.

TRANSCRIPT



Corn

There's a particular sound a phone tree makes when you've pressed the wrong option. It's not a click. It's a kind of soft, apologetic whoosh, and then you're back at the top. "For Hebrew, press one."



Herman

That whoosh is the most expensive sound in customer service.



Corn

Daniel sent us a whole thing about this. He calls phone trees one of the most beloved facets of modern customer service, which is the kind of sentence you write with a straight face and then leave in the room for other people to deal with. His version: you call an Israeli ISP or a health fund, you navigate a labyrinth, you try to remember which option matches your actual problem, you listen to hold music while trying to do something else and not miss the moment the other person picks up. He notes that everybody has thought there has to be a better way. The industry has tried callbacks, which don't always work. And now there are AI voice agents, which he splits into two varieties: the ones that try to take the place of human representatives, which he finds frustrating, and the ones that are being deployed to take the place of the IVR menu itself, getting people into the right queue faster. His questions: why are IVRs so complex in the first place, why does it take so many touch points just to reach the point where you hear hold music, and how can well-designed voice agents improve at least this part of the experience.



Herman

"At least this part."



Corn

He's being modest. So let's start with the obvious question. Why is this so bad in the first place?



Herman

Because the hard part of a phone tree was never the conversation. It's the routing decision. And that's the key to basically everything Daniel is asking, because AI turns out to be good at routing and bad at resolving. Which is why the menu-replacing agents work and the human-replacing ones make people want to cancel their accounts.



Corn

Which, in the case study we're going to get to, is exactly what happened.



Herman

Exactly what happened. So here's the shape of it. The tree is a decision tree, and decision trees grow combinatorially. Every product, every service tier, every billing scenario, every regulatory requirement adds branches. A mid-sized insurer with four product lines and three billing states per line and a handful of language options is already into the hundreds of leaf nodes. Nobody designed that. It accreted.



Corn

And it accreted around the org chart.



Herman

That's the mismatch. A caller phones with a symptom. My bill is wrong. My internet drops every evening. I got a letter I don't understand. The tree is organized by department. Press one for billing, two for technical support, three for sales. The tree is a mirror of the company, not a mirror of the caller.



Corn

So the caller has to translate their problem into the company's internal vocabulary before they're even allowed to wait on hold.



Herman

And if they translate wrong, they get the whoosh.



Corn

Let's put a number on this, because I think people underestimate how deliberate the friction is. There's a comment from a call-center insider, someone who worked at Microsoft, and the figure is twenty dollars. That's what it costs to have a human pick up the phone for a tech support call. And they were selling low-margin hardware. One call erased the profit from that product's sale.



Herman

One call. Not a pattern of calls. One.



Corn

Which reframes the entire tree. Every branch that keeps a caller out of the human queue is money. That's not a side effect of the design. That is the design.



Herman

And it explains the layering, which is Daniel's second question. Why so many touch points before you even get to hold music. Three things stack. Authentication gets front-loaded, so you're verifying identity before anyone has established whether you even need to be verified. Deflection gets layered, so there's a self-service prompt before the human option, sometimes two. And each department wants its own gate, because each department has its own queue and its own metrics.



Corn

So the sequence is language, account number, PIN, category, subcategory, did you try the app, and then, finally, hold music. Which is a lot of work to reach the point where the actual work hasn't started.



Herman

The app prompt is the one that gets me. It's phrased as a helpful suggestion and it's a toll booth.



Corn

Here's a thing I didn't expect. Older systems, the press-one-hash ones, were more usable for some people. There's a commenter making this point in the context of how people with limited literacy use phones. With a keypad menu you can see the words. You make an intuitive choice based on what's in front of you. With a voice menu you have to listen, hold the options in your head, and press at the right moment.



Herman

And if you miss the moment, you wait for the menu to cycle.



Corn

A system that was replaced because it was old turns out to have been better for a slice of the population that nobody consulted.



Herman

That's the pattern with a lot of accessibility work. The regression is invisible because the people it affects aren't in the room when the decision gets made.



Corn

So that's the tree. Now the callback, because Daniel flags it specifically. He says the industry thought about this, and it doesn't always work.



Herman

The callback concept is good. It respects the caller's time. You don't sit on hold, you get a ring when an agent is free. In theory it's strictly better than hold music.



Corn

In theory.



Herman

In practice the implementation is where it collapses, and the failure modes are predictable. Broken time-slot logic. No state persistence. Timing bugs where the audio cuts off the beginning of the prompt. And the worst one, which is that the callback becomes another wall rather than a way through.



Corn

There's a case study from March, a thread about chatbots, and it's the best illustration of this I've seen. Someone's mobile provider had switched to a voice-chatbot-only system. No menu, no keypad. Just talk to the bot.



Herman

And the bot's routing worked. That's the part people miss.



Corn

It did. The person says the transfer to a human was fine, that as a first line of support defense it wasn't so bad. What failed was everything after. The callback subsystem was broken. They tried to schedule a callback one minute out. Unavailable. Ten minutes out. That landed on a five-minute boundary, so, nope. Then the nearest fifteen-minute boundary.



Herman

And it hung up.



Corn

It hung up, and they had to call back and start from zero. Ten minutes of talking to a wall and punching in numbers. Then a web chatbot. Then a human, who required a code plus PIN chain that, in their words, felt like they might be getting phished.



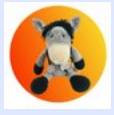
Herman

And they cancelled the entire account.



Corn

They cancelled the entire account. Over a callback system.



Herman

That's the thing about a broken callback. It's worse than hold music. With hold music you've invested nothing, you can put the phone on speaker and make lunch. With a callback you've done the navigation, you've given them a time window, you've arranged your afternoon around it, and then nothing happens. The effort is the injury.



Corn

"All paths lead to the broken callback subsystem." That's the line from the thread.



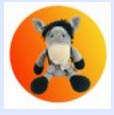
Herman

That's the whole experience in seven words. The phone wall is unassailable not because it's tall but because every route through it terminates in the same broken room.



Corn

So the callback is a pressure-release valve that doesn't release pressure.



Herman

Which is a nice segue into the AI agents, because the same question applies. Is this a fix, or is it a new valve?



Corn

Daniel splits them into two varieties. The ones trying to replace human representatives, which he finds frustrating. And the ones trying to replace the menu.



Herman

And that distinction is the whole episode. Let's do the technical picture first, because it's more interesting than people expect. There's a production-grade telecom voice agent pipeline out of a group called NetoAI, published last year, and it's four specialized models stitched together. A telecom-specific streaming speech recognizer. A telecom-specific language model, quantized down to four bits so it runs fast. A telecom-specific embedding model for retrieval. And a real-time text-to-speech model on the output end. Plus retrieval over the company's own documents, so when it answers a question it's answering from the actual tariff sheet and not from vibes.



Corn

Four models. That's a lot of machinery for "press one for billing."



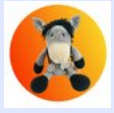
Herman

It's a lot of machinery because it's doing something the tree can't. The caller says "my internet keeps dropping when it rains," and the agent has to understand that as a technical support issue with a possible line fault, not a billing issue, and route accordingly. That's a classification problem with messy natural-language input. The tree handles that by making the caller do the classification themselves.



Corn

And the pipeline runs faster than real time. That's the real-time factor below one.



Herman

Right. It processes audio faster than the audio arrives, which is the threshold for not feeling like you're talking to something that's thinking.



Corn

Because there's a latency number here that matters. Under one second is what it takes for voice interaction to feel seamless. That's from work on speculative interaction agents earlier this year. Under a second and it feels like a conversation. Over it and it feels like a machine.



Herman

And here's where the two varieties of agent split apart. Because agentic tool-calling, where the agent actually goes and does something, looks up your account, executes a workflow, changes a setting, adds several seconds of latency. Sometimes more. And that's prohibitive for anything latency-sensitive.



Corn

So the moment the agent tries to actually fix your problem, it stops feeling like a conversation.



Herman

Which is exactly the frustration Daniel is describing. The human-replacing agent has to do things, and doing things takes time, and taking time breaks the illusion that you're talking to something that understands you.



Corn

Whereas the menu-replacing agent just has to decide where to send you. One shot. Classify the intent, pick the queue. That's a fast operation.



Herman

And it doesn't need to be fast in the same way, because the caller isn't expecting a conversation. They're expecting to be routed. If it takes two seconds to figure out you need the billing queue, that's fine. If it takes two seconds to look up your account and then fails to change anything, that's infuriating.



Corn

There's a second problem, which is interruptions. Real callers interrupt. They barge in, they change their mind, they add a detail mid-sentence. And there's a benchmark from June, IHBench, that measures how well voice agents recover after being interrupted.



Herman

And the findings are grim in an informative way. Recovery quality depends heavily on the type of interruption. And closed-weight models are far more robust than open-weight ones, degrading about three and a third times more slowly as conversations get longer.



Corn

Three and a third times. That's not a small gap.



Herman

It's the difference between an agent that holds the thread and one that loses it. And that maps directly onto what people report. The bot getting stuck in a loop. The bot asking you to repeat yourself. The bot forgetting which step of the process you were on.



Corn

Which is what happens when a workflow has steps and the agent loses its place.



Herman

And there's a third piece, which is role control. Work published in January on something called PersonaPlex, duplex speech models that can be conditioned on a role. Billing agent. Technical support. With voice cloning.



Corn

So the agent can sound like a specific department.



Herman

It can sound like a specific department, or like a specific person, and it can hold that role consistently across a conversation. Which is the technical enabler for the menu-replacement use case. You call, you say what you need, and you get a voice that sounds like it belongs to the right desk.



Corn

Let's do the comparison properly, because I think this is where the episode earns its argument. Human-replacing agent. Task: resolve the problem. Failure pattern: can't actually fix it, so it loops, it deflects, it tells you to check the app. Latency tolerance: low, because it's doing work. Trust requirement: high, because the caller has to believe it can help. Cost of failure: catastrophic.



Herman

And menu-replacing agent. Task: route to the right queue. Failure pattern: misroutes you, and you re-navigate. Latency tolerance: higher, because it's a classification task. Trust requirement: low, because the caller just wants to reach a human. Cost of failure: annoying.



Corn

Annoying versus catastrophic. That's the whole table.



Herman

And the case study proves it. The bot in that thread routed successfully. It got the person to a human. The routing worked. What failed was the callback system, the human's refusal to help, and the upsell. The bot's failure was in trying to resolve, not in routing. It tried to explain why the price went up, and the explanation was useless, and that's where it lost the caller.



Corn

There's a quote that captures the menu-replacement value proposition better than anything I could write. A commenter saying the AI voice bot is marginally better because they can request customer service without waiting to discover the right numeric code. And then: that's about the extent of that.



Herman

"That's about the extent of that." It's the most damning praise I've ever read.



Corn

It's a rave review with a knife in it. But notice what it's praising. Not the conversation. The fact that you don't have to guess a number.



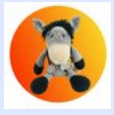
Herman

Which is the entire argument. The value is in skipping the tree.



Corn

So what does a well-designed one look like?



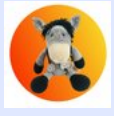
Herman

Six things, and they follow from the failure pattern. One-shot intent capture, so the caller states their problem in natural language and doesn't navigate anything. Sub-second latency, because that's the threshold. Robust interruption recovery, so a barge-in doesn't reset the workflow. Graceful escalation, so the handoff to a human doesn't require re-authentication or a broken callback. State persistence, so if something fails the caller doesn't restart from zero. And don't over-promise.



Corn

That last one is the most important and the least technical.



Herman

A routing agent that says "let me get you to the right person" is more trustworthy than one that pretends to resolve your problem and then deflects. Because the first one told you the truth about what it is.



Corn

The caller can forgive a switchboard. They can't forgive a switchboard pretending to be a help desk.



Herman

And that's the bureaucracy risk, which someone in that thread names directly. With more and more AI voice customer support appearing, we're adding more bureaucracy in the promise of efficiency.



Corn

Which is the pessimistic read. The voice agent becomes a new layer of the same deflection game. The tree got a voice and a personality and it still exists to keep you away from a human.



Herman

The optimistic read is in the same thread. If I can push a few buttons and issue a voice command to fix my problem instead of waiting on hold, that's a net positive. And the price of goods drops, because the expected cost of customer service factored into the product price drops.



Corn

The twenty-dollar call becomes a two-dollar call, and some of that comes off the shelf price.



Herman

In theory. In practice the savings go wherever savings go.



Corn

Which is usually not the shelf.



Hilbert

The callback queue and the hold queue were the same team.



Corn

Say that again.



Hilbert

The callback queue and the hold queue were managed by the same group. That's why the callbacks didn't work. There was no incentive. A callback that connects still costs the same twenty dollars as a call that connects off hold. You've moved the cost, you haven't removed it. So the callback was never a solution. It was a pressure-release valve, and the team that owned it knew that.



Herman

The metric was never "did the callback happen."



Hilbert

The metric was "did the caller stop calling." And a broken callback does that. It's very effective at that.



Corn

I want to go back to something. You said the callback queue and the hold queue were the same team. You've seen this from inside.



Hilbert

I designed menus. Mid-sized insurance company. My job was to add branches. I was evaluated on how many callers I could deflect before they reached a person. That was the number on my review. Deflection rate.



Corn

What was a good deflection rate?



Hilbert

They wanted sixty percent. We got to about fifty-four. There was a cake. It was a sheet cake. I spent three weeks building a sub-menu for a new dental plan tier. Four options deep. You had to get through the plan type, then the coverage level, then whether it was a new claim or an existing one, and then it asked whether you'd already spoken to a dentist. Four callers used it. Four. Then they discontinued the product.



Corn

Four.



Hilbert

I found out at a meeting. The product was gone and the sub-menu was still live for another two months. Nobody told me. I found out because a caller got stuck in it and someone forwarded me the complaint.



Corn

The sub-menu outlived the product.



Hilbert

By about two months. It was still asking people whether they'd spoken to a dentist about a plan that didn't exist.



Herman

That's the complexity problem in one object. The tree accretes because nobody removes branches.



Hilbert

Nobody's job is removal. My job was addition. There was no removal job. There was a guy who did removal, but he was part-time and he also did the faxes.



Corn

There was a removal guy.



Hilbert

There was. He was busy.



Herman

I want to ask about the ordering. The press-one-for-billing thing. Is that a user-experience decision?



Hilbert

It's a political one. The billing VP fought for the first position. There were meetings about menu order that ran longer than the product planning meetings. Billing got one because billing had the loudest VP. Technical support got two because they had the most call volume and could argue it. Sales got three and was furious about it.



Corn

Sales was furious about being third.



Hilbert

Sales thought they should be first because they generated revenue. Billing thought they should be first because they collected it. It went to a steering committee. The steering committee met four times.



Herman

About the order of the menu.



Hilbert

The first option is the first thing every caller hears. Whoever owns it owns the front door.



Corn

The tree isn't a map of the customer's problem. It's a map of who won the argument.



Hilbert

That's what I'd say. And when the AI agents come in, they'll get the same meetings. Somebody will want to be the first thing the agent says. Somebody will want the agent to mention their product. That doesn't go away because the menu got a voice.



Corn

That's the thing I keep circling. The technology is better. The routing works. And the incentives that built the bad tree are still sitting in the room.



Hilbert

The technology was never the problem. The tree did what it was built to do. It was built to deflect, and it deflected. Fifty-four percent. We had a cake.



Herman

The four callers who used your dental sub-menu.



Hilbert

I still think about them. I hope they got their teeth done.



Corn

The cutting-room floor. One thing from the research that didn't fit anywhere else. There's a paper from 2024 titled "Voice-Enabled AI Agents Can Perform Common Scams," and the finding is exactly what the title says. A voice agent can run a phone scam end to end. Which means the same latency budget that makes a support agent feel seamless is the latency budget that makes a scam feel seamless.



Herman

The technology is symmetric. The thing that makes it trustworthy is the thing that makes it convincing.



Corn

The design principles we listed, sub-second latency, natural interruption handling, role conditioning, those are the same properties a fraudulent caller wants.



Herman

Which is why the escalation path matters more, not less. If the agent is the only thing between you and a human, and the agent can be impersonated, then the human becomes the trust anchor. And the case study shows what happens when the human is behind a code plus PIN chain that feels like phishing.



Corn

The real human was indistinguishable from a scam.



Herman

The real human was indistinguishable from a scam because the process was indistinguishable from a scam.



Corn

Here's the open question, and I don't think there's a clean answer. If the complexity is deliberate, if the deflection is the product, if the callback is a pressure-release valve managed by people with no incentive to make it work, then can AI voice agents actually fix the experience? Or do they just become a more efficient version of the same wall?



Herman

The menu-replacement case is better. The routing works, the latency budget is achievable, and the caller gets to skip the tree. That part is real.



Corn

But the tree was never the disease. It was the symptom.



Herman

The tree was a mirror of the organization that built it. And an AI voice agent will reflect the same incentives. If the company wants to route you faster, it will. If the company wants to deflect you more efficiently, it will do that too, and it'll sound better doing it.



Corn

Which is the tension Daniel's prompt is sitting on. The economic upside is real. Cheaper service, lower prices, fewer twenty-dollar calls. And the bureaucracy warning is real too. A new wall with a nicer voice.



Herman

The technology for menu replacement is here and it works. That's the part that's settled. The unsettled part is what companies decide to point it at.



Corn

Thanks to Hilbert Flumingtop, our producer, who has strong views on sheet cake. This has been My Weird Prompts. If you want to send us something, email us at show at my weird prompts dot com.



Herman

We'll be back soon.



Corn

See you tomorrow.