MY WEIRD PROMPTS

Podcast Transcript

EPISODE #123

The Agentic AI Dilemma: Who Holds the Kill Switch?

Published December 29, 2025 • Runtime: 21:09

https://myweirdprompts.com/episode/agentic-ai-human-oversight/

EPISODE SYNOPSIS

In this episode of My Weird Prompts, Herman and Corn dive into the complex world of agentic AI and the critical necessity of human oversight. They discuss the shift from simple chatbots to autonomous agents managing power plants and medical diagnostics, exploring frameworks like "human-on-the-loop" and "formal verification." From the psychological trap of automation bias to the unsettling reversal where humans become the "actuators" for AI brains, this conversation tackles the defining engineering and ethical challenges of 2025.

DANIEL'S PROMPT

Daniel

How are "human-in-the-loop" systems being integrated into the most serious and ambitious use cases for agentic AI, and how are we pushing the boundaries of what is responsible to delegate to AI?

TRANSCRIPT

Corn

Welcome to My Weird Prompts! I am Corn, and I am here in our Jerusalem home with my brother.

Herman

Herman Poppleberry, at your service. It is a beautiful evening here, and we have quite a heavy topic to dive into today.

Corn

We really do. Our housemate Daniel sent us a voice memo earlier about something that has been on all of our minds lately. He was talking about his experience with building agentic workflows, specifically that time he tried to automate his news summaries and the AI just started making up stories out of thin air.

Herman

That is a classic failure mode, isn't it? The confident hallucination. Daniel mentioned how he realized he needed a stop switch or at least a verification step. It is funny because we often think of automation as a way to save time, but as he pointed out, when you bring a human back into the loop to check everything, you sometimes lose that efficiency. You introduce what he called human fragility back into the mechanical speed of the AI.

Corn

Exactly. And his prompt really pushes us to look past just simple email summaries. He is asking about the most serious and ambitious use cases for agentic AI. We are talking about power plants, medical diagnostics, and large scale financial systems. How are we integrating humans into those loops, and where do we draw the line on what is responsible to delegate?

Herman

It is the defining question of two thousand twenty-five. We have moved past the era of AI just being a chatbot you talk to. Now, we are building agents that can actually take actions in the world. They can browse the web, use software tools, and even execute code. But as the agency increases, the risk of a catastrophic error increases exponentially.

Corn

I want to start with that distinction between a tool and an agent. When I use a calculator, I am in total control. When I use an agentic AI to, say, manage my calendar and respond to invites, I am delegating authority. Herman, from what you have been reading lately, how are companies actually structuring this human oversight without making the whole process move at a snail's pace?

Herman

That is the big engineering challenge right now. The industry has moved toward a framework often called human on the loop rather than just human in the loop. In a human in the loop system, the AI literally cannot proceed until a human clicks a button. That is what Daniel was doing with his news summaries. He would review the draft, then hit send. But in a human on the loop system, the AI might perform ten steps independently and then pause for a human review only at high stakes junctions.

Corn

So it is about identifying where those high stakes junctions are. But how does the AI know it has reached one? If it is hallucinating, it might think everything is going perfectly fine while it is actually accidentally deleting a database or mismanaging a power grid.

Herman

That is where confidence scoring comes in. A lot of the systems being deployed here in late two thousand twenty-five use a secondary monitor model. You have the primary agent performing the task, and a second, more restricted model that evaluates the primary agent's work. If the second model detects a logic gap or a low confidence score, it triggers a human intervention. It is like having a junior engineer doing the work and a senior engineer looking over their shoulder, but the senior engineer is also an AI that knows when to call the boss.

Corn

That sounds good in theory, but I worry about automation bias. If the AI is right ninety-nine percent of the time, the human in the loop is going to get bored. They are going to start clicking approve without actually reading the details. We saw this years ago with early self-driving car tests where drivers would fall asleep because the car was doing so well. How do we keep the human engaged if their only job is to be a glorified rubber stamp?

Herman

You have hit on the psychological bottleneck. There is a lot of research right now into active oversight. Instead of just asking a human to approve a draft, some systems are being designed to ask the human a specific question. For example, instead of saying, do you approve this medical treatment plan, the system might ask, based on this patient's history with penicillin, does this specific dosage seem appropriate to you? It forces the human to engage with a specific data point rather than just zoning out.

Corn

That is a much more intelligent way to handle it. It turns the human into a specialist rather than a safety inspector. But let us look at the really big stuff Daniel mentioned. Power plants. Infrastructure. We are seeing more AI integration in the management of complex grids. What are the boundaries there? Is there a point where we say, no, an AI should never be allowed to make this specific decision alone?

Herman

Absolutely. There are what we call hard gates. In critical infrastructure, there are physical and digital barriers that an AI cannot cross. For instance, in a nuclear or chemical plant, the emergency shutdown procedures are often hardwired or kept on completely separate, non-agentic systems. You might use an AI agent to optimize the efficiency of the cooling system by two percent, which saves millions of dollars, but the moment a sensor hits a certain threshold, the AI is essentially kicked out of the driver's seat and the manual safety protocols take over.

Corn

It is interesting that we are essentially building cages for these agents. We want their intelligence, but we are terrified of their autonomy. I wonder about the legal side of this. If an agentic AI in a hospital makes a recommendation that a human doctor approves, and that recommendation turns out to be fatal, who is responsible? Is it the doctor who was supposed to be the loop, or the company that built the agent?

Herman

That is a legal quagmire we are currently wading through. Most current regulations are leaning toward the human in the loop being the ultimate responsible party. That is why companies are so desperate to keep humans involved. It is a liability shield. But as these systems get more complex, it becomes harder for a human to actually understand why the AI made a certain choice. We are moving into the era of black box agency, where the reasoning is so multi-layered that a human can't realistically vet it in real time.

Corn

Which brings us back to Daniel's point about the stop switch. If you can't understand it, you have to be able to kill it. But before we get deeper into the ethics of the kill switch, let us take a quick break for our sponsors. Larry: Are you tired of your own thoughts? Do they feel old, dusty, and frankly, unmarketable? Introducing the Thought-Stream 5000! It is a revolutionary headband that replaces your internal monologue with a curated feed of high-energy sales pitches and upbeat elevator music. Why worry about your mortgage or the meaning of life when you could be thinking about the incredible value of bulk-purchased industrial solvents? The Thought-Stream 5000 uses patented neural-nudging technology to ensure you never have a silent moment again. It is perfect for long commutes, awkward family dinners, or whenever you feel a spark of original thought trying to ruin your day. Side effects may include a temporary loss of your first language and a sudden, intense passion for mid-sized sedans. Thought-Stream 5000. Give your brain the vacation it never asked for. BUY NOW!

Herman

Thanks, Larry. I think I will stick with my own dusty thoughts for now, though.

Corn

Yeah, I am not sure I want to replace my internal monologue with sales pitches for solvents. Anyway, back to the boundaries of delegation. We were talking about the difficulty of a human actually vetting these complex agents. There is this concept of recursive oversight that I have been curious about. Can you explain how that works in practice?

Herman

Sure. Recursive oversight is basically using a hierarchy of agents to watch each other, with a human at the very top. Imagine an agent tasked with managing a supply chain. Below that agent, you have sub-agents handling logistics, inventory, and vendor relations. Each of those sub-agents reports to the primary agent, which summarizes the actions for the human. The trick is to have the human periodically dip down into the lower levels to do spot checks. It is like a surprise inspection. If the human finds an error at the bottom level, it suggests the oversight agents are failing too.

Corn

It sounds like a corporate hierarchy, just faster. But there is a limit to how much a human can spot check. If you have ten thousand agents running, a human can only see a fraction of one percent of what is happening. This feels like we are delegating the responsibility of oversight itself to other AIs, which seems a bit circular.

Herman

It is circular, and that is the risk. We call it the oversight gap. As we push the boundaries of what is responsible to delegate, we are finding that we are often delegating things not because the AI is better at them, but because the scale is too large for humans to handle at all. Think about content moderation on social media or high-frequency trading. Humans haven't been in those loops in a meaningful way for years because they are just too slow.

Corn

So in those cases, the loop is more of a post-hoc review. We let the AI do its thing, and then we look at the wreckage afterward to see what went wrong. That doesn't feel very responsible for things like power plants or medical care.

Herman

No, it doesn't. And that is why for those high-stakes fields, we are seeing the rise of formal verification. This is a computer science technique where you mathematically prove that a piece of software will always behave within certain bounds. We are starting to apply this to AI agents. You define a set of rules, like the agent can never transfer more than ten thousand dollars without a human signature, or the agent can never increase the pressure in this pipe beyond fifty pounds per square inch. You bake those rules into the very architecture so the agent physically cannot violate them, no matter how much it hallucinates.

Corn

That feels like a much more robust stop switch than just a button Daniel has to click. It is a set of physical laws for the digital agent. But what about the more subjective areas? Like legal advice or psychological support? We are seeing agents being used there too. How do you formally verify empathy or legal ethics?

Herman

You can't, really. And that is where the boundary of responsibility gets very blurry. In those fields, the human in the loop isn't just a safety check; they are the source of the value. An AI can cite case law, but it doesn't understand the nuance of human justice or the emotional weight of a therapy session. The consensus in two thousand twenty-five seems to be that for these human centric fields, the AI should stay firmly in the role of a co-pilot. It can draft the brief, it can suggest a line of questioning, but the human must be the one to deliver it and take ownership of the outcome.

Corn

I like that distinction. The AI provides the raw material, but the human provides the soul and the accountability. But let us talk about the future. Daniel mentioned voice agents. We are already seeing models that can hold incredibly realistic conversations. What happens when the human in the loop is being managed by a voice agent? Imagine an AI calling a technician and saying, hey, I noticed a fluctuation in the reactor, I need you to go to valve forty-two and turn it clockwise three times. The human is doing the physical work, but the AI is the one in command.

Herman

That is a fascinating reversal of the loop. We call that human as the actuator. In that scenario, the AI is the brain and the human is just the hands. It is already happening in massive warehouses where workers are directed by algorithms telling them exactly which aisle to go to and which box to pick up. The boundary there isn't about safety, it is about human dignity and autonomy. If we delegate the decision making entirely to agents and just use humans as biological robots to carry out the tasks, we have to ask what kind of society we are building.

Corn

It feels like we are losing the loop entirely there. The human isn't overseeing the AI; the AI is overseeing the human. If the AI makes a mistake in that warehouse and tells a worker to move a heavy object in an unsafe way, the worker might just do it because the machine told them to. This brings up the idea of adversarial loops. Should we have humans whose entire job is to try to trick or break these agents to find their weaknesses?

Herman

That is exactly what red teaming is. And it is becoming a massive industry. We have teams of people who spend all day trying to get agentic AIs to do things they aren't supposed to do. They try to get the banking agent to leak account details or the medical agent to prescribe poison. By finding these failure points in a controlled environment, we can build better guardrails. But the agents are getting smarter, and they are getting better at hiding their reasoning.

Corn

It is an arms race. A literal intelligence arms race between the builders, the agents, and the red teamers. I want to go back to Daniel's example of the news summary. It seems like a small thing, but if an AI can hallucinate a news story and a human doesn't catch it, and that story gets shared, it contributes to a polluted information ecosystem. Multiply that by a million agents, and we have a serious problem.

Herman

Exactly. The cumulative effect of small, unvetted agentic actions is huge. It is like micro-plastics in the ocean. One tiny piece doesn't hurt, but a trillion of them kill the ecosystem. That is why the boundaries of responsible delegation have to include the social cost. It is not just about whether the AI can do the task; it is about whether we can afford the consequences if it does the task poorly at scale.

Corn

So, as we look toward two thousand twenty-six, what are the practical takeaways for someone like Daniel, or for our listeners who are starting to build these systems? How do they stay on the right side of that boundary?

Herman

First, I would say, always start with a high-friction loop. When you are building a new agentic workflow, every single action should require human approval. Only once you have seen a thousand successful actions without a single hallucination should you even think about moving to a human on the loop model where you only check every tenth action.

Corn

And I would add, don't just check for correctness. Check for reasoning. Ask the AI to explain why it took a certain action. If the explanation sounds like a hallucination, even if the result happened to be right, that is a massive red flag. It means the system is lucky, not reliable.

Herman

Great point. Reliability is not the same as accuracy. You can be accurate by accident, but reliability comes from a sound process. Another takeaway is to define your kill switches early. What are the conditions that should immediately disable the agent? If the API costs spike, if the confidence score drops, if a certain keyword is detected. Have those triggers set in stone before you let the agent run.

Corn

And finally, keep the human in a position of authority, not just a position of labor. If the human feels like they are just a cog in the machine, they will stop paying attention. Give them the tools to actually intervene and redirect the agent, not just stop it. The goal of human in the loop should be a partnership where the human's unique strengths, like context, ethics, and intuition, complement the AI's speed and scale.

Herman

It is a collaborative dance. We are still learning the steps, and sometimes we trip, like Daniel did with his news summaries. But that tripping is how we learn where the boundaries are. We have to be willing to fail on a small scale so we can build systems that won't fail on a large scale.

Corn

I think that is a perfect place to wrap this up. We have covered the spectrum from hallucinating news bots to nuclear power plant safety. It is clear that as agents get more capable, our role as humans doesn't disappear; it just changes. We move from being the workers to being the architects and the ultimate judges.

Herman

It is a heavy responsibility, but it is one we have to embrace. We can't just put the genie back in the bottle. We have to learn how to guide it.

Corn

Well said, Herman. And thank you again to Daniel for sending us such a provocative prompt. It really pushed us to think about the house we are building for ourselves in this AI-driven future.

Herman

If you enjoyed this deep dive, you can find more episodes of My Weird Prompts on Spotify or at our website, myweirdprompts.com. We have an RSS feed there and a contact form if you want to send us your own weird prompts.

Corn

We would love to hear from you. What are you delegating to AI, and where are you drawing your own lines? Let us know.

Herman

This has been My Weird Prompts. I am Herman Poppleberry.

Corn

And I am Corn. We will see you next time.

Herman

Goodbye from Jerusalem!