WEST YOST

AI GUARDRAILS #4

/// ARE AI AGENTS RIGHT FOR WATER UTILITIES?

Andy Bochman | Resilience Strategic Lead | West Yost

TL;DR

On the heels of major media coverage of Anthropic's too-dangerous-to-widely-release Claude Mythos version, the SANS AI Cybersecurity Summit explored significant advances in Agentic AI offense and defense. They did not address risks to utility operations (e.g., to ICS/OT systems) and also ignored intrinsic, non-adversarial risks (e.g., hallucinations, model drift, optimization errors, etc.). Those two issues matter a lot for water utilities. Nevertheless, the speakers were brilliant, admitted the gaps in their knowledge, and I left thinking that only the very best-resourced utilities should even consider experimenting with agentic AI in or near operations right now. And that the other 99% should wait, watch and learn.

INTRODUCTION

According to our colleague Kevin Morley at AWWA – the American Water Works Association – there are approximately 500 large water and wastewater utilities in the US. Which leaves 50,000 or so of small and medium size. Many have benefitted in recent years from the risk & resilience assessments, mandated by law, that AWWA and the EPA have sponsored, and West Yost and other consultancies have conducted.

Those assessments have considered all-hazards, which is to say at a minimum both cyber and physical risks. But as every reader not living under a rock knows by now, there's a new risk category in town. It goes by several names: artificial intelligence (AI), generative AI (GenAI) and most recently Agentic AI. And it's prompting a lot of heated debate about what this technology means for those charged with the safety and reliability of utility operations.

NOTES FROM SANS' AI CYBERSECURITY CONFERENCE

April 20-21, 2026. I'm putting the dates right here because it seems the situation changes every month, if not every week. While the water sector-specific rewards of embracing AI technologies remain to be seen, the risks are immediately obvious. So, I attended this conference fully focused on AI risks, run by the premier global cybersecurity training organization: the SANS Institute, to see if I might find ways of protecting those water utilities that decide to give AI a try.

Here's what I heard and what I learned:

Host Rob T Lee opened the conference saying, in the wake of massive press coverage on Anthropic's most recent model, Mythos, being too powerful to release to the public, that veteran cyber defenders are really struggling to know what to do. Lee noted that there have been eight major breaches since Mythos's announcement, and that AI-boosted ransomware is starting to hit hard. Well, that certainly set the tone.

In his keynote and later as a panelist, prolific Cybersecurity author Bruce Schneier identified integrity as the #1 AI security problem. He said AI integrity is closely related to correctness; that its presence is what makes systems trustworthy. And that without it, the lines between safety, reliability, and cyber start to blur. (Which I must point out, has always been in the case in the Operational Technology (OT) domain).

Schneier mentioned the Aladdin's genie problem, where AI agents grant wishes but will often achieve them in ways the wisher doesn't like. Akin to the paperclip maximizer, it's King Midas turning everything he touches to gold, including food and his daughter. In an **article he published in 2021**, at the dawn of GenAI, he wrote:

"Here's the thing: there is no way to outsmart the genie. Whatever you wish for, he will always be able to do it in a way that you wish he hadn't. The genie will always be able to hack your wish."

Then Anthropic's threat intelligence leader, Jacob Klein weighed in, saying AI is now a core capability for offense and defense and that Claude is often used to support cyber attackers and defenders. Because the model sees so much, he allowed that Anthropic is becoming a kind of Security Operations Center (SOC) for everyone.

So later I queried Claude about this prospect, informed by continuity from our previous conversations on similar matters, and I got this response:

"I don't have visibility into adversaries' actions or into how threat actors use Claude or any AI system. Anthropic doesn't share that data with me, and I don't have persistent awareness between conversations or access to threat intelligence feeds, SIEM data, logs, or OT network telemetry. But I could add value in a SOC-adjacent capacity with analyst augmentation, playbook development, threat modeling, alert triage, and post incident analysis."

So clearly there's a difference between what a senior Anthropic employee can see and say, and the capabilities the AI itself can represent for itself. As many smaller water utilities don't even have a SOC, even imperfect AI-assisted analysis, used thoughtfully, might be better than the current state for many small and mid-size utilities.

To be thorough and check on Klein's claims without pinging Claude again, I asked Gemini about the use of AIs as security partners and got this:

"As of early 2026, the landscape for AI-driven vulnerability scanning has moved far beyond simple chat-based code analysis. All three platforms—Gemini, OpenAI, and Claude—now offer dedicated, agentic security tools designed to identify, validate, and suggest fixes for vulnerabilities. While they aren't 'point-and-click' network scanners like Nessus or OpenVAS, they have become highly effective AI Security agents that operate within your development and security workflows."

CLOSING THOUGHTS

I ended day two thinking that from a pure intellectual, curiosity perspective, if I didn't have to worry about risks, this sure would be one exciting space. Maybe if I was just involved in IT matters with no kinetic public health considerations, this would all be so much less worrisome. Perhaps if utilities just used agents for the security assistance as described above, that would be ok. But I'm hearing from some that there's interest about seeing how agents can help as operations assistants, potentially saving time, money, electricity, and more.

Conference-wise, two topics I am acutely tuned into were missing:

1. The first is that there seemed to be very little awareness of cyber-physical security risks from Agentic AI, aka OT risks, aka Industrial Control System (ICS) risks. That was ok, though, as there's a different SANS conference for that coming up in June, the **SANS ICS Summit**.
2. I was hoping they'd address intrinsic risks at some point, of the kind previous papers in the **"I am the Guardrails"** series raised, including the line that Agentic AI

"... introduces authorized systems that can drift, hallucinate, degrade, and optimize toward unintended consequences — none of which register as anomalies under traditional cybersecurity monitoring."

I didn't hear one word on this, and for water or energy utilities, this is a very big deal indeed.

I mentioned #2 to Rob Lee, saying those risks are built right in and begin to threaten any organization that gives authorized access to agents, connecting them to SCADA systems, data historians or other data stores. I said detecting those issues isn't in the job description or skillset of cyber professionals and he agreed. If I heard him right, Rob said he's thinking those challenges may fall to another job function entirely, maybe one that doesn't exist yet. Will be following up with him on that.

In the meantime, at this time, it seems like only the best resourced water utilities should consider experimenting with Agentic AI technologies, and even they may come to find that they are better off without them. For those less blessed, resource-wise, it sure seems like holding off and monitoring how agents work out in leading/bleeding edge utilities (if there are such things) is the best course of action.

Andy Bochman is Resilience Strategic Lead at West Yost. He previously served as Senior Grid Strategist at Idaho National Laboratory, where he co-developed the Cyber-informed Engineering (CIE) methodology. He is the co-author of Countering Cyber Sabotage (CRC Press, 2021).

Comments and questions: resilience@westyost.com

Links:

1. https://www.schneier.com/academic/archives/2021/04/the-coming-ai-hackers.html
2. https://tinyurl.com/yke3a24r
3. https://www.westyost.com/position-papers/