Anomaly Detection

Recognizing Emerging Loss Conditions in Time to Act

Dr. Ron Ross
CEO, RONROSSECURE, LLC
ISTS FELLOW, DARTMOUTH COLLEGE

Abstract

Anomaly Detection is a foundational design principle for trustworthy secure systems. NIST SP 800-160, Volume 1, Revision 1 calls for timely detection of salient anomalies in the system or its environment so that effective response is possible. Applying the principle requires engineers to establish a basis for correctness, collect data that reveal relevant conditions, and interpret those data in the context of system state and mode. This article uses a water treatment example to show how engineers derive detection and response requirements from unacceptable losses, protect the evidence on which decisions depend, and evaluate the design under assumed compromise. It connects Anomaly Detection to the security design principles of Protective Failure and Protective Recovery—helping to ensure that the system can sustain required mission functions under adversity.

1. The Engineering Problem

A system can continue to report normal status while the conditions that justify its operation deteriorate. A controller may acknowledge a valid command even though the physical process responds incorrectly. A redundant service may remain online after losing the independence that allows it to tolerate a failure. A protective mechanism may still be installed even though a change has made it ineffective. In each case, the mission can appear unaffected until the system needs a protection it can no longer rely on.

The design principle of *Anomaly Detection* addresses the interval between a meaningful departure from design intent and the point at which corrective action ceases to be effective. Systems engineers need to know what could go wrong, what evidence would reveal it, and how much time remains to act. A detailed diagnosis delivered after unsafe output reaches a customer has forensic value. A timely indication that the process can no longer be trusted may give an operator enough time to hold that output and preserve service elsewhere.

> ***"The time to detect a loss condition is set by the last effective response."***

The problem occurs in information technology, operational technology, and cyber-physical systems. A transfer service can report that a file was delivered while the receiving application works from an older version. A spacecraft can return valid telemetry while a temperature trend points toward the loss of a required function. Engineers must identify the difference between the success of an individual operation and the outcome the mission requires.

Architecture diagrams often show where status is collected and omit what the status can establish. Several displays can repeat the same measurement from one controller. Agreement among those displays offers little additional evidence about the physical process. The architecture must show where observations originate, which system elements can change them, and which conclusions depend on their continued integrity and availability.

Ordinary operational changes complicate interpretation. Maintenance may legitimately suspend a sensor, startup may produce a transient, and degraded operation may permit a different sequence of actions.

© 2026 RONROSSECURE, LLC | CC BY 4.0.

Engineers must define the expected behavior for these modes and for transitions between them. A design that treats every deviation from steady state as suspicious can overwhelm operators precisely when careful judgment matters most.

Frontier AI adds a related systems engineering problem. An agent can complete a task and produce a plausible summary while exceeding its authority or omitting a required approval. A system should capture consequential tool actions and their effects at the interfaces that enforce them. Where the agent's report and those observations disagree, the difference can warrant investigation before the system relies on the resulting state.

The engineering work begins with the conditions that must remain true for the mission and its protections to continue. From those conditions, engineers derive what needs to be observable, where evidence must be collected, when it must be interpreted, and who or what has authority to respond.

2. What the Principle Means

NIST SP 800-160, Volume 1, Revision 1 states the design principle of *Anomaly Detection* as follows [1]:

> ***"Any salient anomaly in the system or its environment is detected in a timely manner that enables effective response action."***

Salience is a judgment about consequence. A deviation matters when it indicates that an unacceptable loss has occurred, may be developing, or may follow if the condition persists. A small change in a critical measurement can warrant immediate attention. A much larger change during a planned transition may be expected. The basis for deciding is the design intent for the current system state and operating mode.

Timeliness depends on the response and the progression of the loss condition. Early detection may allow graduated action, such as confirming a measurement, reducing operation, or isolating one function before a broader interruption becomes necessary. If a person must make the decision, the design must allow time to receive the indication, interpret the evidence, choose an authorized action, and see that action take effect.

NIST identifies a basis for correctness, data collection, and data interpretation as necessary to determine an appropriate response [1]. The fourth item below makes the timing relationship explicit as an engineering application of the principle.

- **Basis for correctness.** Define the expected behavior, outcomes, and interactions for each relevant state and mode. For conditions outside the system, establish the environmental bounds within which it is designed to operate.
- **Data collection.** Capture observations of behavior, outcomes, and environmental conditions with sufficient accuracy, coverage, and currency. Retain the origin and timing of observations when those details affect a decision.
- **Data interpretation.** Compare observations with the expected condition for the current mode. Distinguish an observed departure, a trend toward a limit, and insufficient evidence to establish the condition.
- **Time for response.** Work backward from the latest point at which action can be effective. Allocate time for observation, interpretation, decision, action, and the physical or logical effect of the action.

A basis for correctness can be an allowed sequence of commands, a process relationship, an approved configuration, or a requirement that a protection remain available. Statistical patterns can help describe

© 2026 RONROSSECURE, LLC | CC BY 4.0.

expected behavior when normal variation matters. The design intent provides the basis for deciding which deviations merit action.

Monitoring also extends to the environment [1]. An upstream change in raw water quality may threaten the treatment process while the facility is still operating within its current limits. A trend in that external condition can justify preparing a response before the plant experiences a departure. Environmental observations and process observations support different conclusions and may require different sources.

"Observe the condition that matters and leave enough time to act."

The fidelity of the evidence must match the consequences of the response. A questionable reading may justify requesting confirmation. Isolating an unaffected process line on the same evidence could interrupt required service. Attributing a condition to a particular person or system element requires evidence suitable for that decision. Immediate protective action can sometimes proceed while the cause remains under investigation, provided the evidence and authority support the action.

3. What the Principle Does Not Mean

Anomaly Detection can be confused with collecting more data, finding attacks, or generating alerts. Its scope is defined by the deviations that matter to the system and by the responses those deviations warrant.

- **Salience depends on the loss.** A rare observation can be harmless, while a frequently recurring deviation can threaten a required function. Frequency can inform interpretation when the design explains its relationship to the loss.
- **Detection and diagnosis serve different decisions.** A facility can recognize that treatment output requires a protective hold before it establishes whether the underlying cause is equipment failure, bad calibration, or hostile action.
- **Alert delivery is one step in the response.** A timely notification is useful when the recipient can interpret it, has authority to act, and has sufficient time to complete the action under realistic operating conditions.
- **A learned baseline needs an engineering basis.** Repeated behavior can reflect gradual deterioration or an existing compromise. Compare learned patterns with approved operating limits and validate changes to the baseline against the mission and protection requirements.
- **An indication has a defined scope.** An instrument reports a property at a particular point under specified conditions. The assurance argument must establish how that observation supports conclusions about the process, especially during transitions or when an instrument is unavailable.
- **Missing evidence requires interpretation.** An outdated value or interrupted status stream can leave a condition unknown. Define when loss of visibility warrants confirmation, a protective response, or continued operation under restrictions.
- **Automated action requires separate justification.** A system detector may have adequate evidence to request an operator review. A response that changes control authority or interrupts service requires appropriate confidence, independent constraints where needed, and a defined recovery path.

A detection capability can combine simple checks with more detailed analysis. A configuration comparison can establish whether a measured value differs from an approved version. A process model can expose an inconsistency among several measurements. Systems engineers should use a method whose assumptions and limitations can be explained and evaluated for the decision it supports.

© 2026 RONROSSECURE, LLC | CC BY 4.0.

Detection and response also operate at different levels. A device can identify a local fault while system-level observations establish whether the fault threatens a specific mission function. Where a local safeguard acts promptly, higher-level monitoring can still help assess remaining service, coordinate maintenance, and confirm that protection continues elsewhere.

Operators need to distinguish uncertainty about the evidence from uncertainty about the underlying process. A plausible but delayed reading has different implications from a current reading outside the approved limit. Preserve source, time, and quality information in the display and records so a person can decide how much confidence to place in the indication.

4. Loss-Driven Interpretation

Anomaly Detection begins with the losses stakeholders consider unacceptable. NIST's systems security engineering approach connects those losses to protection needs and to requirements for system behavior under adversity [1]. Engineers work backward from the loss to the conditions that could produce it, then establish which conditions the design can observe early enough for an effective response.

Consider a water treatment facility with two process lines. Each line must maintain required local control. Supervisory applications request changes within approved limits and provide status for system operators. Stakeholders identify two unacceptable outcomes. The first is release of unsafe treated water. The second is loss of minimum required treatment service across both lines from a disturbance the design is intended to contain to one line. The design needs observations that reveal deteriorating treatment on the affected line and observations that establish whether the other line can continue within its approved operating envelope.[1]

The supervisory display alone cannot establish both conditions. Suppose a line reports that a dosing change completed, while independent measurements indicate that the process response is inconsistent with the change. An operator may have time to hold the affected output and confirm the discrepancy. That response depends on the facility's design, including how long affected water takes to reach the release point and how quickly the hold can take effect.

The following engineering activities connect unacceptable losses to the conditions that must be observed and the responses available.

- **Identify the conditions that can lead to loss.** Trace unsafe output and service interruption to specific process departures, failures of protection, or loss of required observations. Include relevant conditions in the environment.
- **Establish what can be observed.** Identify direct measurements and relationships among observations that provide evidence of a developing loss. Determine which sources are available in each mode and which can be affected by the same failure.
- **Define the response window.** Estimate when corrective action must take effect and what reduces the time available. Include sampling delays, operator decision time, equipment movement, and process transport time.
- **Specify the response and authority.** Decide who or what can confirm, hold, isolate, reduce operation, or restore service. Analyze whether the response itself can create an unacceptable loss.

[1] An approved operating envelope specifies the range of conditions within which a process is authorized to operate. It can include limits on process values, rates of change, and allowed sequences. The applicable limits depend on facility-specific engineering and validation.

© 2026 RONROSSECURE, LLC | CC BY 4.0.

- **Compare feasible designs.** Evaluate alternative observation points, measurement methods, and local protective actions. Where a useful warning cannot be obtained in time, reconsider the process or protection architecture.

The two losses may require different actions. Holding output from one line may be appropriate for a credible indication of unsafe treatment, while stopping both lines could violate the minimum-service requirement. The facility must analyze whether the unaffected line can sustain that service and whether the hold has consequences for shared equipment. A local action is only as local as its actual effects.

Slow departures deserve the same attention as abrupt failures. A gradual change in a sensor or a process relationship can erode the margin available for corrective action. Establish how long a trend must persist to justify intervention and what evidence distinguishes drift from an expected change. A persistence rule consumes time and belongs in the response-time analysis.

For each loss condition, the requirements should identify the earliest credible indication, the latest effective response, and the assumptions that connect them. If the interval is too short for a staffed response, engineers can consider a local automatic protective action or a process change that creates more time. The choice must follow from the loss analysis and be validated for the facility.

5. Engineering the Principle into the Architecture

The architecture allocates observation, interpretation, decision, and response to system elements. It also identifies the dependencies that allow each function to operate. A detector that relies on a supervisory application for measurements, timing, and configuration inherits that application's limitations. Engineers need to determine which observations remain credible when a system element involved in the loss is also involved in reporting it.

Begin with a model of expected behavior and the conditions that support it. Define the approved operating envelope and the relevant modes, including startup, maintenance, degraded operation, and recovery. Identify which conditions apply continuously and which conditions depend on a mode or transition. For each condition, specify the evidence required to distinguish satisfactory operation, a departure requiring action, and an unknown state.

The age and origin of an observation warrant particular scrutiny. A correct value at the time it was measured can be misleading at the time it is used. Define the maximum acceptable age[2] at the decision point and determine whether the source time can be trusted. Consider late arrival, repeated values, clock error, and loss of synchronization. When timing cannot be established, the design should treat the observation according to its defined limits.

Place observations where they can establish the relevant outcome. In the treatment facility, a record of a supervisory request establishes what was requested. A local record can establish what was accepted or applied. A process measurement can provide evidence of what occurred in the water. The design should identify the conclusions each supports and whether the selected combination is sufficient for the intended action.

The following engineering activities help allocate and protect the detection and response functions in the architecture.

[2] The age of an observation is the elapsed time from measurement to use in a decision. The age includes time spent in transmission, buffering, and processing. A newly received value can describe a condition that no longer exists.

© 2026 RONROSSECURE, LLC | CC BY 4.0.

- **Specify observations at their source.** Define the property, measurement point, units, accuracy, currency, and validity information. Trace each observation to the condition it is intended to reveal.
- **Protect the observation path.** Limit access to measurements and records. Control changes to collection and reporting functions. Determine how loss, delay, corruption, or uncertainty in the evidence will be recognized.
- **Allocate capacity for critical observations.** Bound collection, transmission, and processing delays under expected demand and adversity. Reserve enough capacity to preserve time-critical indications when less important reporting is degraded.
- **Choose independent evidence where required.** Identify common instruments, power, software, clocks, and administration. Use independent observations where the design must remain effective after a specified shared dependency fails or is compromised.
- **Constrain response authority.** Define which conclusions can initiate each action. Allocate authority and confirmation according to consequence, and preserve local protections during remote or automated response.
- **Provide for loss of observation.** Establish when the system can continue under restrictions and when it must enter a protective state. Ensure the response remains available when the specified observation path is lost.

A detection rule should identify the system mode, the evidence it uses, the condition it recognizes, and the meaning of its output. For example, an inconsistency rule can compare applied dosing and process response after allowing for a known delay. A valid implementation still depends on the delay being appropriate for the operating mode and on the observations arriving within their specified limits.

Thresholds involve engineering tradeoffs. A sensitive threshold may offer earlier warning while increasing unnecessary interventions. A less sensitive threshold can consume the time available to act. Evaluate plausible process variation, measurement error, transition behavior, and the consequences of wrong responses. Where graduated action is feasible, the design can support an early advisory and reserve an interruption for stronger evidence.

A common monitoring service may improve coordination and records, but local action can have a shorter response window than that service can support. Where the loss analysis requires local detection or response, allocate the necessary evidence and authority to the line. Test whether it retains those capabilities when the monitoring service is unavailable and whether later coordination can occur without relaxing local protection.

For AI-enabled functions, record evidence at consequential boundaries. If an agent proposes a change, the approval record, the operation accepted by the interface, and the resulting state provide different evidence. The architecture can compare those records to identify a skipped approval or an unexpected effect. Evaluate automated interpretation against the same timing, confidence, and authority requirements applied to other detection mechanisms.

Detection itself has a cost. Additional instrumentation requires calibration and maintenance. More observations can increase communications demand and complicate operator displays. Detection requirements should justify each consequential data source and identify how its continued fitness will be assessed. A smaller set of dependable observations can serve a loss-control objective more effectively than a larger collection with uncertain provenance.

“A second display adds confidence only when it brings independent evidence.”

© 2026 RONROSSECURE, LLC | CC BY 4.0.

6. Engineering for Compromise

Engineering for Compromise assumes selected system elements can fail, lose integrity, or act maliciously and asks which security properties the system can still substantiate [2]. For the principle of *Anomaly Detection*, the question extends to the assumptions that support continued operation. Engineers identify which protection conditions must remain observable after compromise and which decisions can still be justified from the remaining evidence.

In the treatment facility, assume that one supervisory application is compromised and continues to display ordinary status. Determine whether local records and independent process observations still reveal a consequential departure within the required time. Evaluate whether the affected line can hold its output using local authority, and whether the other line can establish its own ability to sustain required service.

Now assume a designated observation source loses integrity or becomes unavailable. The response depends on which sources remain, what they can establish, and how much time is left. An independent indication may support continued restricted operation. If the remaining evidence cannot establish the process condition, the design must specify the protective action that is appropriate for that uncertainty.

Compromise of the observed function and compromise of the detection function are distinct cases. NIST distinguishes self-anomaly detection, in which an entity detects anomalies within the scope of its design without depending on another entity, from dependent anomaly detection, in which an entity relies on another entity for some or all of the assessment [1]. A monitor can correctly report the behavior of an untrustworthy controller. A compromised monitor can misrepresent the behavior of an uncompromised controller. If both cases are in scope, analyze the observations and authority that survive each case. State any dependency whose loss defeats the detection claim.

The following compromise scenarios help determine whether the system can still recognize loss conditions and support appropriate response when the evidence or the functions that produce it are in doubt.

- **Compromised observed function.** Evaluate whether independent observations still reveal unacceptable effects when the function's status reports cannot be trusted. Confirm that local action remains available under the stated assumptions.
- **Compromised observation source.** Determine whether invalid, outdated, or inconsistent evidence is recognized. Evaluate how interpretation proceeds when a source cannot be used and whether a required decision remains supportable.
- **Compromised collection or interpretation mechanism.** Assess which local indications and protective actions remain available if a central monitor loses integrity or availability. Identify any decision the central monitor alone can influence.
- **Compromised administrative authority.** Determine whether one authority can change the expected baseline, suppress an indication, and remove the record of the change. Assess independent approval and records where the consequences warrant them.
- **Compromised response or recovery function.** Determine whether a requested action took effect, whether it remained within its authorized scope, and whether evidence of restored protection can be established independently of the function being restored.

Loss of a monitor can itself be a salient anomaly. Establish which protection claims depend on visibility and how long the system may continue without it. The loss of a convenient display may be tolerable if local safeguards and observations remain available. Loss of the only credible indication of a developing process hazard may require immediate protective action.

© 2026 RONROSSECURE, LLC | CC BY 4.0.

Recovery requires observations suited to the decision to return to service. A cleared alarm can follow from a reset or from a return to an acceptable condition. Before releasing held output, confirm that the process is within approved limits, that necessary measurements are credible, and that required protections are again effective. The evidence should address the reason for the hold and any relevant effects of the recovery work.

Anomaly Detection can reveal that a condition supporting the security argument has changed. Examples include loss of a required access mediation mechanism, an unapproved configuration, or loss of a recovery source. This is the *Engineering for Compromise* emphasis on observing the basis for continued trust [2]. The assurance case should identify which conditions are observable and what evidence is required before a claim can be relied on again.

7. Interaction with Other Principles

Anomaly Detection provides evidence that a consequential condition exists or is developing. Other principles help keep that evidence credible and turn it into action that controls loss. The relationships below follow the evidence from its source through the decision to continue, restrict, or restore operation.

The following relationships show how complementary security design principles support observation, interpretation, and response under the conditions for which the system is designed.

- **Domain Separation.** Separate functions with different protection needs. Where the same compromise could affect a process and the evidence about it, allocate independent observations or protective authority across an appropriate boundary.
- **Mediated Access.** Enforce policy on access to observations, baselines, detection rules, and response interfaces. Records at a mediated interface can show what operation was requested and which decision was enforced.
- **Least Privilege.** Limit the authority of collection, analysis, and response functions to their assigned work. A service that reports process state can often do so with authority limited to observation and reporting.
- **Least Functionality.** Provide the functions required to observe, interpret, and respond. Evaluate administrative and remote functions in monitoring components for the influence they introduce over the system being observed.
- **Least Persistence.** Limit the duration of temporary monitoring privileges, maintenance exceptions, and access to sensitive evidence. Set evidence retention according to the decisions and accountability obligations the records must support.
- **Distributed Privilege.** Allocate independent approvals where a consequential change to monitoring or a broad protective action requires more than one authority. Account for any approval delay in the response-time requirement.
- **Protective Failure and Protective Recovery.** Define the protective states available when an anomaly or loss of visibility warrants action. Establish observations that confirm the action took effect and provide the basis for restoring required functions.
- **Compositional Trustworthiness.** Evaluate the assembled system’s detection and response properties. Individually dependable system elements may rely on a common source or consume more response time in combination than the mission can allow.
- **Substantiated Trustworthiness.** Support the detection claim with evidence about coverage, timing, confidence, and the conditions under which the response remains effective. Identify what changes would invalidate the supporting evidence.

© 2026 RONROSSECURE, LLC | CC BY 4.0.

The principle of *Trustworthy System Control* also matters where observation leads to a command. The mechanism enforcing the command must constrain the action to its authorized scope, even when the request comes from an approved detector. In the treatment example, the authority to hold one line's output should be assessed separately from authority to interrupt both lines. A detector's conclusion and the protection mechanism's enforcement decision serve different purposes.

NIST SP 800-160, Volume 2, Revision 1 provides complementary cyber resiliency techniques [3]. *Analytic Monitoring* supports observation and analysis of system properties and behaviors. *Contextual Awareness* connects observations to mission dependencies and system status. *Coordinated Protection* and *Adaptive Response* can help align the actions of different system elements as conditions change. These techniques inform implementation choices. The design principle establishes the engineering obligation to recognize salient departures early enough for an effective response.

8. System Life Cycle Application

Anomaly Detection is a life cycle obligation. The system life cycle processes in NIST SP 800-160, Volume 1, Revision 1 provide a means to translate loss concerns into detection and response requirements, allocate them to the architecture, realize the mechanisms, and maintain the evidence as the system and its environment change [1].

- **Business or Mission Analysis.** Identify unacceptable losses, required service, and the situations in which early recognition can change the outcome. Determine the consequences of failing to act and of acting incorrectly.
- **Stakeholder Needs and Requirements Definition.** Establish expectations for warning, decision authority, continuity of service, and attribution. Identify the operating situations in which people must be able to assess an indication and respond.
- **System Requirements Definition.** Specify the conditions to detect, acceptable uncertainty, applicable modes, and the latest effective response. Allocate measurable limits for observation, reporting, decision, and action.
- **Architecture Definition.** Allocate observation, interpretation, and response to system elements. Identify the boundaries and dependencies that determine whether evidence and response remain available under specified adversity.
- **Design Definition.** Define measurement points, data quality, mode-dependent rules, timing checks, thresholds, and action interfaces. Specify what happens when evidence is missing, inconsistent, or insufficient for the intended decision.
- **System Analysis.** Evaluate detection coverage, time to effect, common dependencies, and the consequences of missed or incorrect indications. Examine how protective actions interact with minimum-service and safety requirements.
- **Implementation.** Realize the observation paths, detection logic, decision interfaces, and protective actions. Enforce the approved authorities for modifying thresholds, modes, and baselines.
- **Integration.** Verify that units, time references, status definitions, and assumptions agree across interfaces. Examine delays and dependencies that become apparent only when system elements are assembled.
- **Verification.** Demonstrate specified detection and timing across representative modes and loads. Exercise departures, missing and outdated evidence, erroneous indications, and the specified loss or compromise of an observation source.

© 2026 RONROSSECURE, LLC | CC BY 4.0.

- **Transition.** Establish the approved baseline and confirm that instruments, clocks, procedures, and response authorities are ready for use. Remove commissioning exceptions or explicitly bound their use.
- **Validation.** Determine whether the realized design enables effective action in representative mission conditions. Include operators, communications, protective equipment, and the process response in the evaluation.
- **Operation and Maintenance.** Track changes in process behavior, evidence quality, detection performance, and response time. Reassess calibration, thresholds, authority, and procedures after consequential changes.
- **Disposal.** Remove retired instruments and reporting paths from monitoring and decision logic. Revoke associated access, preserve required records, and verify that retained functions can still establish the conditions they depend on.

The *Configuration Management* process preserves the relationship between approved modes, baselines, thresholds, and response logic. The *Information Management* process preserves the meaning, origin, and retention of observations and evidence. Together these two processes make it possible to explain which configuration produced an indication and to reassess whether historical evidence still supports the current design.

The *Acquisition* process should establish what supplier-provided status can actually demonstrate. Ask which measurements it represents, how often it updates, what conditions invalidate it, and what dependencies support it. Obtain enough information to evaluate the consequences of lost or incorrect status. Supplier descriptions of a monitoring feature should be traceable to observable behavior and the intended mission decision.

Change analysis must cover the response as well as the detector. A replacement sensor can change measurement delay. A revised operating mode can change normal process behavior. A remote staffing arrangement can change decision time. Each can invalidate timing or interpretation evidence even when the detection rule itself stays the same.

9. Failure Patterns and Anti-Patterns

Detection failures often arise from a mismatch between the evidence a system collects, the conclusion a decision maker draws from it, and the time available to act.

- **A threshold lacks a loss basis.** A limit is inherited from a product default. Determine which loss condition it reveals and whether the warning leaves time for an effective response.
- **Controller availability is mistaken for correct process operation.** The interface shows that a controller is reachable. Verify separately whether the process is within its approved envelope and required protection is available.
- **Several indications share one source.** A dashboard, report, and alarm agree because each uses the same measurement. Establish the common dependency before treating agreement as corroboration.
- **Silence is interpreted as normal.** A stopped status stream leaves the last valid value on the display. Show observation age and define the action required when the source is unavailable.
- **Steady-state rules govern every mode.** Startup, maintenance, or recovery produces expected transitions that resemble faults. Define the permitted behavior, evidence, and protection requirements for those modes.

© 2026 RONROSSECURE, LLC | CC BY 4.0.

- **Confirmation consumes the available time.** A rule requires repeated observations or sequential approvals. Verify that the total time to protective effect fits the fastest credible progression of the loss condition.
- **An alert has no decision owner.** An indication reaches a shared queue without established response authority. Assign ownership and demonstrate the handoff under representative staffing and communications conditions.
- **One response can spread the loss.** A local anomaly initiates a facility-wide interruption. Analyze the action's effect on unaffected functions and constrain its scope to the justified protective need.
- **A baseline follows an unsafe change.** An adaptive model gradually accepts a deteriorating condition. Use approved limits, independent checks, and reviewed changes to preserve the engineering basis for correctness.
- **Monitoring competes with essential functions for resources.** Collection or analysis consumes capacity required by the mission. Evaluate resource demand and protect essential functions during normal monitoring and high event volume.
- **Clearing an alarm substitutes for recovery evidence.** An indication disappears after restart while the original condition remains uncertain. Require evidence that the process and its protections satisfy the criteria for return to service.

These patterns can combine. A common status source may stop updating during maintenance, leaving system operators with plausible values. If the system has no indication of observation age and treats an acknowledged command as evidence of the physical outcome, the discrepancy may become visible only after the time for a local response has passed. NIST cautions that poorly designed lines of defense can conceal emerging dangerous system states and conditions, especially from human observers, and that the design must minimize the difference between estimated and actual system states and conditions [1].

> ***"A normal indication is useful only when the evidence behind it remains credible."***

Operational pressure can weaken detection gradually. A noisy rule may be suppressed, a temporary exception may become routine, or an instrument may remain out of service longer than assumed. Review recurring exceptions against the loss and timing analyses. If the design depends on visibility that operations can no longer maintain, revise the monitoring arrangement, the permitted operation, or the assurance claim.

10. Evidence and Assurance

Anomaly Detection claims require evidence that the system recognizes salient anomalies with sufficient confidence and timeliness for effective response. Establish the scope of the claim before selecting tests. Identify the relevant losses, modes, environmental conditions, assumed failures or compromises, and protective actions. Specify what can be concluded from available evidence and which cases require a different design or a narrower claim.

A top-level claim might be:

> ***Within the defined system scope, operating modes, and environmental bounds, salient anomalies in the system or its environment of operation are detected with sufficient confidence and time to enable effective response action, including under the failure and compromise conditions specified in the assurance scope.***

© 2026 RONROSSECURE, LLC | CC BY 4.0.

The claim concerns detection of salient anomalies and the ability to respond effectively in time. Its scope depends on the documented system models, response authority, availability of response mechanisms, and justified assumptions about the independence of observations. The assurance argument must connect those assumptions to evidence and account for the effects of response actions on required mission functions.

The following supporting subclaims organize the assurance argument and identify the kinds of engineering evidence needed to support it.

- **Subclaim 1: Salient anomalies and applicable conditions are defined.** The loss analysis identifies the salient anomalies within the assurance scope and traces them to unacceptable losses and required mission or protection outcomes. System models identify relevant states, modes, environmental conditions, transitions, and expected behavior against which observations are evaluated.
- **Subclaim 2: Observations support the decisions.** Instrumentation specifications, calibration records, interface definitions, and dependency analyses establish what is measured and whether observation quality, origin, and timing are adequate for the intended response.
- **Subclaim 3: Interpretation is correct for the specified conditions.** Analysis and verification demonstrate that decision logic recognizes required departures and distinguishes relevant normal variation, degraded observation, and uncertain results. Evaluation addresses missed indications and incorrect indications.
- **Subclaim 4: Response can take effect in time.** Timing analyses and demonstrations of the complete detection and response sequence establish that authorized response actions can achieve their intended protective effect within the required time. The evidence accounts for observation, communication, interpretation, decision, action, and resulting changes in system conditions, including effects on required mission functions.
- **Subclaim 5: Detection and response remain effective under specified adversity.** Dependency analyses and tests demonstrate that the required detection and response capabilities remain effective under the failure and compromise conditions specified in the assurance scope. The evidence identifies limitations, assumptions, and dependencies.
- **Subclaim 6: Protection is maintained through change and recovery.** Configuration records, change analyses, and recovery demonstrations establish that approved changes preserve the detection and response basis and that return to service requires evidence of restored protection.

Verification should exercise each defined condition at the point where it can first be observed and at the point where response time is most constrained. Vary the phase of occurrence relative to sampling and reporting. Include expected processing and communications loads. Measure time to recognition and time to protective effect separately so the engineering team can locate a timing shortfall and adequately assess its consequences.

Evidence about detection performance should include representative conditions near decision boundaries. Test permitted transients, degraded modes, missing data, and observations that disagree. Record the circumstances of missed and incorrect indications and assess them against the consequences of delay or wrong action. Average performance alone provides little evidence for a rare, time-critical case unless the evaluation covers that case.

Where operators participate, validation should include the decisions they actually face. Give them the information available through normal interfaces and use realistic staffing, workload, and communications conditions. Observe whether they can distinguish an unknown condition from a confirmed departure, select an authorized response, and complete it in time.

© 2026 RONROSSECURE, LLC | CC BY 4.0.

In the water treatment example, assume the facility must detect a specified divergence between reported treatment and independent process observations on one line despite compromise of its supervisory application. The engineering evidence should identify the independent measurements, show how their timing and quality support the conclusion, and establish which system elements can authorize and carry out a hold. The facility must derive the numerical limits and response times from its process engineering and validate them for the applicable operating conditions.

For the unsafe-output loss, demonstrate that the departure is recognized and that the designated action takes effect before affected water reaches the release point. Include plausible measurement delay and equipment response. Record the achieved process state as well as the indication and command. A successful command transmission is evidence of an intermediate step in the argument.

For the minimum-service loss, demonstrate that the action remains within the affected line's scope and that the other line can sustain the required service under the stated conditions. The evidence should cover relevant shared resources and the measurements on which the unaffected line relies. Include the effect of any common reporting failure that falls within the scope of the claim.

Recovery testing should confirm that the affected line can be evaluated and restored while the other line continues its assigned function. The test should establish that temporary measures are removed or controlled, that measurements support the release decision, and that the required detection and protection functions are effective in the restored configuration.

> ***"Assurance follows the anomaly from its first credible indication to the effect of the response."***

The assurance argument links claims, reasoning, evidence, assumptions, and limitations. Make clear which parts of the argument are supported by analysis, which are demonstrated in tests, and which depend on facility conditions established during validation. If a single dependency can invalidate the expected behavior, the observations, and the response, its effect belongs in the claim scope and in the loss analysis.

Maintain the argument as the system evolves. A new collection service, modified threshold, replacement sensor, or revised operating procedure can change coverage and timing. Trace each change to the claims and evidence it affects. Use the results to determine which analyses, verification activities, and validation exercises need to be repeated before relying on the revised design.

11. Key Engineering Takeaways

The following takeaways summarize the engineering decisions and assurance obligations needed to apply the design principle of *Anomaly Detection*.

1. Derive detection requirements from unacceptable losses and the conditions that can cause or reveal them.
2. Specify expected system behavior and environmental bounds for relevant modes, including transition, degradation, maintenance, and recovery.
3. Apply NIST's three aspects of anomaly detection. Establish a basis for correctness, collect adequate data, and interpret them for the decision at hand.
4. Allocate time from the latest effective response backward through actuation, decision, interpretation, reporting, and observation.
5. Assess the independence, integrity, and currency of evidence. Treat insufficient visibility according to its consequences for the mission and its protections.

© 2026 RONROSSECURE, LLC | CC BY 4.0.

6. Use evidence and response authority commensurate with the consequences of the action. Establish attribution rigor separately where the decision requires it.
7. Evaluate detection and response under *Engineering for Compromise* assumptions, including loss of integrity in the observed function, the observation source, or the monitor.
8. Compose *Anomaly Detection* with *Trustworthy System Control*, *Protective Failure*, and *Protective Recovery* so observations support constrained action and substantiated return to service.
9. Preserve the basis for correctness and the response window through acquisition, integration, maintenance, configuration changes, and recovery.
10. Substantiate coverage, confidence, timeliness, and response effect within the defined assurance scope. State the dependencies and conditions that limit the claim.

12. Conclusion

Anomaly Detection makes early recognition of consequential departures an engineering obligation. The obligation applies to the behavior and outcomes of the system and to conditions in its environment. It begins with the losses stakeholders need to control and leads to specific requirements for what can be observed, what can be concluded, and what action can take effect in time.

NIST provides the foundation in a basis for correctness, data collection, and data interpretation [1]. Engineers apply these aspects to actual states, modes, and transitions. They choose observation points that establish the conditions of interest, protect the evidence, and allocate the time and authority needed for a proportionate response.

Engineering for Compromise extends that work to the conditions on which continued trust depends [2]. Assume selected system elements lose integrity or availability, then determine what remains observable and what protective action remains possible. The answers identify where independence is required and where the assurance claim must depend on explicit assumptions.

Producers should make those assumptions and limits visible. Consumers should ask which conditions the system can recognize, how quickly it can recognize them, and what response is possible when the evidence is uncertain. Both should be able to trace the answers to the mission outcomes and protections they depend on.

A trustworthy secure system gives its operators and protective mechanisms sufficient time to act on credible evidence. Applying the principle of *Anomaly Detection* throughout the system life cycle helps preserve that capability as operating conditions, system elements, and sources of adversity change.

“Recognize the departure while there is still time to protect the mission. Substantiate the response.”

Acknowledgments

The author gratefully acknowledges the support provided by Dartmouth College’s Institute for Security, Technology, and Society (ISTS). The resources provided by the ISTS contributed to the research and development of this paper. The views and conclusions expressed herein are those of the author and do not necessarily represent the views of Dartmouth College.

© 2026 RONROSSECURE, LLC | CC BY 4.0.

References

[1] R. Ross, M. Winstead, and M. McEvilley, Engineering Trustworthy Secure Systems, NIST Special Publication 800-160, Volume 1, Revision 1, National Institute of Standards and Technology, November 2022. https://doi.org/10.6028/NIST.SP.800-160v1r1

[2] R. Ross, "Engineering for Compromise: Designing Systems That Remain Trustworthy Secure Under Adversity," RONROSSECURE, LLC, August 2026. https://ronrossecure.com/library

[3] R. Ross, V. Pillitteri, R. Graubart, D. Bodeau, and R. McQuaid, Developing Cyber-Resilient Systems: A Systems Security Engineering Approach, NIST Special Publication 800-160, Volume 2, Revision 1, National Institute of Standards and Technology, December 2021. https://doi.org/10.6028/NIST.SP.800-160v2r1

© 2026 RONROSSECURE, LLC | CC BY 4.0.