WEST YOST

AI GUARDRAILS #3

/// ENGINEERING THE GUARDRAILS

Andy Bochman | Resilience Strategic Lead | West Yost

TL;DR

Water utilities must build their own AI safety governance now, before deployment, because neither the federal government nor the AI industry will do it for them. This paper provides a practical framework built on Cyber-informed Engineering (CIE) principles — an emerging engineering discipline in the water sector— covering foundational prerequisites, the three most critical CIE principles for AI governance, agentic AI risks, and a standing caution about operational dependency that utility leaders must internalize before they flip the switch.

INTRODUCTION

The first paper in this series explained why GenAI risks are categorically different from traditional cybersecurity threats. The second made the economic case that these pressures are largely irresistible. This third and final paper delivers the governance framework — practical, actionable, and built on engineering principles that already exist in the water sector.

Let us begin with the immortal risk management observation of cybersecurity legend Dr. Dan Geer: "The wellspring of risk is dependency."

Geer's observation should serve as a standing caution throughout everything that follows: do not let your human-based operational capabilities atrophy even as GenAI technologies begin to prove their worth. That caution is not a reason to avoid deployment' it is the most important design constraint you will bring to it.

Though many may not know it, some water utilities still retain the ability to operate in fully manual or close-to-manual mode for extended periods. That is especially good news now that GenAI-based solutions leveraging Claude, Gemini, OpenAI, and others — including some built on open-source models — are knocking at the door. Water's close cousin, the electric sector, digitized its operations more briskly and in so doing became wholly dependent on computers, networks, and the cloud to perform its operational functions. Their manual operations capability left the barn a long time ago. Yours has not. Protect it.

SIX FOUNDATIONAL PREREQUISITES

As with building a house, you need to start with a solid foundation. Here are six foundational prerequisites, in order of urgency:

1. **Build governance capacity.** If your organization lacks it today, get it. For smaller utilities, this means bringing in external expertise and ensuring those voices are embedded in governance and decision-making processes. A governance structure that exists only on paper protects no one.
2. **Get your data house in order.** GenAI is only as good as what it can reach. Data must be scrubbed for accuracy, organized into repositories accessible to AI systems, and structured to yield insights across assets, teams, and time frames. This is not optional prep work — it is the single biggest determinant of whether your AI deployment succeeds or becomes an expensive hallucination engine.
3. **Get procurement right.** Determine what questions you will ask vendors, what contractual protections you will demand, and what testing you will require before any system touches production. If your vendor cannot explain how their model was trained, what happens when it fails, and who bears liability when it does, walk away.
4. **Pilot relentlessly.** Run AI-enhanced systems in advisory mode first, with recurrent red teaming, adversarial testing, and honest evaluation against ground truth. Do not promote any system to production until it has been stress-tested under conditions designed to make it fail.
5. **Plan for failure modes.** Before deployment, know how you will execute Plans B and C. Know how you will operate in a degraded mode for an extended duration. If you cannot answer these questions before you flip the switch, you are not ready to flip the switch. As prominent cybersecurity expert Josh Corman is known to say: "If you can't protect it, don't connect it."

6. **Build AI-specific incident response.** When the AI makes a bad recommendation and causes a failure, how do you diagnose it? How do you recover? How do you determine whether it was a technical error, a data quality problem, or an attack? How do you prevent recurrence? These are not questions you want to be answering for the first time during a crisis.

BUILDING FROM ENGINEERING PRINCIPLES YOU ALREADY KNOW

The good news is that water utilities do not need to invent a governance framework for GenAI from scratch. The intellectual foundation already exists, as evidenced by the definitive book on CIE published in 2025: Building Cyber Resilience in the Water Sector. Developed at Idaho National Laboratory and now being adopted across the water sector with support from the Department of Energy (DOE) and the AWWA, CIE includes twelve principles for embedding security and resilience into engineered systems from the earliest design phases forward.

CIE was built for a world where sophisticated adversaries are assumed to already have access to your networks — where the question isn't whether your systems can be compromised, but what happens when they are. That assumption maps directly onto the challenge of governing GenAI in water operations. The question isn't whether your AI systems will produce a flawed recommendation, hallucinate a maintenance history, or drift from their training baseline. They will. The question is what happens to your operations, your infrastructure, and your community when they do, and how your organization responds.

Though there are twelve CIE principles, three in particular provide the structural backbone for an AI safety governance framework that water utilities can build today.

CIE PRINCIPLE: ENGINEERED CONTROLS

In the CIE framework, engineered controls are physical or mechanical safeguards that function independently of digital systems. They cannot be hacked, corrupted, or persuaded. A pressure relief valve does not care what the SCADA system thinks the pressure is. A check valve prevents backflow whether or not the AI monitoring system is online. These are deterministic protections: they operate on the laws of physics, not the recommendations of algorithms.

When applied to AI safety governance, this principle demands that utilities identify every critical function where an AI system might influence operations and ask a deceptively simple question:

What is the engineered control that prevents catastrophe if the AI is wrong?

If an AI system optimizes chemical dosing, there must be a physical upper bound — a maximum feed rate that cannot be exceeded regardless of what the algorithm recommends. If an AI system manages pump scheduling, there must be mechanical pressure limits that prevent the kind of transient-induced fatigue described in the scenarios from the first paper in this series. If an AI system advises on valve configurations, the consequence of the worst possible recommendation must be survivable without human intervention.

This is not about distrusting the technology. It is about designing systems where trust is not required for safety. The most critical functions in a water system — maintaining safe pressure, preventing contamination, ensuring minimum flows — should be protected by engineered controls that are entirely indifferent to whether the AI is performing brilliantly or has become untrustworthy.

In practice, this means that before any GenAI system is granted the ability to influence a physical process, engineers must map the consequence space of its potential recommendations and ensure that hard physical limits bound the worst outcomes. No software guardrail, no matter how sophisticated, substitutes for a mechanical one on a function where failure means public harm.

CIE PRINCIPLE: DESIGN SIMPLIFICATION

This CIE principle holds that simpler systems are inherently more resilient, more auditable, and harder to attack. Complexity is the enemy of security because every additional component, connection, and dependency creates a potential failure point and an additional surface for exploitation.

Applied to AI safety governance, design simplification becomes the most counterintuitive — and most important — discipline a utility can adopt. The AI vendor's pitch is, by nature, additive: more sensors, more data streams, more integration points, more cloud connectivity, more analytical capability layered on top of existing operations. Each addition may be individually justified. But the cumulative effect is an exponential increase in system complexity, and with it, an exponential increase in the number of ways things can fail in ways no one anticipated.

Design simplification for AI governance means resisting the temptation to connect everything to everything. It means asking, for each proposed integration: what is the minimum viable architecture that delivers the operational benefit while preserving the ability to understand, audit, and override every decision the system makes? It means keeping AI advisory systems decisively separate from control systems, so that a corrupted recommendation cannot propagate directly into a physical action without a human checkpoint. It means limiting the number of third-party API connections, cloud dependencies, and data pathways, because each one is a potential vector for data exfiltration, model corruption, or cascading failure.

Most importantly, it means preserving operational simplicity at the human layer. When an AI system makes operations more complex for operators to understand — when the reasoning behind a recommendation is opaque, when the number of variables in play exceeds human comprehension, when the operator's role shifts from decision-maker to button-pusher — the system has become less resilient, not more, regardless of how sophisticated the algorithm is.

CIE PRINCIPLE: PLANNED RESILIENCE

TThe "Planned Resilience with No Assumed Security" principle may be the most directly applicable to AI safety governance. It requires organizations to plan for the failure of every security measure and every digital system, and to ensure that critical functions can continue when those systems are unavailable, unreliable, or are actively working against you. As Sarah Freeman, Chief Engineering at MITRE's Cyber Threat Intelligence and Modeling group, says it best:

Resilient systems (i.e., those that can endure disruption) do not succeed solely because they are digitally secure. They succeed because they are engineered to remain controllable when digital systems fail, degrade, or behave unexpectedly.

MAINTAIN MANUAL OPERATIONAL CAPABILITY

Maintain manual operational capability for every critical function that AI touches. This is where the water sector's legendary conservatism becomes a genuine strategic advantage. Many utilities still retain the ability to run their systems in manual or semi-manual mode for extended periods — a capability that their counterparts in more aggressively digitized sectors have in many cases lost. That capability must be preserved deliberately and exercised regularly, not allowed to atrophy as AI systems prove their value and operators grow comfortable relying on them.

A graduated stress testing framework captures this well: A Day without SCADA. A Week without Power. A Month without AI. These are not hypothetical disaster scenarios. They are training exercises that should be conducted routinely, with increasing duration and complexity, to ensure that your organization can function when its most sophisticated tools are unavailable. The cognitive offloading risk is real and well-documented in aviation, healthcare, and military contexts. Water utilities must learn from those sectors rather than repeat their mistakes.

BUILD AI-SPECIFIC INCIDENT RESPONSE

Build incident response plans specifically designed for AI failure modes. Traditional incident response assumes you can identify the failure, isolate the affected system, and restore normal operations. But AI failures can be subtle, cumulative, and difficult to diagnose. When the AI has been providing slightly degraded recommendations for months, what does "normal operations" even mean? When operators have been following AI guidance that has been slowly eroding system resilience, how do you establish a baseline to recover to? These questions demand answers before deployment, not discovery during a crisis.

STRUCTURE PROCUREMENT AROUND FAILURE

Structure procurement and vendor relationships around the assumption of failure. Contracts must specify what happens when the AI system produces harmful recommendations. They must require transparency into model architecture, training data, and update processes. They must guarantee the utility's right to audit, override, and disconnect. And they must address the insurance and liability questions that no current framework adequately covers: when the AI is wrong and the consequence is a $50 million main break, who pays?

FILLING THE GOVERNANCE GAPS FOR GENAI AND AGENTIC AI

Beyond these three principles, utilities must confront several governance gaps that existing frameworks leave wide open.

NEW ATTACK SURFACES

When GenAI systems are connected to OT, they create attack vectors that didn't previously exist. Adversarial attacks on AI models — prompt injection, data poisoning, and model inversion — are categorically distinct from traditional cyber threats and require different controls. This is precisely where CIE's full twelve-principle framework plays a critical role.

AI-SPECIFIC FAILURE ATTRIBUTION

When an AI system makes a consequential wrong recommendation, utilities need a clear process to determine whether the failure was a technical error, a data issue, a model drift problem, or a deliberate attack. Without that diagnostic capability, you cannot fix what broke, and you cannot determine who bears responsibility. Today we are very far from having the ability to do this reliably.

THIRD-PARTY DEPENDENCY RISK

Routing utility-critical decisions through third-party AI systems creates an unacceptable single point of failure. When those systems are unavailable or compromised — and they will be — the consequences cascade through your operations. Plan accordingly.

AGENTIC AI IN OT ENVIRONMENTS

AI agents that directly issue control signals to OT — adjusting pump schedules, modifying chemical dosing, or reconfiguring valve positions without human intervention — represent a categorically different governance challenge than AI systems that draft reports or summarize data. The accountability and oversight requirements must be proportionally more rigorous.

A qualified human operator must always be in the loop for any AI action that can affect public health or safety.

The International Society for Automation's authoritative briefing "AI Risks to Critical Infrastructure" states this clearly:

"Generative AI is not permissible for autonomous control in high-consequence control systems. However, it may be acceptable in these systems where a human in the command loop supplies the ultimate decision on an action. The indeterministic nature of generative LLM models makes the variability of GenAI responses unacceptable for autonomous actions."

Also be on the lookout for the Water Research Foundation's Project 5394, titled "Evaluating Scalability, Reproducibility, & Impact of GenAI & Agentic AI in the Water and Wastewater Sector," which will focus on how utilities may overcome barriers to adoption of GenAI and agentic AI, including "actionable guardrails."

THE GUARDRAILS ARE YOURS TO BUILD

The three CIE principles — engineered controls, design simplification, and planned resilience — do not constitute the entire governance framework a utility needs. Nine additional CIE principles, from consequence-focused design to cybersecurity culture, each have direct applications to AI safety governance. But these three establish the essential posture: protect the critical functions with physics, not just software; keep the system simple enough to understand and override; and plan for the day when everything fails.

The deeper insight of CIE, and the reason it translates so powerfully to AI safety governance, is that it treats security and resilience as engineering problems first and technology problems second. You do not solve an engineering problem by buying a better firewall, and you do not solve an AI governance problem by buying a better AI. You solve it by understanding the consequences, simplifying the design, and building in the ability to survive the worst case. Water engineers have been doing exactly this for more than a century. The challenge now is to apply those same disciplines to the newest and most powerful tools arriving at the plant gate.

The economics are irresistible. The technology is here. The regulatory cavalry is not coming. The vendors will tell you what their systems can do; it is your job to determine what happens when they don't function as advertised. The insurance industry hasn't caught up. The liability frameworks don't exist. The only entity with both the obligation and the ability to protect your ratepayers is you.

Resilience is now one of water utilities' most critical assets. You are the one who must build it and maintain it. And when it comes to GenAI, remember:

YOU ARE THE GUARDRAILS.

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