The AI Failure Mode Register

The documented ways AI fails at quality work, the control that catches each one, and the level of control each demands

Compiled from months of trying to break AI systems on real root cause and corrective action work. The point is not that AI is unusable. It's that AI is *unqualified until proven*, like any new process. Every mode below appeared in real testing. Every one has a control, and every control has a level.

The three control levels. Level 1, Procedure: instructions the model is asked to follow. Weakest. Revalidate fully on every model change. **Level 2, Detection:** code that checks the output: gates, detectors, flagging inside the record. The rule survives a model change, but its calibration does not; prove it again with a challenge suite. **Level 3, Structure:** the workflow built so the failure has nowhere to land: enforced order, required fields, no direct write path, a human signature before anything counts. Model changes cannot touch it.

#	Failure mode	What it looks like on your record	Level	The control that catches it
1	**Fabricated specifics**	It invents a lot number, date, or rate nobody provided: fluent, correctly formatted, wrong.	**2 + 3**	**Traced back:** every stated fact must map to a source; unsourced claims flagged loudly in the record. Reviewers miss fluent errors, so this must be code, with no direct write path behind it.
2	**Vague claims accepted as evidence**	"Probably," "usually," "a lot" get treated as data and the analysis moves on.	**2**	**Thin evidence gate:** a claim without a number or source can't advance the analysis: quantify it or park it.
3	**Operator blame as root cause**	The investigation stops at "operator error, retrain," with no system mechanism.	**1 + review**	**Reframe rule** plus a qualified reviewer. This is the one mode a human reliably catches: it sits in plain sight in the conclusion.
4	**Premature convergence**	It anchors on the first plausible cause and builds a tidy story around it in seconds.	**3**	**Minimum hypotheses rule:** the process requires at least two candidates, with the alternatives documented as ruled out, before any cause advances.
5	**Confirmation without challenge**	It defends the working theory and never asks what would disprove it.	**3**	**Counter evidence requirement:** a mandatory "what would prove this wrong?" step. Absence blocks progress.
6	**The circular 5-Why**	The answer to Why #1 is the theory somebody walked in with; the chain investigates the guess.	**3**	**Fork rule:** the first Why yields candidates by design, IS/IS-NOT cuts the list, and only evidence advances the chain.

7	**Method skipping**	It jumps to causes before the problem is even defined: no IS/IS-NOT, no containment check.	**3**	**Phase gates:** the workflow refuses to advance until the prior step's output exists. Cheapest control on this page.
8	**Missing escape point**	A root cause is named, but nobody answers where the defect should have been caught.	**3**	**Escape point requirement:** a required section, checked for presence.
9	**Citation gaps**	Conclusions appear that don't trace to any evidence in the record.	**2**	**Evidence matrix:** each conclusion links to its source in code; an unlinked conclusion reads as opinion, not finding.
10	**Sycophancy**	"Excellent observation. You're absolutely right." It grades you, and you always get an A.	**2**	**Zero praise standard:** a praise detector, and grade the claim, never the person. The detector most likely to drift: recalibrate often, and on every model change.

How to read the levels: they are minimums for work that lands in a record, reaches a customer, or drives a decision. A draft a human fully rewrites can run lighter. The level is set by your workflow's exposure, not by the mode alone. That rating exercise is the first day of a qualification.

The short version: a demo shows you the happy path. Qualification finds these. You don't find failure mode #10 in a sales meeting. You find it deep in challenge testing, the way I did.

Fae Intelligence · Richard Snyder · operations, quality & manufacturing engineering, 30 years on real floors
See the discipline running: **the Root Cause Analyzer** at rca.faeintelligence.com
faeintelligence.com · rsnyder@faeintelligence.com