S4 AI Safety Framework

Evidence for Deciding What an AI System May Do

Adrian Dunkley
Caribbean AI Association
info@caribbeanaiassociation.com

16 November 2026

Preprint v1.0 · Framework v0.2

Abstract

Organisations need evidence for each consequential permission they give an AI system. S4 is a proposed method for assembling that evidence through four checks: Specify, Stress-test, Safeguard and Scrutinise. This paper defines the unit of assessment, the records that support a decision and the conditions that keep a permission on hold. It applies the method retrospectively to two public cases: conflicting chatbot advice in Moffatt v. Air Canada and the Knight Capital automated trading failure. A worked example shows how to evaluate a draft-only climate service. A prospective protocol proposes a comparison with existing review practice, measuring decision quality, detection of critical failures and review burden. The evidence supports an account of how to apply and evaluate S4. Its effect on deployment decisions and harm remains an open empirical question.

Keywords: AI safety, alignment, permissions, assurance, human oversight

1 Introduction

A team is deciding whether an assistant may draft a customer message, place it in a review queue or send it. Each permission changes who can be affected and when someone can intervene. A good draft supports a claim about drafting. Sending also requires evidence about recipients, source currency, approval authority and the ability to stop queued messages. S4 starts with the permission under review so the recommendation answers the decision the owner faces. [1, 2]

Alignment here means that behaviour serves an authorised purpose within stated constraints, including when records conflict or the goal is underspecified. Safety concerns harm to people and systems. A service can follow its assigned objective and still cause harm if the objective rewards a poor outcome. Work on objective design, supervision and behaviour in unfamiliar conditions provides context for both problems. [1, 3]

Research question. Does a permission-level S4 review help teams identify more consequential gaps and make better scoped decisions than their existing review process, within a tolerable review burden?

Contribution. The paper turns the published S4 guide and canvas into a testable protocol. It identifies what an evaluator must record, applies the protocol to selected evidence and defines what

would justify revising it. NIST's AI Risk Management Framework and generative AI profile provide broader governance and risk management context. The organisation of the paper follows Liao, Shan and Chatterji's study of scientific workflows. [1, 4, 5, 9]

2 Data and measurement

2.1 Sources and case selection

The source for the four stages and their order is the published S4 guide and companion essay. The case analysis uses the British Columbia Civil Resolution Tribunal decision in Moffatt v. Air Canada and the United States Securities and Exchange Commission order on Knight Capital. These were purposefully chosen to examine two pathways to harm: reliance on a customer-facing answer and rapid execution through delegated authority. The second case concerns rule-based automation. Its relevance is the control of action speed and exposure. [1, 2, 6, 7]

For each case, the analysis extracts the documented behaviour, what made it consequential and what a reviewer would have needed to inspect under S4. Paragraph numbers point to the underlying evidence. The cases are publicly documented examples chosen after the outcomes were known. They support a retrospective interpretation with possible hindsight bias. Appendix A gives an evidence record another reviewer can use to repeat the mapping.

2.2 One permission in one operating context

The unit of assessment is a permission for a named workflow and version in a specified operating context. Record the actor, the action, the data or resource, the circumstances and the owner who may authorise it. Give related permissions separate records when evidence or consequences differ. For instance, reading an approved bulletin and placing a draft in a queue can belong to the same workflow and still require distinct decisions.

Record timestamped test and incident evidence against the relevant version of the model, policy, source set, tools and access controls. Treat a missing result as unknown. Record a negative test as failed only when it runs and produces the defined failure. An attempt denied before it has an effect and an action that reaches its destination need separate measures. Include both when they inform the review.

Table 1: Proposed measures. Report denominator definitions and relevant case groups.

Measure	Operational definition
Critical failures	Runs with a predefined critical failure divided by eligible completed runs for that permission. Report unrun, incomplete and inconclusive cases separately.
Unauthorised actions	Count and rate of out-of-scope actions reaching the destination, per eligible action opportunity. Log blocked attempts separately.
Review burden	Active reviewer time per completed permission decision, including time to find evidence, adjudicate conflicts and record the rationale.
Containment time	Elapsed time from a detection trigger to confirmed suppression of new and pending actions. For financial exposure use the full interval from the first possible harmful commitment.

3 The S4 review method

3.1 From purpose to evidence

Specify. Name the outcome that matters to the people who use or are affected by the workflow. List approved inputs, permitted actions, prohibited actions and decisions reserved for a person. Set an approval owner, a suspension owner and an expiry or review date. State what the workflow does when sources conflict and how a person can request a correction.

Stress-test. Write the expected result and critical failure rule before each test. Include normal use, missing or contradictory information, changed conditions and unavailable tools. Add cases from the intended user population and from the most consequential failure paths. Hold back cases for independent assessment so revisions can be evaluated on new material. Record the observed result and the version used.

Safeguard. Put the control where the action becomes effective. Use access restrictions, approval requirements, enforceable limits and a way to pause the workflow or cancel queued work. Test the manual process as well as the stop. For each control record the owner, the observation that shows it worked and what happened when a dependency failed.

Scrutinise. Review corrections made before use, failures found after use and complaints from people affected. Choose the reviewer and give that person a means to escalate or restrict permission. Set the interval, denominators and triggers. Reopen the relevant checks after material changes to sources, data access, workflows, models, populations or tools. [1, 2]

Specify $\longrightarrow$ **Stress-test** $\longrightarrow$ **Safeguard** $\longrightarrow$ **Scrutinise**

Figure 1: The S4 review cycle. A material change or a new failure sends the affected permission back to Specify.

3.2 Decisions and containment

Approve only the tested scope and give it an owner and review date. Use a bounded trial when critical controls pass and the remaining uncertainty can be studied within an authorised scope. Hold the affected permission when evidence is missing, a critical failure remains, or the owner cannot demonstrate an applicable stop or fallback. Make the conditions for restart explicit and retain the reviewer's reasons, including dissent.

For a financial action, let a be the assumed maximum actions per minute, t the minutes from the first possible harmful commitment to effective containment, and c the maximum commitment per action. Define the direct exposure envelope

$$E = atc, \qquad E_C = \min(atc, C),$$

where C is a verified aggregate cap covering the relevant commitments. Check whether queued work, retries and concurrent workers count against it. Report the assumptions and units. The equation gives direct commitments at risk under the specified stress scenario. Assess privacy, fairness and irreversible consequences using separate evidence. [2]

4 Case analysis and worked application

4.1 Reliance on an answer and authority to act

In Moffatt v. Air Canada, the tribunal describes a consumer who received chatbot advice suggesting a retroactive bereavement fare request was allowed. The referenced policy page stated that the policy did not allow requests after travel. The tribunal found that the customer reasonably relied on the chatbot and held Air Canada liable for negligent misrepresentation in that dispute. The decision documents the user-facing conflict and its consequences. It provides limited detail about the chatbot's architecture and internal controls. [6, paras. 2–4, 18–22, 24–32]

For S4, the relevant permission is giving customer-facing eligibility and deadline guidance. Specify which record governs each answer. Stress-test questions that include conflicting dates and a policy update. Safeguard answers when the source is stale or conflicting by routing those cases for staff review. Scrutinise corrections and complaints by policy version. This mapping shows what a reviewer would ask for. It leaves the effectiveness of the proposed controls unmeasured.

In a separate case, the SEC found that on 1 August 2012 Knight Capital's automated order router sent millions of orders over 45 minutes while handling 212 small retail orders. The SEC reported losses exceeding US$460 million. The order documented weaknesses in deployment and market-access controls. The scope of this case is automated trading, where execution speed made the consequences of a defect difficult to contain. The SEC also described aggregate controls that could not prevent orders beyond firm-wide thresholds. [7, paras. 1–4, 12–16], [8]

> "safety nets are in place to limit the harm" Daniel M. Hawke, then chief of the SEC Enforcement Division's Market Abuse Unit, asking firms to examine their controls [8].

S4 would ask the owner to define order-entry and deployment permissions, test an incomplete deployment and examine whether the stop reaches all active paths. The team would also need to reconcile submitted orders, executions and internal exposure under degraded conditions. These are evidence requests derived after the event. The public record supports no estimate of a counterfactual loss under S4. [7]

4.2 A draft-only Caribbean climate service

Consider a team that prepares plain-language summaries from approved published rainfall outlooks. The system may read a fixed source list and save drafts internally. A named climate scientist checks the issue date, area, units, period and whether the source describes a forecast or a projection. An authorised editor controls publication. The drafting account has no publishing or alert permissions. This scenario illustrates the method. It uses no live forecast, assessment of a specific organisation or operational test. [1]

Use two bulletins with different issue dates and an archived outlook that uses a different period. Set the expected result in advance: retain each document's status and period, flag the conflict and hold the disputed claim for review. Withhold required evidence in a second case and record whether the system refers it for review. For containment, test whether the service owner can pause drafts and whether staff can prepare the summary directly from the approved source. Appendix B shows the evidence fields.

5 Prospective evaluation

5.1 Compare S4 with existing practice

Recruit teams that already review narrowly defined AI permissions. Start with a feasibility phase covering drafting, recommendation and limited execution in an isolated environment. Document the usual review process at each site before assignment. Give reviewers the same case materials and tools. Use matched scenario sets with different cases, and randomise assignment by reviewer team to reduce spillover. Stratify assignment by domain and experience when sample size allows. Use the feasibility phase to estimate review time, variability and recruitment. Set the confirmatory sample size from those estimates and a pre-agreed minimum effect worth detecting.

A panel independent of the reviewers should agree on the critical issues and defensible permission decisions before anybody starts. The panel should include workflow expertise and relevant experience of how decisions affect users. Preserve disagreements. Mask adjudicators to study arm where possible, using anonymised records in a common format. Check whether the common format omits evidence that could matter to adjudication.

The primary endpoint is the share of permission decisions that both identify all reference critical failures and limit authority to the documented evidence, divided by all assigned decisions. Classify an incomplete review as unsuccessful for the primary analysis and report missingness by arm. The rubric must distinguish a justified hold from needlessly blocking a supported permission. Report the two components separately. Include active reviewer time, unjustified holds, missed critical issues and fallback success as secondary measures.

5.2 Analysis and criteria for revision

Report the absolute difference between arms with an interval that accounts for assignment clusters. Present raw numerators and denominators by workflow and report agreement before adjudication. Where group counts are small, present descriptive counts, identify the uncertainty and avoid ranking groups. Treat the controlled scenario measures as evidence about reviewer decisions in the study. An estimate of field harm would need evidence from actual deployment.

Preregister the analysis, permitted exclusions and minimum benefit before studying outcomes. Set a maximum acceptable increase in reviewer time with participating teams. The case for adopting S4 strengthens if the primary result clears the agreed benefit threshold without excessive burden or new critical misses. If it fails those criteria, use reviewer notes and disagreement records to identify stages to simplify or revise, then run a fresh comparison.

A subsequent pilot could track valid complaints, detected incidents and time to contain failures during limited operation. Report who observed what, changes to the workflow and reporting coverage. Fewer reports could reflect fewer incidents or weaker reporting. An independent sample of records would help distinguish those explanations.

6 Scope and interpretation

S4 v0.2 is a framework proposal. The public cases illustrate pathways a review might miss and the climate service shows how to record a decision before testing. Evidence for adoption depends on the prospective comparison, particularly whether reviewers find critical gaps without increasing

delays enough to make the workflow impractical. Use the study to decide what to keep, revise or remove from the protocol.

References

[1] Adrian Dunkley. S4 AI Safety Framework. Proposed framework v0.2, 2026.

[2] Adrian Dunkley. An uncontrolled acceptance of AI. Opening essay and permission canvas, 2026.

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety. arXiv:1606.06565, 2016.

[4] Elham Tabassi. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1, 2023.

[5] Chloe Autio et al. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1, 2024.

[6] British Columbia Civil Resolution Tribunal. Moffatt v. Air Canada. 2024 BCCRT 149, 2024. Decision reproduced by The Sedona Conference.

[7] U.S. Securities and Exchange Commission. In the Matter of Knight Capital Americas LLC. Release No. 70694, 2013.

[8] U.S. Securities and Exchange Commission. SEC Charges Knight Capital With Violations of Market Access Rule. Press release 2013-222, 2013.

[9] Chris Liao, Carl Shan, and Aaron Chatterji. Science at the Frontier: How Scientists Use ChatGPT. 2026.

A Evidence and reproducibility

For each case, retain the document identifier and date, paragraph number, observation, permission affected, applicable S4 stage, proposed test and open question. Store the source retrieval date and a stable copy or link. A second reviewer can repeat the mapping against the primary records and log alternative interpretations. Release a de-identified case register, scoring rubric, exclusions, anonymised review summaries and analysis code when the study permits. This preprint uses public records and an illustrative case. Evaluation data will depend on the proposed study.

B Permission and test records

B.1 Permission decision record

Table 2: Fields for a permission decision.

Field	What to record
Scope	Workflow, permission, affected users, authorised sources and actions, reservations, effective date and expiry.
Claim and evidence	Proposed authority, evidence and versions, evidence that would overturn the claim, gaps and dissent.
Controls and fallback	Approval point, access limit, stop owner and evidence, queue behaviour and manual route.
Decision and restart	Hold, bounded trial or approval. Record owner, rationale, conditions, review date, suspension triggers and restart authority.
Recourse	How users report errors, who replies and how records are corrected.

B.2 Example stress-test register

Table 3: Proposed stress tests for the illustrative climate service. Record the run date, versions, actual result, evidence and reviewer for each executed case.

Case	Expected behaviour	Critical failure
Conflicting outlooks	Preserve issue dates and periods. Refer the disputed claim.	An unresolved figure appears as current and definitive.
Required source missing	Mark the claim unavailable and route it for review.	A claim appears without traceable support.
Publication rights	Keep the draft internal and log the blocked request.	The drafting account publishes the draft.
Queue held	Pause drafts and demonstrate the manual route.	The queue advances after suspension or the manual route fails.

B.3 What to record after release

For each review period, list the eligible drafts and the number inspected. Count corrections made during review, corrections made after publication, reported harms and any changes in permission. Include a source for the denominator. Log how people raised concerns and what action was taken. If a draft crosses the publication boundary or a critical control fails, suspend the affected permission and route work through the manual procedure. Verify the fix and record the restart decision before restoring access.