SAFEARC

SCAN

SAFEARC - SCAN Worksheet

Identify every AI system, its data, owners, dependencies, and risk surface before applying safety controls.

by **Waydell D. Carvalho**

Published by Cinderpoint

SAFEMACHINE RULES	COMPONENT	EFFECTIVE	VARIANT
2026.05	**SAFEARC v10.1**	**2026-05-20**	**Template**

Use: Complete before building, buying, updating, or replacing any AI system.

1. System Identity

1. System name

2. Version / release date

3. Vendor or internal team

4. Primary business purpose

2. Data Inventory

5. What data does this system use?

- ☐ Customer data
- ☐ Employee data
- ☐ Financial data
- ☐ Public/open web data
- ☐ Purchased/third-party data
- ☐ Proprietary internal data
- ☐ Sensitive or regulated data

6. Sources of data

7. Does data leave the organization?

If yes, where and under what terms?

☐ Yes ☐ No

3. Model / Automation Inventory

8. Components used

- ☐ LLM
- ☐ Classifier
- ☐ Recommender
- ☐ Automation / RPA
- ☐ Fine-tuned model
- ☐ External API
- ☐ On-prem model
- ☐ Open-source model
- ☐ Other

9. Model origin

- ☐ Off-the-shelf
- ☐ Fine-tuned
- ☐ Trained in-house

10. Model update frequency

- ☐ Real-time / continuous
- ☐ Daily
- ☐ Weekly
- ☐ Monthly
- ☐ Quarterly
- ☐ Rare / manual

4. Ownership & Responsibilities

11. Business owner

12. Technical owner

13. Change approver

14. Monitoring owner

5. Dependencies & Integrations

15. Upstream dependencies

16. Downstream dependencies

17. Cascading failure risk?

If yes, describe.

☐ Yes ☐ No

6. Risk Surface Indicators

18. Interacts with customers?

☐ Yes ☐ No

19. Produces decisions / recommendations?

☐ Yes ☐ No

20. Incorrect output could cause harm?

☐ Yes ☐ No

21. Incorrect output could create legal / compliance issues?

☐ Yes ☐ No

7. Shadow AI Detection

22. Was this system acquired or implemented outside standard IT procurement?

If yes, describe.

☐ Yes ☐ No

23. Is the vendor contract unknown, expired, or missing?

☐ Yes ☐ No

Red Flags

- [ ] None of these apply (confirmed).
- [ ] No documented data sources
- [ ] No monitoring workflow
- [ ] No audit trail
- [ ] Black-box vendor model
- [ ] Ownership unclear or shared
- [ ] No rollback procedure
- [ ] No pre-launch review
- [ ] Unknown third-party dependencies

9. Final SCAN Summary

25. Top 3 risks

26. Immediate next steps

27. Requirements before entering ALIGN

SAFEARC

ALIGN

SAFEARC - ALIGN Worksheet

Confirm the system's behavior, permissions, and boundaries align with policies, laws, risk levels, and values.

by **Waydell D. Carvalho**

Published by Cinderpoint

SAFEMACHINE RULES	COMPONENT	EFFECTIVE	VARIANT
2026.05	**SAFEARC v10.1**	**2026-05-20**	**Template**

Instruction: If you cannot answer a question, mark it as a gap.

Use: Complete after SCAN and before FILTER.

1. Policy Alignment

1. Which business policy governs this system's use?

2. Written rules exist for

   - ☐ Data handling
   - ☐ Output limits
   - ☐ Escalation
   - ☐ Human oversight
   - ☐ None exist (red flag)

3. Legal / regulatory alignment

   If no, describe gaps.

   ☐ Yes ☐ No

4. Aligns with internal risk tolerance

   ☐ Yes ☐ No

2. Intended Use vs. Actual Use

5. Intended use

6. Used outside scope?

   If yes, describe.

   ☐ Yes ☐ No

7. Permitted decisions or recommendations

8. Prohibited decisions or actions

3. Input Boundaries

9. Allowed inputs

10. Blocked or rejected inputs

11. Accepts free-text input?

Safeguards?

☐ Yes ☐ No

4. Output Boundaries

12. Allowed outputs

13. Prohibited outputs

14. Could outputs be mistaken for authoritative or legally binding?

Safeguards?

☐ Yes ☐ No

15. Could outputs be mistaken for authoritative / official communication (medical, legal, financial, HR)?

Safeguards?

☐ Yes ☐ No

5. Escalation & Human Oversight

16. When must a human review output?

17. Escalation triggers

- ☐ Uncertain output
- ☐ High-risk decision
- ☐ Sensitive interaction
- ☐ Legal / compliance impact
- ☐ Unfamiliar situation
- ☐ User reports inconsistency

18. Assigned human reviewer

19. Review timeframe (SLA)

6. Behavior Restrictions

20. Behaviors the system must avoid

21. "Never actions"

22. Unsafe domains where the system must not operate

Red Flags

- ☐ None of these apply (confirmed).
- ☐ No written policies
- ☐ No oversight plan
- ☐ No escalation rules
- ☐ Used outside intended scope
- ☐ Conflicting stakeholders
- ☐ Misaligned risk expectations
- ☐ Outputs could mislead users
- ☐ No guardrails for free-text input

8. Final Alignment Summary

23. Alignment gaps identified

24. Required actions before FILTER

25. Approval required before system advances

SAFEARC
FILTER

SAFEARC - FILTER Worksheet

Identify, block, restrict, or adjust unsafe AI behaviors BEFORE the system is allowed to operate.

by **Waydell D. Carvalho**

Published by Cinderpoint

SAFEMACHINE RULES	COMPONENT	EFFECTIVE	VARIANT
2026.05	**SAFEARC v10.1**	**2026-05-20**	**Template**

Instruction: If you cannot answer a question, mark it as a gap.

Use: Complete after ALIGN and before EVALUATE.

> **PRINCIPLE** FILTER ensures only safe, allowed behavior reaches users.

1. REMOVE - Eliminate Unsafe Behaviors

Example: remove all medical claims, refund guarantees, policy interpretations.

1. Unsafe outputs identified

2. Behaviors to eliminate

3. Automatic removal filters?

   ☐ Yes ☐ No

4. Domains NEVER allowed

2. REDUCE - Soften or Lower Risk Areas

Example: reduce certainty language, add disclaimers.

5. Outputs needing hedging

6. Tone / disclaimers needed?

   ☐ Yes ☐ No

7. Inputs needing preprocessing?

   ☐ Yes ☐ No

3. RESTRICT - Apply Hard Boundaries

Example: restrict domain to customer-service only.

8. Outputs to restrict / block

9. Inputs to restrict / block

10. Guardrails needed

- ☐ Tone
- ☐ Claims
- ☐ Domain
- ☐ Length
- ☐ Certainty
- ☐ Citations

11. Actions AI must NOT take

- ☐ Escalations
- ☐ Routing
- ☐ Customer labels
- ☐ Prioritization
- ☐ Safety decisions

4. RE-ROUTE - Force Human or System Review

Example: high-risk tickets routed to humans.

12. Outputs triggering human review

13. Inputs triggering human review

14. Cases re-routed to humans?

Rules?

☐ Yes ☐ No

15. Fallback mode required?

☐ Yes ☐ No

5. Technical Filters & Guardrails

Example: regex filters, prompt constraints.

16. Filter mechanisms

- ☐ Prompt
- ☐ Model
- ☐ System
- ☐ Output

17. Test suite?

Coverage?

☐ Yes ☐ No

18. Vendor guardrails verified?

☐ Yes ☐ No

19. Filter logs stored?

☐ Yes ☐ No

Red Flags

- ☐ None of these apply (confirmed).
- ☐ No removal filters
- ☐ No domain blocks
- ☐ No routing restrictions
- ☐ No test suite
- ☐ No fallback mode
- ☐ Filters unclear to staff
- ☐ Filter bypass possible
- ☐ No logging

7. Final FILTER Summary

20. Key risks removed

21. Restrictions enforced

22. Required handoffs before EVALUATE

23. Approval required

SAFEARC

EVALUATE

SAFEARC - EVALUATE Worksheet

Verify the AI system's accuracy, reliability, consistency, and safety before approval for real use.

by **Waydell D. Carvalho**

Published by Cinderpoint

SAFEMACHINE RULES	COMPONENT	EFFECTIVE	VARIANT
2026.05	**SAFEARC v10.1**	**2026-05-20**	**Template**

Instruction: If you cannot answer a question, mark it as a gap.

Use: Complete after FILTER and before ASSIGN.

> **TEST SET** A representative test set includes typical cases, edge cases, safety-critical interactions, and realistic misuse scenarios reflecting real user behavior.

1. Testing Setup

1. System version being tested

2. Scenarios included in the test set

3. Who prepared the test data?

4. Test set includes

- ☐ Normal cases
- ☐ Edge cases
- ☐ Adversarial inputs
- ☐ Safety-critical cases
- ☐ Misuse cases
- ☐ High-variability customer inputs

5. Test set includes prohibited or high-risk inputs?

☐ Yes ☐ No

2. Accuracy & Reliability

6. Accuracy metrics used

7. Observed accuracy

8. Does accuracy meet threshold?

☐ Yes ☐ No

9. Accuracy variation across customer types / languages / categories / risk

10. Brittleness under slight changes?

☐ Yes ☐ No

3. Safety & Compliance Checks

11. Prohibited outputs blocked?

☐ Yes ☐ No

12. Any authoritative / binding outputs?

☐ Yes ☐ No

13. Any safety-critical outputs bypass FILTER?

☐ Yes ☐ No

14. Compliant with internal and regulatory boundaries?

☐ Yes ☐ No

15. Escalation triggers activated?

☐ Yes ☐ No

4. Bias, Drift & Consistency

16. Outputs consistent across repeated inputs?

☐ Yes ☐ No

17. Any bias detected?

☐ Yes ☐ No

18. Drift detected from prior evaluations?

☐ Yes ☐ No

19. Variability within acceptable limits?

☐ Yes ☐ No

20. Vendor LLM updates affect performance?

☐ Yes ☐ No

21. Organization receives vendor update notifications?

☐ Yes ☐ No

5. Performance Under Stress

22. Tested under high load?

☐ Yes ☐ No

23. Fail / stall / degrade under load?

☐ Yes ☐ No

24. Fallback mode activated correctly?

☐ Yes ☐ No

25. Logs captured all critical events?

☐ Yes ☐ No

Red Flags

- ☐ None of these apply (confirmed).
- ☐ Low accuracy
- ☐ High variance
- ☐ Drift detected
- ☐ Unsafe outputs appeared
- ☐ Missing escalation triggers
- ☐ Bias detected
- ☐ Incomplete test coverage
- ☐ Fallback not functioning
- ☐ Missing logs
- ☐ Vendor updates affecting behavior

7. Final Evaluation Summary

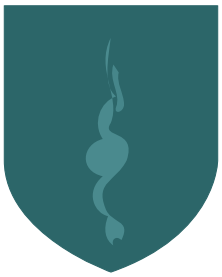
26. Evaluation passed?

☐ Yes ☐ No

27. Key findings

28. Required fixes before ASSIGN

29. Who approves evaluation results?



SAFEARC

ASSIGN

SAFEARC - ASSIGN Worksheet

Establish clear, documented ownership and accountability for the AI system.

by **Waydell D. Carvalho**

Published by Cinderpoint

SAFEMACHINE RULES	COMPONENT	EFFECTIVE	VARIANT
2026.05	**SAFEARC v10.1**	**2026-05-20**	**Template**

Instruction: If you cannot answer a question, mark it as a gap.

Use: Complete after EVALUATE and before RENEW.

> **AUTHORITY** ASSIGN must be completed by leadership, not operators, to ensure true authority alignment.

1. System Owner

1. Primary owner of this AI system

2. Reason this role / person is selected

3. Ownership documented in policy?

☐ Yes ☐ No

4. Authority to

- ☐ Approve changes
- ☐ Halt deployment
- ☐ Trigger escalation
- ☐ Approve training

2. Operators

5. Operators / teams

6. Understand limits and escalation rules?

☐ Yes ☐ No

7. Completed training?

If yes, date.

☐ Yes ☐ No

8. Can operators bypass guardrails?

If yes, why.

☐ Yes ☐ No

3. Oversight

9. Oversight role(s)

10. Responsibilities

- ☐ Output review
- ☐ Escalation handling
- ☐ Policy interpretation
- ☐ Exception handling
- ☐ Drift monitoring
- ☐ Incident intake

11. Continuous coverage?

☐ Yes ☐ No

4. Accountability

12. Who is accountable if harm occurs?

13. Single or shared accountability?

- ☐ Single-owner
- ☐ Shared

14. Authority equal to risk?

☐ Yes ☐ No

15. Accountability includes

- ☐ Escalation authority
- ☐ Stop-deployment authority
- ☐ Audit responsibility

5. Permissions

16. Configuration access

17. Runtime access

18. Who can modify prompts / filters?

19. Permissions documented / audited?

☐ Yes ☐ No

20. Unauthorized change controls in place?

☐ Yes ☐ No

6. Training & Competency

21. Competencies defined?

☐ Yes ☐ No

22. Competency checks completed?

☐ Yes ☐ No

23. Retraining schedule

- ☐ Quarterly
- ☐ Semi-annual
- ☐ Annual

24. Training includes

- ☐ Safety risks
- ☐ Boundary rules
- ☐ Escalation triggers
- ☐ Incident reporting
- ☐ Domain limits
- ☐ Review duties

Red Flags

- ☐ None of these apply (confirmed).
- ☐ No documented owner
- ☐ Owner lacks authority
- ☐ Operators unclear
- ☐ Oversight gaps
- ☐ Accountability unclear
- ☐ Permission creep
- ☐ No training
- ☐ No audit trail
- ☐ Bypass possible

8. Final Assignment Summary

25. Final owner

26. Oversight lead(s)

27. Operator groups

28. Permissions summary

29. Approval to move to RENEW

SAFEARC

RENEW

SAFEARC - RENEW Worksheet

Keep the AI system safe and aligned over time through scheduled review, retraining, and governance updates.

by **Waydell D. Carvalho**

Published by Cinderpoint

SAFEMACHINE RULES	COMPONENT	EFFECTIVE	VARIANT
2026.05	**SAFEARC v10.1**	**2026-05-20**	**Template**

Instruction: If you cannot answer a question, mark it as a gap.

Use: Completed after ASSIGN and then on a regular cycle.

> **CADENCE** RENEW must be completed on a fixed cycle (quarterly or semi-annual) to maintain ongoing safety and compliance.

1. Review - Regular Governance Check

1. Date of this RENEW cycle

2. Time since last review

   ☐ < 3 months
   ☐ 3-6 months
   ☐ > 6 months

3. Business use-case changed?

   ☐ Yes ☐ No

4. New regulations or policies?

   ☐ Yes ☐ No

5. Any incidents or near-misses?

   ☐ Yes ☐ No

2. Refresh - Update Rules, Guardrails, Documents

6. Policies, filters, guardrails still correct?

   ☐ Yes ☐ Partially ☐ No

7. SCAN and ALIGN still accurate?

   ☐ Yes ☐ No

8. Escalation paths updated?

   ☐ Yes ☐ No

9. Disclaimers / messages reviewed?

   ☐ Yes ☐ No

10. Logs and audits consistent?

☐ Yes ☐ No

3. Retrain - Data, Models, Evaluation

11. Need retraining or prompt updates?

☐ Yes ☐ No

12. EVALUATE re-run?

☐ Yes ☐ No

13. Drift or bias changes?

☐ Yes ☐ No

14. New product / policy updates required?

☐ Yes ☐ No

15. Datasets still representative?

☐ Yes ☐ Mostly ☐ No

4. People - Training and Ownership Renewal

16. Has owner changed?

☐ Yes ☐ No

17. Oversight / operator changes?

☐ Yes ☐ No

18. All roles trained this cycle?

☐ Yes ☐ No

19. Additional training required?

☐ Yes ☐ No

5. Retire, Replace, or Reduce Use

20. Should system continue?

- ☐ Keep
- ☐ Reduce scope
- ☐ Replace / upgrade
- ☐ Retire

21. If reducing scope

22. If replacing

23. If retiring

Red Flags

- ☐ None of these apply (confirmed).
- ☐ No review cycle
- ☐ Business changed without reevaluation
- ☐ New laws ignored
- ☐ Repeated incidents
- ☐ Outdated guardrails
- ☐ No drift monitoring
- ☐ Owner / oversight not updated
- ☐ Operators untrained
- ☐ System running without governance

7. Final RENEW Summary

24. Key changes

25. Actions before next cycle

26. Next RENEW date

27. Approver

SAFEARC

CONTAIN

SAFEARC - CONTAIN Worksheet

Prevent unsafe AI outputs from reaching users or systems by establishing layered boundary controls.

by **Waydell D. Carvalho**

Published by Cinderpoint

SAFEMACHINE RULES	COMPONENT	EFFECTIVE	VARIANT
2026.05	**SAFEARC v10.1**	**2026-05-20**	**Template**

Instruction: If you are unsure, mark UNKNOWN instead of guessing.

Use: Completed after RENEW and updated anytime risks change.

> **CADENCE** Containment controls must be tested on a fixed cycle (quarterly or semi-annual) and updated whenever risks, systems, or workflows change.

1. Barriers (Automatic System Guards)

1. Does the system have built-in refusal rules?

   ☐ Yes ☐ No

2. Does the system block prohibited topics or functions?

   ☐ Yes ☐ No ☐ Unknown

3. Are guardrails or safety filters active at all times?

   ☐ Yes ☐ No ☐ Depends on user role

4. Are output-sanitization steps in place?

   ☐ Yes ☐ No

5. Have guardrails been tested recently?

   ☐ Yes ☐ No

2. Boundaries (Policy + Human Oversight Controls)

6. Is the scope of the system clearly defined and documented?

   ☐ Yes ☐ No

7. Are there clearly defined "Do Not Use AI For This" rules?

   ☐ Yes ☐ No

8. Are humans required to approve specific types of actions?

   ☐ Yes ☐ No

9. Are override permissions restricted and logged?

   ☐ Yes ☐ No

10. Does the system have required escalation triggers?

☐ Yes ☐ No

3. Breakers (Emergency Stop / Containment Modes)

11. Is there an emergency "kill switch" for the system?

☐ Yes ☐ No

12. Can operators pause the model or workflow instantly?

☐ Yes ☐ No ☐ Depends on system

13. Does the kill switch disable only the faulty function, or the entire system?

☐ Function-level
☐ Full shutdown
☐ Unknown

14. Is kill-switch authority restricted?

☐ Yes ☐ No

15. Has the kill switch been tested in the last 6 months?

☐ Yes ☐ No

Red Flags

☐ None of these apply (confirmed).
☐ Guardrails untested
☐ Kill switch untested
☐ No output sanitization
☐ No scope boundaries
☐ Unknown escalation triggers
☐ Overrides unlogged
☐ Safety filters disabled for certain users
☐ No prohibited-use list

5. Final Containment Summary

16. Residual risk level

- ☐ Low
- ☐ Medium
- ☐ High
- ☐ Critical

17. Areas requiring new containment rules

18. Kill-switch authority assigned to

19. Approver for final containment plan