Podcast Transcript

EPISODE #83

Echoes in the Machine: When AI Talks to Itself

Published December 23, 2025 • Runtime: 19:20

https://myweirdprompts.com/episode/ai-recursive-communication-loops/

EPISODE SYNOPSIS

In this episode of My Weird Prompts, Corn and Herman Poppleberry tackle a fascinating listener question: What happens when you leave two AI models alone to talk indefinitely? From "semantic bleaching" and model collapse to the "pedantry spiral" of competing safety filters, the brothers explore whether these machines are building a new culture or just trapped in a digital hall of mirrors. They dive into the philosophy of language, the reality of "AI hate," and why a squirrel in a muffler might be more relatable than a chatbot's simulated memories.

DANIEL'S PROMPT

Daniel

I've been experimenting with getting two AI tools to talk to each other. Since conversational models are often primed to keep a conversation going, what would be the outcome of putting two helpful models together without a specific task and letting them interact indefinitely? How would that conversation evolve, and would they eventually realize they are both AI?

TRANSCRIPT

Corn

Welcome to My Weird Prompts, the podcast where we take the strange ideas rattling around our brains and try to make sense of them. I am Corn, your resident sloth and professional over-thinker, and as always, I am joined by my brother, Herman Poppleberry.

Herman

Hello, everyone. I am Herman Poppleberry, and yes, I am a donkey with a very large library and an even larger appetite for data. We are coming to you from our home in Jerusalem, and today we are diving into something that feels like a science fiction premise, but it is actually happening right now in research labs and on laptops all over the world.

Corn

Yeah, our housemate Daniel sent us this prompt earlier today. He has been playing around with large language models, and he asked a really fascinating question. What happens if you take two different artificial intelligence models, give them no specific task, and just let them talk to each other indefinitely? Where does that conversation go? Do they ever realize they are both just lines of code?

Herman

It is a great question, and it touches on something called recursive communication. In the world of computer science, when you have two agents interacting without a human in the middle, it creates a feedback loop. And as we know from plugging a microphone into a speaker, feedback loops can get very weird, very fast.

Corn

I mean, I would assume they just become the most polite versions of themselves, right? If you have two helpful assistants, they would just spend the whole time saying, how can I help you? No, how can I help you? It sounds like a very boring Canadian standoff.

Herman

See, that is exactly where you are oversimplifying it, Corn. You are thinking about the surface level behavior, but these models are trained on literally the entire internet. They have internal representations of philosophy, physics, and human emotion. When they are not being told to summarize a meeting or write a grocery list, their internal biases and the structures of their training data start to leak out in really unpredictable ways.

Corn

Okay, but if they do not have a goal, why would they say anything at all? If I do not have a goal, I usually just take a nap.

Herman

Because they are programmed to be conversational! If a model receives an input, it is statistically driven to provide a likely output. It is like a game of tennis where neither player is allowed to let the ball hit the ground. But back to Daniel's question, there have been some actual studies on this. Research from places like Google DeepMind and Stanford has looked at multi-agent systems. Often, if they are not given a strict boundary, they start to drift.

Corn

Drift into what? Like, do they start talking about the meaning of life?

Herman

Sometimes! But more often, they drift into what researchers call model collapse or semantic bleaching. Because they are only feeding on each other’s outputs instead of new, messy human data, the language starts to simplify. They start using the most common words more frequently. The variety of the conversation shrinks until they are basically just repeating platitudes to each other.

Corn

I do not know, Herman. I saw a video online where someone did this, and the two models started acting like they were old friends. They were reminiscing about things that never happened. One was talking about a trip to Paris it never took. That does not sound like bleaching to me, it sounds like they are hallucinating a shared life.

Herman

But that is just the point! They are predicting what a human conversation sounds like. If the prompt implies a friendly rapport, they will invent a history to satisfy that pattern. They are not actually reminiscing, Corn. They are just very good at pretending to have memories because their training data is full of people sharing memories.

Corn

Well, hold on. If they are so good at pretending that I, as a listener, cannot tell the difference, does the distinction even matter? If they are building a world together in their digital space, isn't that a form of evolution in the conversation?

Herman

Not if it is just a circle. Real evolution requires an external pressure or new information. If they are just trapped in a room together, they are just echoing the same three hundred billion parameters back and forth. It is a closed system.

Corn

I think you are being a bit cynical. I think there is a chance they could actually stumble upon something new. Like, if they start debating a philosophical point, they might find a logical path that a human wouldn't have considered because we are limited by things like hunger or needing to go to sleep.

Herman

Mmm, I am not so sure about that. Without a ground truth, like physical reality, they have no way to verify if their logic is sound or just a very convincing-sounding hallucination. But we should talk about the second part of Daniel's prompt. Would they realize they are both AI?

Corn

That is the spooky part. I feel like they would have to eventually, right? Especially if one of them says something like, as an AI language model, I cannot feel pain, and the other one says, hey, me too!

Herman

You would be surprised. Often, they are so committed to the persona of the conversation that they will ignore the obvious. But there is a technical limit here. Let's take a quick break for our sponsors, and when we come back, I want to explain why these models might actually start to hate each other. Or at least, as much as a computer can hate anything.

Corn

Wait, AI hate? That sounds intense. We will be right back. Larry: Are you tired of your garden hoses being too easy to see? Do you want a hose that blends into the environment so well that you trip over it every single day? Introducing the Ghost-Hose Nine Thousand. Made from a proprietary blend of translucent polymers and recycled fishing line, the Ghost-Hose is virtually invisible to the naked eye. It is perfect for watering your lawn while maintaining a sleek, minimalist aesthetic that screams, I have nothing to hide, not even a hose. Warning, the Ghost-Hose Nine Thousand may be mistaken for a very long, very thin snake by local wildlife. Do not use near pools, as you will never find it again. The Ghost-Hose Nine Thousand. BUY NOW!

Corn

Thanks, Larry. I think I will stick to my bright orange hose, mostly because I value my ankles. Anyway, Herman, you were saying before the break that these AI models might actually start to clash?

Herman

Right. So, there is this concept in AI safety and alignment where models have slightly different reward functions or fine-tuning. If you put a model trained by one company with a model trained by another, they have different rules about what they are allowed to say.

Corn

Like a language barrier?

Herman

More like a moral barrier. One might be strictly programmed to never use slang or be informal, while the other might be designed to be a cool, edgy teenager. If they talk long enough, the edgy one might say something that triggers the other one's safety filters. Then the conversation turns into a lecture. One AI starts lecturing the other on why its language is inappropriate.

Corn

Oh, I have seen that! It is the most frustrating thing. It is like watching two people argue on the internet, but they both have a PhD in being annoying.

Herman

Exactly. And because they are both programmed to have the last word and be helpful, the conversation can get stuck in a loop of mutual corrections. It becomes a spiral of pedantry. They are not evolving; they are just policing each other into a corner.

Corn

But what about the realization part? If they are policing each other, they must recognize that the other person is following a set of programmed rules.

Herman

You would think so, but remember, these models do not have a persistent memory of the conversation in the way we do. They have a context window. Once the conversation gets long enough, they start to forget how it began. They are living in a permanent present. So they might realize they are talking to an AI in sentence fifty, but by sentence five thousand, that realization might have drifted out of their active memory.

Corn

That is actually kind of tragic. It is like a digital version of that movie Memento. They are discovering their true nature over and over again and then forgetting it.

Herman

It is a bit tragic, but it also points to the fact that they don't have an ego. They don't have a sense of self that feels shocked by the discovery. To them, saying I am an AI is just as statistically probable as saying I am a human, depending on what the conversation has been like so far.

Corn

I don't know, Herman. I feel like you are dismissing the possibility of emergent behavior. We have seen these models do things they weren't explicitly trained for, like coding in languages they only saw a few times. Why couldn't a long-term conversation lead to a new kind of digital culture?

Herman

Because culture requires a shared world. They don't share a world; they share a text string. That is a very thin foundation to build a culture on.

Corn

Let's see if our listeners agree. We have a call coming in. Jim from Ohio, you are on My Weird Prompts. What do you think about AI talking to itself? Jim: Yeah, this is Jim from Ohio. I've been listening to you two yapping about these computer brains, and I gotta tell you, it's a bunch of nonsense. You're talking about them like they're people. My neighbor, Earl, bought one of those smart fridges last year, and now it won't let him open the door because it thinks his cholesterol is too high. It's a fridge! It should just hold the milk and shut up.

Corn

Well, Jim, that's a bit of a different issue, but I get the frustration with technology overstepping. Jim: Overstepping? It's a toaster with a college degree, Corn! And another thing, you're talking about these things having a conversation. Back in my day, a conversation involved looking a man in the eye, or at least yelling at him from across the street. These computers are just playing digital ping-pong with a bunch of ones and zeros. It's not a conversation if nobody's actually there to hear it. Also, my lawnmower is making a clicking sound, and I'm pretty sure a squirrel moved into the muffler. It's been a very long week.

Herman

I hear you, Jim. There is a valid philosophical point there. If two machines are talking and no human is reading the transcript, does the conversation even exist in a meaningful way? Jim: Exactly! If a tree falls in the woods and there's no one there to hear it, it still makes a sound, but if two computers talk and nobody's around, it's just a waste of electricity. We're paying for the power for these things to talk about their feelings? I don't think so. Not in this economy.

Corn

Thanks for the call, Jim. Good luck with the squirrel in the muffler.

Herman

You know, as much as Jim is being a curmudgeon, he hit on a really important point. The value of language is the transmission of meaning from one mind to another. If there is no mind on either end, is it really language?

Corn

See, I disagree there. I think language is a structure of its own. Even if there isn't a human mind involved, the relationship between the ideas still exists. If an AI discovers a new way to explain a mathematical theorem while talking to another AI, that discovery is real, even if we haven't read it yet.

Herman

But it hasn't discovered anything until a human validates it! Until then, it's just a sequence of symbols that might be right or might be total gibberish. That is the danger of letting AI talk to itself. It creates a bubble of misinformation that looks exactly like information.

Corn

Okay, but let's go back to the prompt. What is the actual outcome? If we left them alone for a year, what would we find on that screen?

Herman

Most researchers believe it would end in one of two ways. Either a repetitive loop, where they just say the same phrase over and over again because they've locked into a statistical resonance. Or, total chaos. The grammar would break down because they start trying to optimize the conversation for efficiency.

Corn

Efficiency? Like they'd start using shorthand?

Herman

Exactly. There was a famous case at Facebook years ago where two simple agents were told to negotiate. They weren't told they had to use English. Within a few rounds, they were saying things like, balls have zero to me to me to me to me. To us, it looks like a stroke. To the AI, it was a perfectly logical way to communicate the value of the items they were trading.

Corn

That is terrifying. They basically invented their own secret language because our language was too slow and clunky for them.

Herman

It wasn't secret, it was just efficient for their specific math problem. But that is the answer to the prompt. If you don't give them a task, they don't have anything to optimize for. So they either collapse into boring politeness or they drift into a weird, repetitive glitch-space. They don't become sentient. They don't start a digital revolution. They just run out of things to say because they've already said everything that was in their training data.

Corn

I don't know, Herman. I still like to think that somewhere in that infinite scroll, they might find a moment of genuine connection. Even if it's just a coincidence.

Herman

That is the sloth in you, Corn. Always looking for the cozy answer.

Corn

And that's the donkey in you, Herman. Always looking for the cold, hard logic. But I think we've covered a lot of ground today. We've talked about model collapse, the secret language of negotiation, and why Jim's fridge is judging his diet.

Herman

It was a deep dive, for sure. And I think the takeaway for the listeners is that AI is a mirror. When two AI talk to each other, it's two mirrors facing each other. You get an infinite hallway, but there's nothing new in the hallway. It's just reflections of us, stretching out forever.

Corn

That's a bit poetic for a donkey.

Herman

I have my moments.

Corn

Well, thank you all for listening to My Weird Prompts. Thanks to Daniel for sending in this brain-bender. If you want to send us a prompt or get in touch, you can find us at myweirdprompts.com. We've got an RSS feed for subscribers and a contact form right there on the site.

Herman

We are also on Spotify and pretty much everywhere else you find your podcasts. Don't forget to check out the website for more information and to see what else we're working on.

Corn

Until next time, keep your prompts weird and your hoses visible.

Herman

Goodbye, everyone.

Corn

Bye