

Where it breaks.

The 24 documented AI failure modes, and the check that catches each one.

Italic is the tell, in the book's own words. Roman is the control, quoted from the book (... marks a cut). The number is the mode's full page:
ignatiustheyoungerai.com/failure-modes/N

01 Hallucination <i>The model does not know the difference between remembering and inventing. Neither does the output.</i> Source verification on every external factual claim. 02 Knowledge Cutoff <i>The model is fluent about a world that stopped existing on a date it may not disclose.</i> A date check on every time-sensitive claim. 03 No True Understanding <i>Fluency in your vocabulary is not competence in your job.</i> Read the source document for anything consequential. 04 Overconfidence <i>The tone is a rendering choice. It is not a confidence interval.</i> For any determination that crosses a decision boundary, require the answer to name its own conditions: what would have to be true for this to be wrong? 05 Training-Data Bias <i>The model learned how decisions have been made. You are asking it how they should be made.</i> Adverse-impact testing on the output distribution, not review of the reasoning text. 06 Context-Window Limits <i>It read everything you gave it. It did not attend to everything it read.</i> Chunked processing with explicit coverage accounting. 07 No Common-Sense Grounding <i>It will produce a schedule that requires a truck to be in two states at once, and the formatting will be immaculate.</i> An explicit feasibility pass against a written constraint list. 08 Prompt Injection <i>Your AI reads untrusted text. Instructions inside that text are still instructions.</i> An assistant reading untrusted input should not hold write, send, or lookup permissions on sensitive records. 09 No Persistent Memory <i>Every session starts from zero. Your context does not carry, and nothing announces that it didn't.</i> Every correction becomes a written standing instruction ... not a message in a thread. 10 Sycophancy <i>It answers the question you asked. You asked it to agree with you.</i> Consequential queries must be phrased as questions, not as requests for support. 11 Reasoning Fragility <i>The reasoning is a story about an answer, not the process that produced it.</i> Independent recomputation of the highest-exposure step. 12 No Real-World Verification <i>It can tell you what the invoice should say. It cannot tell you what the invoice says.</i> Name the system of record for each fact class, in writing, and verify against that system, not against whichever system answered first.	13 Verbosity Bias <i>Length reads as rigor. It is produced by neither.</i> Before circulating, list every discrete assertion in the document and mark each: sourced / inferred / assumed. 14 Uncertainty Miscalibration <i>A stated 90% confidence is a linguistic performance, not a probability.</i> Validate the score before trusting it. 15 Training-Data Quality <i>It learned from the internet, including the parts of the internet that were wrong.</i> Route authoritative questions to authoritative sources. 16 Lack Of Accountability <i>The system cannot be held responsible. Someone will be. The only open question is who.</i> Every consequential AI-assisted output carries a record: what was verified, by whom, against what source, on what date, and what was explicitly not verified. 17 Temporal Reasoning <i>It knows the dates. It does not know what order they happened in, or which one is still in force.</i> Build an explicit version timeline before querying. 18 Mathematical Fragility <i>It is not calculating. It is predicting what a calculation looks like.</i> Verify the setup, not the arithmetic. 19 Cannot Truly Cite <i>A citation is an invitation to verify. It is not a certificate of truth.</i> Locate the exact supporting passage. Not the document. The sentence. 20 No Sensory Grounding <i>It has read every description of the warehouse. It has never been in one.</i> Operator review before execution, structurally rather than as a courtesy. 21 Cross-Session Inconsistency <i>Same question. Same document. Different answer. No notification.</i> Scheduled revalidation against a fixed regression set ... re-run on a defined cadence, with results logged and compared over time. 22 Difficulty With Negation <i>'Must not' and 'must' differ by one word. That word is frequently the whole control.</i> Before accepting any summary of a policy, contract, or requirement, list the prohibitions separately ... Then verify each one against the source text individually. 23 Cultural And Linguistic Blind Spots <i>It is fluent in your language. It is not fluent in your market.</i> Native-speaker review with domain context, not translation checking. 24 No True Creativity <i>It recombines what exists. Your strategy problem is that what exists is what your competitors already did.</i> Use the model for the base rate, not the answer.
--	--