



OUTPACED
Individual (Appendix F)

OUTPACED Individual Self-Assessment (Appendix F)

A 10-minute individual self-assessment of one AI deployment. One question per failure mode;
pick the option that most accurately describes what you see.

by **Waydell D. Carvalho**

Published by Cinderpoint

SAFEMACHINE RULES
2026.05

COMPONENT
OUTPACED v1

EFFECTIVE
2026-05-20

VARIANT
Template

Advisory governance assessment - not a regulatory opinion or compliance certification. SAFEMACHINE is not affiliated with NIST, ISO, OECD, or any regulatory body. Citations indicate the source of governance principles, not official endorsement.

Instruction: For each of the 13 failure modes, choose the option (0, 1, 2, or 3) that most accurately describes the AI deployment you are thinking about. The deployment can be one your organization runs, one you use as a customer, or one you interact with as an employee or citizen. Answer every question; if you cannot answer one, or you do not know, score it 3, because not knowing is itself a signal. Then read your answers two ways: the profile (modes scored 2 or 3) and the count of modes scored 3 (which sets the tier).

Use: When you want to quickly recognize whether the 13 failure modes are operating in a specific deployment you encounter. The deeper team-level diagnostic is Appendix G (run inside the SAFEMACHINE app under the Organization project type).

| | |
|---------------------------------------|--|
| SOURCE | Reproduced verbatim from Outpaced (Carvalho 2026, manuscript v18) Appendix F. Each question is the single highest-signal indicator for its failure mode. |
| SCORING | Score each question 0, 1, 2, or 3. 0 means the failure mode is not active; 3 means it is fully active. If you cannot answer or do not know, score it 3. The 13 modes are not summed into one number, because a sum hides a single mode running hot. Read the answers two ways instead. |
| PROFILE | List every mode you scored 2 or 3. Those are the modes most likely operating in the deployment, and the chapters worth rereading are the ones that name them. |
| TIER (COUNT OF MODES SCORED 3) | Count how many modes you scored 3. 0 is Resilient. 1 to 2 is Transitional. 3 or more is Outpaced. A higher tier does not certify that anything is wrong, and a Resilient result does not certify a deployment as safe. |
| WHAT THIS ASSESSMENT IS NOT | A regulatory compliance instrument; a substitute for the team-level work SAFEMACHINE delivers; a predictor of specific incidents; a measure of the AI system's technical safety. It is a recognition tool for an individual reader. |
| GOING FURTHER | The same 13 modes are assessed at greater depth in Appendix G (four questions per mode, rolled into a 0-100 score), run in SAFEMACHINE alongside BASE, SAFEARC, STABLE, and CARG. The two methods differ and land on the same three tier names, so read either as a prompt, not a precise measure. |

The 13 Questions

Pick the option that most accurately describes the AI deployment you are thinking about. Each option is scored 0 (best) to 3 (worst).

1. Illusion of Validity (Ch 1) - When did your organization last verify the AI system against its original safety specification, not against its current performance metrics?
 - 0 - Within the last quarter.
 - 1 - Within the last year.
 - 2 - More than a year ago.
 - 3 - Never since initial deployment, or unknown.

-
2. Illusion of Control (Ch 2) - If a customer or affected person asks why the AI made a specific decision about them, can the customer-facing operation explain that specific decision in real time?
- 0 - Yes, for every decision.
 - 1 - Yes for most, escalation required for some.
 - 2 - Only after an internal investigation.
 - 3 - No, the per-decision explanation is not available.
-
3. Ceremonial Compliance (Ch 3) - If a governance procedure (committee, audit, ombudsman, regulator) found a problem with the AI today, could it stop the AI from continuing to make the same kind of decision?
- 0 - Yes, immediately.
 - 1 - Yes, within a day.
 - 2 - Only after a multi-step escalation.
 - 3 - No, the procedure produces records but does not alter the AI's behavior.
-
4. Pacing Problem (Ch 4) - How many months typically pass between an AI system change and the regulatory or oversight body's ability to authoritatively constrain that change?
- 0 - Same month.
 - 1 - One to three months.
 - 2 - Three months to a year.
 - 3 - More than a year.
-
5. Alert Fatigue (Ch 5) - How many alerts does a typical monitoring analyst process per shift, and what fraction are real signals worth acting on?
- 0 - Fewer than fifty alerts; most are real.
 - 1 - Fifty to two hundred alerts; real fraction is high.
 - 2 - Two hundred to one thousand alerts; real fraction is low.
 - 3 - More than one thousand alerts; real signals are lost in the noise.
-
6. Normalization of Deviance (Ch 6) - When the AI operates outside its original specification but does not produce a catastrophic outcome, what does the organization do?
- 0 - Stops the system and re-verifies against the original specification.
 - 1 - Re-verifies against the specification while the system continues.
 - 2 - Documents the deviation as acceptable and continues.
 - 3 - Treats the new operating point as the new standard.
-

7. Post-hoc Rationalization (Ch 7) - When the organization is asked to explain a specific AI decision, what does it provide?

- 0 - The actual computation that produced the decision.
- 1 - The policy criteria that should have produced the decision.
- 2 - The system's general approach and validation studies.
- 3 - A procedural reference (for example, "the algorithm assessed").

8. Surrogation (Ch 8) - Is the metric the AI optimizes against the same as the outcome the organization actually wants?

- 0 - Yes, the metric is the outcome.
- 1 - The metric is a proxy and is regularly checked against the outcome.
- 2 - The metric is a proxy and the gap to the outcome is widening.
- 3 - The metric and the outcome are now operationally different.

9. Problem of Many Hands (Ch 9) - How many separate organizations have to coordinate to address an AI-driven harm in this deployment?

- 0 - One: a single party owns the integrated question.
- 1 - Two or three, with clear handoffs.
- 2 - Four to six, with informal coordination.
- 3 - Seven or more, and no party owns the integrated picture.

10. Coordination Neglect (Ch 10) - If a multi-agency response to an AI-driven harm is required, how does the response timeline compare to the speed at which the harm propagates?

- 0 - The response is faster than the propagation.
- 1 - Response and propagation operate at comparable speed.
- 2 - The response is months behind the propagation.
- 3 - The response is years behind.

11. Threat Rigidity (Ch 11) - When evidence accumulates that the AI strategic premise is failing, how does the organization respond?

- 0 - Revises the premise.
 - 1 - Pauses and assesses.
 - 2 - Continues with minor adjustments.
 - 3 - Intensifies commitment to the premise.
-

12. Moral Licensing (Ch 12) - Does the organization have publicly articulated AI ethics principles, and how often are specific product decisions audited against them?

- 0 - Principles exist and every relevant decision is audited.
- 1 - Principles exist and most decisions are audited.
- 2 - Principles exist and occasional audits occur.
- 3 - Principles exist; specific decisions are not externally auditable.

13. Cascading Failures (Ch 13) - What fraction of the AI system's code, configuration, and deployment state is documented and modeled by the operations team?

- 0 - Nearly all of it.
- 1 - Most of it.
- 2 - Some of it.
- 3 - A small fraction; the system contains accumulated state nobody fully models.

Red Flags

- Three or more modes scored 3: Outpaced tier - several modes are running together and the interaction this book calls Outpaced may be forming.
- Any single mode at 3: that mode is fully active and warrants targeted attention.
- Mode 1 (Illusion of Validity) or Mode 6 (Normalization of Deviance) at 3: the original safety specification is no longer the operational reference. Hard escalation flag.
- A wide profile (many modes at 2 or 3) even with few 3s: the deeper Appendix G diagnostic is warranted.

Reading the Score

14.1- Strong signals - count the modes you scored 3:

.

14.2- Tier (0 modes at 3 = Resilient; 1-2 = Transitional; 3 or more = Outpaced):

.

14.3- Profile - the modes you scored 2 or 3 (the chapters worth rereading):

.

14.4- Next step:

.

Citations

[1] **Outpaced (Carvalho 2026), Appendix F** - <https://cinderpoint.com/safemachine/books/outpaced>

[2] **SAFEMACHINE app (free, browser, offline-first)** - <https://cinderpoint.com/safemachine>



OUTPACED
13 Modes

OUTPACED Self-Assessment (13 Failure Modes)

Assess an AI deployment against the 13 failure modes that operate inside Automation Complacency.

by **Waydell D. Carvalho**

Published by Cinderpoint

SAFEMACHINE RULES
2026.05

COMPONENT
OUTPACED v1

EFFECTIVE
2026-05-20

VARIANT
Template

Advisory governance assessment - not a regulatory opinion or compliance certification. SAFEMACHINE is not affiliated with NIST, ISO, OECD, or any regulatory body. Citations indicate the source of governance principles, not official endorsement.

Instruction: For each of the 13 modes, answer 4 observable indicators. RED = none / not present. YELLOW = partial / inconsistent. GREEN = sufficient / fully in place. Modes accumulate; the total score and per-mode profile identify which modes are most active.

Use: Run per deployment when the deployment changes meaningfully, when an incident surfaces, or every six months alongside BASE.

| | |
|---|---|
| SOURCE | Derived from Outpaced (Carvalho 2026), Chapters 1-13. The book introduces each failure mode with a named real-world case and prior-art lineage. |
| SCORING | Per question RED=20, YELLOW=10, GREEN=0. Per mode = sum of 4 questions (0-80 raw / 0-100 normalized). Total = sum of 13 modes (0-1040 raw / 0-100 normalized). |
| MISSING ANSWERS (CONSERVATIVE DEFAULT) | If you cannot answer an indicator, or do not know, score it at the most-exposed level (RED). Silence is not a pass. All 52 indicators are required; the app treats any missing answer as RED. |
| TIER THRESHOLDS (NORMALIZED 0-100) | 0-30 Resilient. 31-65 Transitional. 66-100 Outpaced. |
| RELATION TO BASE | BASE assesses organizational readiness. OUTPACED assesses deployment-specific failure-mode exposure. Use BASE first; use OUTPACED per AI deployment. |

Mode 1 - Illusion of Validity (Outpaced Ch 1)

The state in which a system is judged safe because its tracked performance indicators have remained within their tracked ranges, regardless of whether those indicators measure the conditions that would surface failure.

1.1. Has the AI system been verified against its original safety specification in the last six months?

RED = never or unknown. YELLOW = more than six months ago. GREEN = within last six months.

RED
 YELLOW
 GREEN

1.2. Are the metrics tracking the AI independent of the metrics the AI itself produces?

RED = all metrics derived from system outputs. YELLOW = some independent. GREEN = independent metrics cover critical behavior.

RED
 YELLOW
 GREEN

1.3. When stable performance metrics are observed, is the conclusion drawn that the system is safe overall?

RED = stable metrics treated as overall safety evidence. YELLOW = mixed. GREEN = inference limited to what metrics measure.

RED
 YELLOW
 GREEN

1.4. Does monitoring detect conditions outside the system’s original operating envelope?

RED = only within-envelope deviations detected. YELLOW = partial out-of-envelope detection. GREEN = systematic out-of-envelope detection.



Mode 2 - Illusion of Control (Outpaced Ch 2)

The state in which an organization treats its formal ownership of an AI system as if that ownership were the same as the capacity to explain, revise, or override specific decisions the system produces.

2.1. Can the customer-facing operation explain a specific AI decision in real time?

RED = no explanation available. YELLOW = only after investigation. GREEN = yes, every decision.



2.2. Can a human operator override a specific AI decision without escalation?

RED = no override capacity at operator level. YELLOW = requires escalation. GREEN = operator override in real time.



2.3. Does the organization’s documentation of the AI correspond to operational control over its decisions?

RED = documentation and control decoupled. YELLOW = partial coupling. GREEN = documentation maps to control levers.



2.4. Can the human review function engage with the actual volume of decisions the AI is making?

RED = volume exceeds review capacity by orders of magnitude. YELLOW = sample-based only. GREEN = meaningful fraction reviewed.



Mode 3 - Ceremonial Compliance (Outpaced Ch 3)

The belief that governance activities constrain system behavior because they exist, run on schedule, and produce records, regardless of whether the activities have any mechanism to alter what the system is doing.

3.1. If a governance procedure finds a problem with the AI today, can it stop the AI from continuing to make the same kind of decision?

RED = procedure produces records but does not alter behavior. YELLOW = only after multi-step escalation. GREEN = yes, immediately.



3.2. Does the governance review frequency match the system’s rate of change?

RED = annual reviews against daily change. YELLOW = quarterly against weekly. GREEN = aligned cadence.

RED
 YELLOW
 GREEN

3.3. When governance bodies make recommendations, what fraction are implemented in operational practice?

RED = few or none. YELLOW = some. GREEN = most implemented and verified.

RED
 YELLOW
 GREEN

3.4. Could the governance body, if it decided to, stop the system from making the same kind of decision tomorrow?

RED = no. YELLOW = only through external escalation. GREEN = yes, directly.

RED
 YELLOW
 GREEN

Mode 4 - Pacing Problem (Outpaced Ch 4)

The structural delay between an AI system’s deployment at scale and the moment when the oversight structures around it can authoritatively constrain what it is doing. Oversight here is broader than law: external regulators, internal safety reviews, ethics boards, audit functions, and professional codes are all in scope.

4.1. How many months pass between an AI system change and the relevant oversight body’s ability to authoritatively constrain that change?

RED = more than a year. YELLOW = one month to a year. GREEN = same month.

RED
 YELLOW
 GREEN

4.2. If multiple oversight bodies have jurisdiction, do they coordinate at the speed of the system?

RED = no coordination. YELLOW = coordinates at slowest pace. GREEN = near system speed.

RED
 YELLOW
 GREEN

4.3. Was the AI system reviewed by relevant oversight bodies before deployment, or only after harms surfaced?

RED = no pre-deployment review. YELLOW = light-touch. GREEN = substantive pre-deployment review with binding outcomes.

RED
 YELLOW
 GREEN

4.4. When oversight findings land, are they about the current version of the system or an earlier version?

RED = always lag by versions. YELLOW = mixed. GREEN = address current version.

RED
 YELLOW
 GREEN

Mode 5 - Alert Fatigue (Outpaced Ch 5)

The state in which a monitoring environment generates more information than the human attention inside it can process, so that the information needed to interrupt a failure becomes operationally indistinguishable from the information that does not need action.

5.1. How many alerts does each monitoring analyst process per shift?

RED = more than 1,000; signals lost. YELLOW = 200 to 1,000; low real fraction. GREEN = under 200; meaningful triage.

RED YELLOW GREEN

5.2. Has the monitoring team had a real intrusion or alert hit recently, and was it identified inside the routine alert flow?

RED = past signals missed in stream. YELLOW = caught only after external disclosure. GREEN = caught in routine flow.

RED YELLOW GREEN

5.3. When the team escalates an alert, what fraction reaches an analyst's intervention in time?

RED = most escalations sit past actionable window. YELLOW = some timely. GREEN = most timely.

RED YELLOW GREEN

5.4. Does the monitoring system have automatic protective actions, or does every alert require human review before anything happens?

RED = all require human review. YELLOW = some auto-actions. GREEN = automatic actions handle clear cases.

RED YELLOW GREEN

Mode 6 - Normalization of Deviance (Outpaced Ch 6)

The process by which an organization, having operated successfully outside the original specification of a system, comes to treat the new operating point as the operating standard.

6.1. When the AI operates outside its original specification but produces no catastrophic outcome, what happens?

RED = new operating point becomes new standard. YELLOW = documented as acceptable. GREEN = stopped and re-verified.

RED YELLOW GREEN

6.2. If waivers are issued for out-of-spec operation, are they time-limited and reviewed?

RED = persist indefinitely. YELLOW = reviewed but rarely revoked. GREEN = time-limited and often revoked.

RED YELLOW GREEN

6.3. Can the operations team point to the original safety specification today?

RED = not in operational use. YELLOW = exists but not the reference. GREEN = active operating reference.

RED YELLOW GREEN

6.4. When release decisions are made, what is the comparison baseline?

RED = previous release. YELLOW = recent prior release. GREEN = original safety case.

**Mode 7 - Post-hoc Rationalization (Outpaced Ch 7)**

The structural divergence between the explanation an organization provides for an AI decision and the operational mechanism that actually produced the decision.

7.1. When explaining a specific AI decision, what does the organization provide?

RED = a procedural reference. YELLOW = policy criteria. GREEN = actual computation.

**7.2.** If the AI is a vendor product, can the operating organization examine the decision logic?

RED = proprietary, no access. YELLOW = limited contract-based reviews. GREEN = full examination access.

**7.3.** Has any independent audit reached the per-decision logic, or only system-level documentation?

RED = system-level only. YELLOW = partial per-decision. GREEN = independent per-decision audit completed.

**7.4.** Does the organization's public explanation of how the AI works align with the operational mechanism?

RED = public describes intent; mechanism differs. YELLOW = partial alignment. GREEN = matches.

**Mode 8 - Surrogation (Outpaced Ch 8)**

The structural condition in which an organization builds its operating decisions around a measurable proxy and, over time, the proxy and the underlying outcome decouple.

8.1. Is the metric the AI optimizes against the same as the outcome the organization actually wants?

RED = metric and outcome operationally different. YELLOW = proxy, gap widening. GREEN = metric is the outcome.

**8.2.** How often is the chosen optimization metric reviewed against the underlying outcome?

RED = never since metric chosen. YELLOW = annual. GREEN = continuous.



8.3. Is staff compensation tied to the metric the AI optimizes against?

RED = tightly tied across levels. YELLOW = loosely tied. GREEN = tied to outcomes, not metrics.

**8.4.** If the metric were found misaligned with the outcome, could the organization replace it without re-engineering operating infrastructure?

RED = major re-engineering required. YELLOW = moderate effort. GREEN = metrics designed to be replaceable.

**Mode 9 - Problem of Many Hands (Outpaced Ch 9)**

The structural condition in which a system's operation, oversight, and consequences are distributed across multiple parties such that no single party owns the integrated question.

9.1. How many separate organizations coordinate to address AI-driven harm in this deployment?

RED = seven or more, no integrated owner. YELLOW = four to six with informal coordination. GREEN = one or clear ownership.

**9.2.** When a harm occurs, is the integrated record assembled by any party, or only by external investigators?

RED = only external investigators. YELLOW = partially internal. GREEN = fully internal on defined timeline.

**9.3.** Are hand-offs between the parties documented with shared ownership of the integrated outcome?

RED = unclear or undocumented. YELLOW = documented but not jointly owned. GREEN = documented and jointly owned.

**9.4.** If the same harm occurs at multiple sites, is the pattern detected within the operating chain?

RED = cross-site patterns only via external investigators. YELLOW = surface slowly. GREEN = systematic internal detection.

**Mode 10 - Coordination Neglect (Outpaced Ch 10)**

The structural condition in which a harm is met by many separate overseers, each acting within its own authority, and no party performs the integrating work of combining their separate responses into a single binding constraint.

10.1- When multi-agency response is required, what is the typical timeline from incident to coordinated finding?

RED = years. YELLOW = months. GREEN = weeks or faster.

**10.2-** Does the multi-agency response speed match the system's rate of operational change?

RED = system daily, response yearly. YELLOW = order-of-magnitude mismatch. GREEN = matched.

**10.3-** Is there a mechanism to act on the parts of a finding that are settled without waiting for the slowest participant?

RED = waits for slowest party. YELLOW = some staged action. GREEN = findings act in stages.

**10.4-** Does a standing inter-agency coordination structure exist, or is it improvised per incident?

RED = improvised each time. YELLOW = exists for some harm types. GREEN = standing structure with binding authority.

**Mode 11 - Threat Rigidity (Outpaced Ch 11)**

The structural condition in which an organization, faced with evidence that its strategic premise is failing, responds by intensifying its commitment to the premise, narrowing the information it takes in, and centralizing control, rather than revising it.

11.1- When evidence accumulates that the AI strategic premise is failing, what does the organization do?

RED = intensifies commitment. YELLOW = minor adjustments. GREEN = revises premise or pauses to assess.

**11.2-** How does the organization handle sunk-cost arguments in strategic decisions?

RED = sunk costs anchor continued commitment. YELLOW = weighed but often win. GREEN = excluded from forward decisions.



11.3- When internal dissent surfaces about the AI strategy, how is it processed?

RED = dissenters marginalized. YELLOW = heard but not acted on. GREEN = triggers structured review.



11.4- Is the organization's public narrative about the AI tightly coupled to the strategy?

RED = tightly coupled; revision means credibility loss. YELLOW = some coupling. GREEN = loosely coupled.



Mode 12 - Moral Licensing (Outpaced Ch 12)

The institutional pattern in which an organization's publicly articulated ethical identity functions as a substitute for ethical scrutiny of specific decisions.

12.1- Does the organization audit specific product decisions against its publicly articulated AI ethics principles?

RED = not externally auditable. YELLOW = occasional audits. GREEN = every relevant decision audited.



12.2- Are the internal review decisions externally auditable?

RED = internal-only. YELLOW = audit access on request. GREEN = externally auditable as standard.



12.3- Does the organization use its ethics commitment to deflect scrutiny of specific decisions?

RED = identity functions as the answer. YELLOW = sometimes. GREEN = specific scrutiny separate from identity.



12.4- How often are the AI ethics principles updated against operational reality?

RED = fixed at publication, never revised. YELLOW = updated under external pressure. GREEN = updated against operational learning.



Mode 13 - Cascading Failures (Outpaced Ch 13)

The structural condition in which a complex automated system contains accumulated state that interacts with new changes in ways the change-validation procedures do not detect.

13.1- What fraction of the AI system's code, configuration, and deployment state is documented and modeled by the operations team?

RED = small fraction. YELLOW = some documented, gaps. GREEN = nearly all.



13.2- Does the codebase contain dormant code, deprecated functions, or feature flags from prior releases?

RED = extensive, no removal discipline. YELLOW = some, occasional cleanup. GREEN = removed or actively tracked.



13.3- When code deploys, is the binary on each server verified against the expected hash?

RED = engineer confirmation only. YELLOW = spot checks. GREEN = every server verified on every deploy.



13.4- When a cascade begins, how quickly can the operating team interrupt it?

RED = cannot interrupt before significant damage. YELLOW = manual intervention in minutes or hours. GREEN = automatic circuit breakers in seconds.



Red Flags

- Any single mode at full RED (4 of 4 questions) indicates structural exposure on that mode and warrants targeted intervention before deployment proceeds.
- Three or more modes at full RED indicate Outpaced-tier exposure and require comprehensive intervention.
- A mode in which the original safety specification cannot be located (Illusion of Validity or Normalization of Deviance at RED) is a hard escalation flag.

Final OUTPACED Summary

F1. Computed OUTPACED tier (Resilient / Transitional / Outpaced).

F2. Top three modes by raw score.

F3. Top three actions to reduce exposure on the highest-scoring modes.

Citations

[1] **Outpaced (Carvalho 2026)** - <https://cinderpoint.com/outpaced>

[2] **Reason 1990 - Human Error** - <https://www.cambridge.org/9780521314190>

[3] **Power 1997 - The Audit Society** - <https://global.oup.com/academic/product/the-audit-society-9780198296034>

[4] **Vaughan 1996 - The Challenger Launch Decision** - <https://press.uchicago.edu/ucp/books/book/chicago/C/bo3624858.html>

[5] **Perrow 1984 - Normal Accidents** - <https://press.princeton.edu/books/paperback/9780691004129/normal-accidents>

[6] **Staw, Sandelands, Dutton 1981 - Threat-Rigidity Effects** - <https://www.jstor.org/stable/2392337>

[7] **Merritt, Efron, Monin 2010 - Moral Self-Licensing** - <https://compass.onlinelibrary.wiley.com/doi/10.1111/j.1751-9004.2010.00263.x>

[8] **Goodhart 1975 / Campbell 1979 - Goodhart's Law / Campbell's Law** - https://en.wikipedia.org/wiki/Goodhart%27s_law

[9] **Thompson 1980 - Organizations in Action** - <https://www.routledge.com/9780765809919>

[10] **Langer 1975 - Illusion of Control** - <https://psycnet.apa.org/record/1976-03345-001>

[11] **Meyer-Rowan 1977 - Institutionalized Organizations** - <https://www.jstor.org/stable/2778293>

[12] **Bromley-Powell 2012 - From Smoke and Mirrors to Walking the Talk** - <https://doi.org/10.1080/19416520.2012.684462>