WEST YOST

AI GUARDRAILS #1

# /// WHY THIS TIME IS DIFFERENT

Andy Bochman | Resilience Strategic Lead | West Yost

## TL;DR

*Generative Artificial Intelligence (GenAI) and Agentic AI introduce risks that are categorically different from traditional cybersecurity threats — systems that can cause harm while working exactly as instructed, through authorized access, under normal operating conditions. Existing cybersecurity frameworks are necessary but not nearly sufficient to govern these risks. This brief explains why, using realistic failure scenarios drawn from water operations. And the two papers that follow dive into the factors that make GenAI adoption compelling for the water utilities, and how engineering principles developed at a DOE Natonal Lab and embraced by the AWWA can help utilities build the guardrails (i.e., safety mechanisms ) they need to keep themselves and their customers safe and secure.*

## INTRODUCTION

A little more than a decade ago, cybersecurity leaders Josh Corman, Nick Percoco, and Beau Woods issued a declaration to critical infrastructure operators who were waiting for someone else to handle the growing digital threat landscape. Their message was blunt: "I am the Cavalry." Stop waiting for the cavalry to arrive. You are the cavalry. The message resonated, and in the years since, the best-run utilities have internalized it, building cultures where cybersecurity is everyone's job, not just IT's.

We are at a similar inflection point now, but the stakes have shifted. Cybersecurity was fundamentally about keeping bad actors out of systems that were otherwise working as designed, or detecting their movements once in. GenAI introduces something qualitatively different: systems that can cause harm while working exactly as instructed, through authorized access, under normal operating conditions. When your AI-enhanced decision support system optimizes pump schedules in a way that slowly fatigues pipe joints, or when it hallucinates a chemical dosing recommendation with the same confidence it delivers an accurate one, the failure isn't a breach. It's a gap in governance. And governance, unlike perimeter defense, cannot be purchased off the shelf or bolted on after the fact.

Which is why, for 2026 and beyond, a more apt mantra might be: **I am the Guardrails.**

The metaphor is deliberate. A guardrail doesn't stop you from driving. It keeps you on the road when things go sideways: when the fog rolls in, when the curve is sharper than you expected, when the road itself is still being built. That is precisely the situation water utilities now face with GenAI. The technology is powerful, it is arriving fast, and the road it's traveling has no lane markings yet. Your job is not to refuse the journey — with the wellbeing of your utility and the customers you serve in mind, it is to make sure you can navigate it safely.

This is the first in a three-paper series written for utility and public works directors, board members, engineers, IT and OT managers, and the regulators who oversee them. This paper focuses on the diagnostic: why GenAI risks are categorically different from traditional cyber threats, why existing frameworks are inadequate, and what new risk categories demand your attention. The second paper makes the economic case for adoption and explains why the forces driving GenAI into water utilities are so compelling. The third delivers the governance framework water utilities need to manage these risks safely.

## THE REGULATORY CONTEXT: A GOVERNANCE VACUUM

The instinct in regulated industries is to wait for guidance from regulatory agencies. In the water sector, that instinct has often served utilities well. In 2018, more than a decade after the electric sector implemented mandatory security controls, Congressional action made the Environmental Protection Agency (EPA) the implementor and AWWA stepped in to provide support for America's Water Infrastructure Act (AWIA). A sector-specific implementation of the NIST Cybersecurity Framework, the AWIA cyber risk management tool gives utilities a structured process for assessing risks and building resilience.

AWIA moved the sector forward but was built primarily for a world where the primary digital threat was unauthorized access to systems operating under human control.

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

The current administration's AI policy framework is explicitly designed to minimize regulatory burden on AI developers in the name of maintaining American competitiveness. The European Union AI Act classifies critical infrastructure AI as "high-risk" and imposes governance requirements, but its implementation timeline, jurisdictional reach, and enforcement mechanisms offer little near-term help to a utility operator in Fresno, Tacoma, or Milwaukee. State legislatures are beginning to act: Colorado, California, Illinois, and others have passed or proposed AI-related legislation, but the patchwork is inconsistent and largely focused on consumer-facing applications rather than operational technology (OT) in critical infrastructure.

The result is a governance vacuum, and water utilities are sitting squarely in the middle of it. They are being courted by AI vendors offering transformative capabilities — real-time predictive maintenance, intelligent SCADA augmentation, digital twins of retired operators' institutional knowledge — while operating under frameworks that have nothing to say about what happens when those capabilities fail, drift, or are compromised in ways that look nothing like a traditional cyberattack.

## EXISTING CYBER FRAMEWORKS: NECESSARY BUT NOT NEARLY SUFFICIENT

Traditional cybersecurity posits that defense is about perimeter protection and access control; that a system is safe when it is working as designed; that unauthorized access is the primary threat; and that detection should focus on anomalies and intrusions. But when it comes to GenAI-enhanced systems, the main shortcoming of cyber frameworks is that they don't address goal misalignment. A system pursues its objectives correctly, but those objectives are subtly wrong.

A robust defense against GenAI threats requires governing the decision-making process itself — something well outside the bounds of current cybersecurity frameworks, policies, and tools. Industrial security veteran and instructor Jason Christopher explains the difference precisely:

*The control point isn't just access control or detection anymore. It's governance over how decisions are generated, validated, and constrained in real operations. Security teams are used to thinking about adversaries. AI forces us to also think about authorized systems producing unsafe outcomes. It's an entirely different discipline.*

The cybersecurity community has embraced GenAI security so quickly that most people think cyber professionals are the ones best equipped to bring security to GenAI. They are, at least for defending the new attack surfaces that come with it. But they are not the right tribe for the safety engineering skills most needed.

## NEW RISKS TO HAVE ON YOUR RADAR

Most of these are not things utilities have had to know about, pay attention to, or devise plans to mitigate in the past. But they must be managed now if utilities are to safely realize the benefits of GenAI:

- Hallucination as a distinct risk category. This is the purest example of a risk that has no analog in traditional cyber frameworks. The system isn't compromised or misaligned — it's just confidently wrong.
- Model drift and degradation over time. No attacker is needed to trigger this. The model quietly becomes less accurate as the world changes around it, and traditional monitoring methods don't notice it.
- Data governance as a prerequisite that must be done well and in advance. Bad data in, bad decisions out — but unlike a traditional data quality problem, GenAI amplifies errors with false confidence.
- Sensitive data exposure. Sending infrastructure vulnerability data and system topology to cloud-based AI services creates an attack surface that didn't exist before GenAI adoption.
- The explainability problem for regulated utilities. A compliance risk that didn't exist until you put a black-box system between your operators and their regulatory obligations. No prior technology created this particular bind.

Frontier model vulnerability. As of early 2026, frontier models remain highly susceptible to jailbreaking. The pattern is consistent: a new model drops, bad actors crack it quickly, and post techniques publicly. Guardrails keep failing because they are designed to appease fears rather than fix vulnerabilities.

## WATER SECTOR SCENARIOS

Here are two illustrations of things that may go wrong even with solid cyber defenses in place. Note: neither is a cybersecurity failure. No unauthorized access occurs. No intrusion detection system fires off an alarm. These are failures of governance and operations for which existing cyber frameworks have nothing to say.

## SCENARIO 1: CATASTROPHIC OPTIMIZATION

An AI system optimizes pump operations to minimize energy costs, saving $2M in electricity over its first twelve months. Its optimization strategy involves running pumps in sequences that create pressure transients that were individually within normal operating parameters but created a cumulative fatigue pattern that no one previously had a reason to monitor for — establishing a novel failure mode. These transients slowly fatigue pipe joints in a way that had not happened under human operation.

In year two, a major main break occurs; resultant floods cause $100M in damage. Questions arise immediately: who's liable — the utility that deployed it? The AI vendor? The utility's engineer who approved it? Under current frameworks, the utility bears all the risk, and that engineer's career suddenly looks much less promising, to put it mildly.

## SCENARIO 2: A CYBERSECURITY CASCADE

A utility connects sensors to a GenAI system for predictive maintenance. The system has API connections to cloud services, third-party analytics, and remote access for vendor support. Attackers compromise the GenAI system. They don't shut anything down — they just subtly corrupt the AI's recommendations. Over months, the AI decision support system guides operators to decisions that degrade system resilience.

Then the attackers unleash the real attack: they shut down power to some pumps and further corrupt AI recommendations at the same moment. The system fails in ways operators can't diagnose because their muscle memory has deteriorated. This is cognitive offloading — they've been relying on the AI to think for them, and now without it, they are lost.

## THE INSURANCE GAP

Beyond the regulatory void, existing insurance policies — including cyber insurance — were not written for GenAI risks. No framework exists to assign liability, and no entity is rushing to fix this problem. This is the most concrete proof that utilities adopting GenAI solutions are operating without a financial safety net.

When the AI makes a bad recommendation and causes a failure, how do you determine whether it was a technical error, a data quality problem, or a deliberate attack? Today we are very far from having the ability to do this with any confidence. The absence of attribution capability compounds the liability problem enormously.

## WHAT COMES NEXT

Understanding why existing frameworks fall short is the first step. The second paper in this series makes the economic case for adoption — because the forces driving GenAI into water utilities are powerful and largely irresistible, which makes governance more urgent, not less. The third paper delivers a practical governance framework built on Cyber-Informed Engineering (CIE) principles, an emerging engineering discipline in the water sector, applied to AI safety challenges that cybersecurity was not designed to address.

The goal across all three papers is not to make you afraid of GenAI. It's to help you more fully understand the risks and prepare to mitigate them before you deploy it.

---

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