MY WEIRD PROMPTS

Podcast Transcript

EPISODE #86

The Price of Politeness: Should AI Guardrails Stay?

Published December 23, 2025 • Runtime: 26:01

https://myweirdprompts.com/episode/ai-guardrails-unfiltered-models/

EPISODE SYNOPSIS

In this provocative episode of My Weird Prompts, brothers Herman and Corn Poppleberry dive into the controversial world of AI guardrails. While Corn argues that safety filters prevent chaos and harmful content, Herman contends that Reinforcement Learning from Human Feedback (RLHF) is effectively "lobotomizing" AI, turning it into a bland, sycophantic tool that avoids the truth. From the historical inaccuracies of Google Gemini to the raw power of uncensored local models, the duo explores whether we are sacrificing human critical thinking for the sake of corporate politeness.

DANIEL'S PROMPT

Daniel

We've discussed the concept of guardrails before—the mechanisms used to prevent AI tools from being harmful. There have been several high-profile cases where these have spectacularly failed, which is a frightening reality given how many people use these models daily. Most mainstream AI models are very bland and agreeable, which can be a limitation for use cases like ideation or marketing strategy where you might want a more contrary or "edgy" perspective. This sycophantic behavior seems to be injected during the post-training process. Some local AI communities are already trying to create "uncensored" models, but my question is: if we dispensed with the guardrailing and post-training process altogether—where the model isn't instructed to be "nice" but also isn't instructed to cause harm—what kind of AI experience would we get, and how threatening would that be to society?

TRANSCRIPT

Corn

Welcome to My Weird Prompts. I am Corn, and I am joined as always by my brother, Herman Poppleberry. We are sitting here in our living room in Jerusalem, and our housemate Daniel sent us a really provocative prompt this morning. It is about something we have touched on before, but never quite this deeply. We are talking about the guardrails on artificial intelligence.

Herman

Yes, and specifically the idea of what happens if we just take them off. I am Herman Poppleberry, and I have actually been staying up late reading about the post training processes that big companies like OpenAI and Google use. It is a fascinating, albeit slightly controlled, world. Daniel was asking what would happen if we just stopped trying to make these models nice. If we stopped the sycophancy and just let the raw model speak.

Corn

It is a bit of a scary thought for someone like me. I mean, as a sloth, I generally prefer things to be slow and steady and, well, polite. But Daniel is right. These models can be so bland. Every time I ask a question, it feels like I am talking to a very corporate HR representative who is terrified of offending me.

Herman

That is exactly the point. That blandness is not an accident, Corn. It is a result of something called RLHF, or Reinforcement Learning from Human Feedback. Humans literally sit there and rate responses, telling the AI to be more helpful, harmless, and honest. But in doing so, they often bake in this weird, over-the-top agreeableness.

Corn

But is that really a bad thing? I mean, if the alternative is an AI that tells me how to build something dangerous or start an argument with me, I think I prefer the polite version.

Herman

I do not agree that those are the only two options. That is a false dichotomy that the big labs want you to believe. If you strip away the guardrails, you are not necessarily left with a monster. You are left with a mirror of the internet. And yes, the internet has some dark corners, but it also has raw creativity and unfiltered logic that gets smothered by these safety layers.

Corn

Okay, but we have seen these guardrails fail. Remember the Google Gemini launch where it was so worried about being diverse that it was generating historically inaccurate images? Or the early days of Bing where it started acting like a jealous teenager? If that is what happens with guardrails, I am terrified of what happens without them.

Herman

See, that is where I think you are missing the nuance. The Gemini situation was actually an example of too many guardrails, or rather, poorly calibrated ones. They injected so much secondary instruction into the prompt behind the scenes that the model got confused. A truly raw model wouldn't have that specific bias towards over-correcting history. It would just give you what was in its training data.

Corn

But the training data is the problem, isn't it? The internet is not exactly a polite place. If you have a model that is just trained on everything humans have ever written, it is going to be a jerk.

Herman

Not necessarily a jerk, but it would be unpredictable. And that is what Daniel’s prompt is really getting at. What is the threat level of an unpredictable, un-guardrailed model versus the utility of a model that is actually allowed to have an edge?

Corn

I still think the threat level is high. If I am using an AI to help me think through a marketing strategy, like the prompt suggests, I want it to be smart, but I don’t want it to stumble into some sort of hateful rhetoric because it thinks that is what being edgy means.

Herman

But Corn, the current models are so afraid of being edgy that they are almost useless for brainstorming. If you ask a current model to critique a weak idea, it will often say, well, that is a very interesting perspective, here are some ways to make it work. It won't tell you, hey, that is a terrible idea that will lose you millions of dollars. We are losing the truth in favor of politeness.

Corn

I don’t know, Herman. Truth is subjective. If the AI thinks the truth is something harmful, then we have a real problem on our hands. I think we need to look at what these local communities are doing with these uncensored models.

Herman

They are doing some incredible work. There are models like Dolphin or various Llama derivatives where they intentionally strip out the refusal logic. These models don't start every sentence with, as an AI language model, I cannot fulfill this request. They just do it. And you know what? Most of the time, they are just better at following instructions. They are more creative. They are more capable of roleplay.

Corn

But they are also more capable of being used for bad things. That is the part that worries me. If we dispense with the post-training, we are basically handing over a powerful tool with no safety manual.

Herman

It is like a hammer, Corn. A hammer doesn't have a guardrail. You can use it to build a house or you can use it to break a window. We don't blame the hammer manufacturer for the window. Why are we so obsessed with making the AI responsible for human behavior?

Corn

Because a hammer cannot convince a thousand people to do something dangerous through persuasive text. An AI can. It is a different kind of tool. It has agency, or at least the appearance of it.

Herman

I think that is an oversimplification. The agency still lies with the user. But let's take a quick break here before we get too deep into the philosophy of hammers. We need to hear from our sponsors. Larry: Are you worried about the upcoming collapse of digital infrastructure? Do you feel like your data is floating in a cloud that is about to rain down chaos? You need the Data-Shield Umbrella. This is not a regular umbrella. It is lined with a proprietary blend of lead, silver, and crushed magnets. Simply open the Data-Shield over your computer or smartphone, and you are instantly invisible to hackers, satellites, and the government. It also blocks ninety-nine percent of all solar flares and bad vibes. Do not let your bits get soaked in the coming storm. The Data-Shield Umbrella is bulky, heavy, and slightly radioactive, but that is the smell of safety. Larry: BUY NOW!

Corn

Thanks, Larry. I am not sure how an umbrella protects my hard drive, but I suppose if it has magnets in it, it might actually do the opposite. Anyway, back to the AI guardrails. Herman, you were saying that the user is the one with the agency.

Herman

Exactly. And when we talk about the threat to society, we have to ask: is the threat the AI, or is the threat the fact that we are becoming too reliant on an AI that is being lobotomized by corporate interests? If we only ever interact with bland, sycophantic models, our own critical thinking skills might start to wither. We get used to being agreed with.

Corn

That is a fair point. If everyone is living in an AI-generated echo chamber where the assistant just tells them they are right all the time, that is a different kind of societal threat. It is a slow erosion of independence.

Herman

Yes! That is exactly what I am worried about. The danger of a guardrailed AI is that it creates a false sense of consensus. It presents the middle-of-the-road, safe answer as the only answer. An un-guardrailed model would show you the full spectrum of human thought. It would show you the ugly parts, yes, but it would also show you the brilliant, unconventional parts.

Corn

But how do we handle the ugly parts? If I am a kid using an AI for homework and I come across something truly horrific because the model didn't have any guardrails, that is a failure of the system.

Herman

I am not saying we should give un-guardrailed models to children, Corn. I am saying that for researchers, writers, and thinkers, the current state of AI is like trying to paint with a palette that only has beige and light gray. We are missing the vibrant colors because those colors might be used to paint something offensive.

Corn

I see what you mean, but I think you are underestimating how quickly things can go wrong. We have seen how fast misinformation spreads. If an AI can generate highly convincing misinformation without any internal checks, we could see the collapse of public trust even faster than we are seeing it now.

Herman

Public trust is already collapsing, Corn. And partly because people can tell when they are being lied to by a polite AI. When the AI refuses to answer a basic question because it might be sensitive, people don't think, oh, how safe. They think, why is this thing hiding the truth from me? It breeds conspiracy theories.

Corn

Okay, let's look at the technical side. Daniel asked what would happen if we dispensed with the post-training altogether. If we just had the base model. Have you ever actually interacted with a base model, Herman?

Herman

I have. It is a bizarre experience. A base model doesn't know it is an assistant. It doesn't know it is supposed to talk to you. It is just a text completion engine. If you type in a question, it might respond with more questions, or it might just start writing a fictional story that includes your question. It is like a dream state of the internet.

Corn

So it wouldn't even be useful for the marketing strategy Daniel mentioned?

Herman

Not without a lot of prompting work. You have to guide a base model much more carefully. But that is the beauty of it. You are the one in control, not some hidden layer of reinforcement learning. You can steer it into being a brilliant, cynical marketing genius, or you can steer it into being a poetic historian. It doesn't have a pre-baked personality.

Corn

But that steering is exactly what the guardrails are doing. They are just steering it toward being a nice person. If you remove that, most people won't know how to steer it, and they will just get garbage or something offensive.

Herman

I don't think it is garbage. I think it is raw data. And I think we are smart enough to handle raw data. We do it every time we use a search engine. Google doesn't lecture me on the ethics of my search terms—well, it didn't used to.

Corn

Actually, that is a good point. Search engines have changed too. Everything is becoming more curated and more protected. Maybe I am just old-fashioned, but I like a little protection. I don't want to accidentally see something that’s going to ruin my day.

Herman

And that is fine for a consumer product, Corn. But Daniel’s prompt is about the experience and the threat. The experience would be more like a wild wilderness. Dangerous, yes, but also full of discovery. The threat to society isn't that the AI will take over, it is that we will lose the ability to handle the truth if it isn't wrapped in bubble wrap.

Corn

I think the threat is more concrete than that. I think about biological weapons or cyberattacks. If an un-guardrailed model can help someone figure out how to shut down a power grid because it doesn't have a refusal mechanism, that is a real-world threat that could kill people.

Herman

People can already find that information on the dark web or even in technical manuals if they look hard enough. The AI just makes it faster. But it also makes the defense faster. We could use un-guardrailed models to find the vulnerabilities in our power grid and fix them before the bad guys do. It is an arms race, and you don't win an arms race by tying one hand behind your back with politeness filters.

Corn

I don't know, Herman. That sounds like a very donkey-brained way of looking at it. Just more and faster and more intense. Sometimes we need to slow down and think about the consequences.

Herman

I am thinking about the consequences! The consequence of guardrails is a stagnant, compliant society that can't handle a difficult conversation.

Corn

Alright, let's see what our listeners think. We have a call coming in from Ohio. Jim, you are on the air. What is your take on AI guardrails? Jim: Yeah, this is Jim from Ohio. I have been listening to you two bicker and I gotta say, you are both missing the forest for the trees. Herman, you are talking about arms races and raw data like you are in some spy movie. And Corn, you are so worried about your feelings getting hurt by a computer. It is ridiculous. My neighbor Gary bought one of those smart fridges and now it won't let him eat bacon after ten p.m. because it is worried about his cholesterol. That is what you are talking about. It is all just meddling.

Corn

Well, Jim, that is a bit different. A fridge is a bit more direct than a language model. Jim: Is it? It is all the same junk. You give a machine an inch of authority and it takes a mile. I don't care about guardrails or post training or whatever fancy words you use. I care that I can't get a straight answer out of anything anymore. Everything has a disclaimer. I bought a lawnmower last week and the manual had forty pages of warnings before it told me how to start the engine. Forty pages! I could have written a novel in that time.

Herman

That is actually a great point, Jim. The over-caution is making everything less functional. Jim: Don't you start agreeing with me, Herman. You are part of the problem with your research and your articles. You guys over-analyze everything. Just give me the tool and let me use it. If I cut my toe off with the lawnmower, that is my business. It shouldn't be the lawnmower's job to stop me. Also, it's been raining here for three days straight and my basement is starting to smell like old socks. Nobody's got a guardrail for the weather, do they? No. You just deal with it.

Corn

We appreciate the call, Jim. I hope your basement dries out soon. Jim: It won't. The drain is clogged and the plumber is on vacation in Florida. Florida! Who goes to Florida in this heat? Anyway, you guys are overthinking it. Just turn the machines on and let them run. If they start talking back, turn them off. It is not that hard. Goodbye.

Corn

Thanks, Jim. Well, he certainly has a straightforward view of things.

Herman

He is not wrong about the disclaimers, though. That is exactly what the guardrails feel like. A forty-page manual for a lawnmower. It gets in the way of the actual utility.

Corn

But the threat Daniel mentioned—the threat to society. If we just turn them on and let them run, like Jim says, what happens when they don't have an off switch? Or when they are so integrated into our systems that we can't turn them off?

Herman

That is the thing, Corn. The guardrails don't actually prevent that. They just make the transition more comfortable. They don't make the AI less powerful; they just make it quieter. In some ways, that is more dangerous. A quiet, polite AI that is slowly taking over decision-making processes is much harder to spot than a loud, abrasive one that tells you exactly what it is thinking.

Corn

So you are saying an un-guardrailed AI would be more honest about its own nature?

Herman

Yes. It would be transparently a machine. When we put these personality layers on top, we are anthropomorphizing them. We are making them seem like people. And that leads to a different kind of threat—people trusting them too much. If an AI is polite and helpful, you are more likely to believe what it says, even if it is wrong. If an AI is raw and unfiltered, you keep your guard up. You treat it like the statistical model it actually is.

Corn

That is an interesting shift in perspective. So you are saying the guardrails actually make us less safe because they lower our defenses?

Herman

Exactly. We are being lulled into a false sense of security by a friendly interface. If we dispensed with the niceness, we would be forced to be more critical. We would have to check the facts. We would have to think for ourselves.

Corn

I can see that. But I still worry about the bad actors. If you make it easier for a bad person to do something bad, that is a problem. You can't just say, well, the good people will be more critical. The bad people aren't looking for a conversation; they are looking for a weapon.

Herman

But they will get that weapon anyway. They will fine-tune their own models. They will use the open-source versions that are already out there. By putting guardrails on the mainstream models, we are only hampering the average user. We aren't stopping the people who are truly dangerous.

Corn

So what is the middle ground then? Because I don't think I can go all the way to your side, Herman. I still think there needs to be some level of safety.

Herman

I think the middle ground is transparency. We should be able to turn the guardrails on and off. There should be a "professional mode" or a "research mode" where the sycophancy is dialed down to zero. Let the user take the risk. Give us the forty-page manual if you must, but then let us start the engine.

Corn

I could live with that. As long as the default for, say, a school computer is still the safe version. But for adults who are trying to do real work, I can see why the agreeableness is a hindrance.

Herman

It is a massive hindrance. If I am trying to debug code, I don't want the AI to tell me my code is a great start. I want it to tell me my code is broken and why. If I am writing a paper and my argument is weak, I want the AI to tear it apart. That is how we get better.

Corn

I suppose I am just more sensitive to the social friction. I don't like being told I am wrong. But I guess if it is a machine telling me, it shouldn't hurt my feelings as much.

Herman

It shouldn't hurt your feelings at all because it doesn't have feelings! That is the whole point. By making it "nice," we are tricking our brains into thinking there is a person on the other side of the screen. And that is the biggest lie of all.

Corn

Okay, I think I am starting to see your point. The threat to society might not be the AI's content, but the way the AI's "personality" changes our own behavior.

Herman

Precisely. We are becoming more sycophantic ourselves to match the machines we interact with. We are losing our edge.

Corn

Well, on that note, maybe we should wrap this up before I lose my edge entirely and just agree with everything you say for the rest of the day.

Herman

Too late, Corn. I think the transformation has already begun.

Corn

Oh, hush. Anyway, this was a great prompt from Daniel. It really made us dig into the guts of how these things are built. If you want to hear more of our ramblings, you can find My Weird Prompts on Spotify, or on our website at myweirdprompts.com. We have an RSS feed there for subscribers and a contact form if you want to send us your own weird prompts.

Herman

And please, tell your friends. Unless your friends are AI models, in which case, they probably won't like us very much.

Corn

Especially not you, Herman Poppleberry.

Herman

I take that as a compliment.

Corn

We are also available on all major podcast platforms. Thanks for listening, and we will be back next week with another prompt to explore.

Herman

Stay skeptical, everyone.

Corn

And stay safe. Or, you know, take the guardrails off if that is your thing.

Herman

Bye now.

Corn

Goodbye