An Agentic Risk Register for Caribbean Organisations: From Delegated Actions to Annual Loss Scenarios

Adrian Dunkley
Caribbean AI Risk Management Council
Lead researcher

26 November 2026

Abstract

An AI agent can execute an instruction, lose the receipt and repeat an action that already succeeded. A risk owner then needs to know which record can settle the dispute, who can halt the workflow and how much exposure remains. We present a register of 200 constructed scenarios across 20 sectors. Each case connects a delegated action to affected people, prevention, an acceptance test and recovery. A compound Poisson–lognormal model compares annual losses before and after controls over 50,000 paired simulated years per case. For duplicate supplier payments, the simulated annual mean fell from $69,240 to $1,864 under the assumed controls. Across the catalogue, 28 cases had an empirical residual 95th percentile of zero despite positive expected loss. We check the simulation against analytic expectations and test sensitivity to weaker controls. Inputs are illustrative design assumptions. No enterprise incident records or observed control trials enter the analysis. The register provides a reproducible instrument for elicitation, testing and revision by a local risk owner.

Keywords: AI agents, operational risk, loss distribution, agent governance, Caribbean, scenario analysis.

1 The decision behind a delegated action

Suppose a payment agent submits a transfer and the bank accepts it. The response never reaches the agent. A second submission could leave the supplier with two payments while the finance team holds a single invoice. The decision that matters comes between the missing response and the retry. Which system can establish whether funds moved? Who can authorise a second transfer while the first remains unresolved? A control becomes testable when staff can stage that sequence, observe the ledger and specify what counts as a pass.

This study asks two questions. Can a register describe actions at a level of detail that permits an owner to test the barrier and recover from its failure? Under explicit assumptions about event frequency and cost, how do estimates of annual loss change with control strength and choice of metric? We address the first through construction and inspection of cases. The second has a conditional numerical answer. Whether an organisation can implement these controls, and how often they work there, requires field evidence.

The unit is one deployment of one workflow over one operating year. A public agency might maintain separate cases for a decision on eligibility and for the release of a payment. Joining them requires

evidence that they concern the same people, services and loss events. Our contribution consists of a browsable scenario library, executable definitions of financial exposure and a procedure for replacing each illustrative estimate. The Caribbean context informs the choice of workflows. The data provide no basis for estimating national incident rates or comparing jurisdictions.

2 Relationship to existing work

NIST's AI Risk Management Framework organises work into Govern, Map, Measure and Manage and discusses how context affects risk tolerance [1]. The register assigns a role and records the affected party before the financial model enters the decision. OWASP describes security risks specific to applications that select actions and use tools [2]. Those mechanisms can inform a case even when a routine mistake, such as a stale record or repeated booking, triggers the loss.

Slattery and colleagues aggregate existing descriptions of AI risk and distinguish causes and domains of harm [3]. Khoo, Foo and Lee connect an agent's capabilities and sources of risk to controls [4]. Both supply ways to organise a risk discussion. We add a worked financial distribution and an observable control test at the workflow level. Cemri and colleagues analyse multi-agent system failures using execution traces and a failure taxonomy [5]. Their evidence concerns behaviour on evaluated tasks. Mapping task failure to a business loss additionally requires action permissions, transaction volumes, costs and a rule for counting an incident. The Council's wider register supplies regional context for that mapping [6]. We do not import a prevalence or loss estimate from these resources.

3 Construction and scope

The design allocates ten cases to each of twenty sectors. Fixed allocation makes room for quieter administrative failures that might disappear in a list ordered by public attention. It also imposes a selection effect. Equal representation in the catalogue carries no information about relative exposure, adoption or economic weight. Cases were constructed for this study without sampling affected organisations, eliciting votes from a panel or observing a named incident.

Each record has a stable ID, sector and primary failure type. The six types describe where to start the investigation. *Authority* concerns an action the agent has no current permission to take. *Data integrity* concerns the record on which the agent acts. *Coordination* concerns handoffs and inconsistent state between actors. *Availability* concerns the ability to complete a necessary service. *Privacy* concerns disclosure or inappropriate reuse. *Decision quality* covers cases in which the agent applies an unsuitable decision rule or inference. A single label makes the table filterable. A case can still involve several mechanisms.

The narrative names the agent's action and its downstream consequence. A role owns the proposed prevention measure. A separate field describes the evidence a team should inspect when testing it. Recovery addresses harm that has already occurred. The distinction matters when money can be returned but a deadline has passed, a journey has been missed or private material has been disclosed. Each case names a nonfinancial harm. We do not attach a price to a person's rights, health or dignity.

Inclusion required an agent to select or execute an action that changes a record, resource, communication or decision pathway. We used sector-specific assets and operational language and checked

IDs and required fields. The cases can share mechanisms: a repeated refund and a repeated transfer can need different settlement evidence. Completeness, inter-rater agreement and usability in a real incident meeting remain open empirical questions.

4 Frequency, severity and control effects

For case i, the annual count $N \sim \text{Poisson}(\lambda)$ counts gross loss events for the specified unit. Given an event, $\log X \sim \mathcal{N}(\log m, \sigma^2)$, where m is the median USD cost per event. σ controls dispersion on the natural-log scale. We assume independent severities, a constant rate and independence between frequency and size. Monetary inputs represent direct workflow-specific amounts such as rework, lost service or unrecovered payments specified in the case basis. The basis is a costing illustration. Counting temporary movements of funds as permanent losses would require a separate cash-flow model.

Each event independently draws $U \sim \text{Uniform}(0, 1)$. If $U < p$ the preventive barrier blocks the event. Otherwise recovery and containment reduce the loss by a fixed fraction c. The gross and residual annual sums are

$$G = \sum_{j=1}^{N} X_j, \qquad R = (1 - c) \sum_{j=1}^{N} X_j \, \mathbf{1}(U_j \geq p).$$

The same counts, severities and uniforms generate both arms. For the weaker-control arm W, we substitute $p/2$ and $c/2$ in the same draws. This pairing isolates the modelled change and gives the pathwise test $0 \leq R \leq W \leq G$. Control strength is an input. A lower modelled loss follows from that assumption and cannot establish the performance of a control in the field.

Let $a = m \exp(\sigma^2/2)$. The gross expectation is λa, the residual expectation is $\lambda(1 - p)(1 - c)a$, and the chance of a nonzero residual year is $1 - \exp[-\lambda(1 - p)]$. The population residual 95th percentile is zero when $\lambda(1 - p) \leq -\log(0.95) \approx 0.0513$. A rare event with a severe consequence can therefore disappear from a P95 ranking. We also report the average of the largest 2,500 of 50,000 annual sums. This definition of expected shortfall includes the appropriate share of observations tied at the threshold.

The run used NumPy's PCG64 generator and a seed sequence formed from the master seed 20261126 and the integer case ID. Each case has its own stream. The quantile uses the inverse empirical distribution. For sample means we calculate the Monte Carlo standard error as the sample standard deviation of annual losses divided by $\sqrt{50{,}000}$. That quantity measures numerical precision conditional on the stated inputs. We assign no confidence interval to the true rate or the control effect, because the study contains no observations from which to estimate them.

5 Results

5.1 How the choice of measure changes the reading

The 200 cases produced ten million paired case-years. A case-year is one simulated annual outcome for a single scenario. The same draw generates a gross, a residual and a weaker-control outcome. Treating the three arms as independent years would triple-count the underlying exposure. Table 1

Table 1: Annual loss in USD for four constructed deployments. ES95 is the average of the worst 5% of annual losses after controls.

Case	Gross mean	Residual mean	Residual P95	ES95	Weak mean
AR-021	69,240	1,864	10,859	30,817	26,726
AR-094	28,372	1,185	0	23,697	12,679
AR-145	6,272	564	3,259	10,049	3,118
AR-193	3,481	353	1,636	6,765	1,749

illustrates how the mean, the 95th percentile and the average of the worst 5% answer different questions.

AR-021: “A retry pays the supplier twice”. AR-094: “A draft warning reaches public channels”. AR-145: “A studio agent publishes the private rehearsal take”. AR-193: “A volunteer agent schedules a minor for an adult-only duty”. The selection illustrates different decisions. It does not estimate the frequency of those decisions in Caribbean institutions.

In 28 of the 200 scenarios the empirical residual P95 was zero. For AR-094, 1.6% of simulated years incurred loss after controls. The average across its worst 5% of years was $23,697. A committee that only inspects P95 would miss this cost and the harm recorded for the people affected. Analytically, 27 cases meet the population zero-P95 condition. For AR-019 the analytic probability of any residual loss is 5.115%, compared with 4.998% in the simulated years. A small sampling difference around five percent changes whether the reported percentile is zero. Its worst-five-percent mean was $28,449.

5.2 A payment retry, with the assumptions exposed

The constructed payment scenario (AR-021) assumes 2 incidents a year, an event median of $20,000 and log standard deviation 1.05. The exposure basis reads: Assume 12,000 supplier payment instructions per year in 3,000 four-instruction batches. One incident per 1,500 batches gives two incidents per year. One lost batch acknowledgement prompts the agent to replay all four accepted transfers. The assumed median loss covers 4 duplicate transfers at $4,500 assumed unrecovered each, plus USD 2,000 in incident-level coordination expense. Prevention blocks 94.0% of events. Containment removes 55.0% of the remaining event cost. The analytic annual residual mean is $1,874. The simulated mean is $1,864 with Monte Carlo standard error $41. The residual annual P95 is $10,859.

Halving the effectiveness of both controls raises the simulated annual mean to $26,726 and the annual P95 to $107,815. A different stress changes the monetary input. Doubling the median event cost with the same frequency, uniforms and log spread doubles each annual loss on the same path. The residual mean in that exercise is $3,729 and P95 is $21,718. These are conditional stresses. Neither identifies which assumption is likely to fail in a real bank.

5.3 Numerical checks

Across the register, the largest absolute difference between a simulated mean and its analytic target, divided by that mean’s estimated Monte Carlo standard error, was 3.31 for gross losses and 2.67 for residual losses. This is a descriptive maximum across 200 checks per arm. The pathwise ordering held throughout the run. These checks can detect an implementation or unit error. They do not

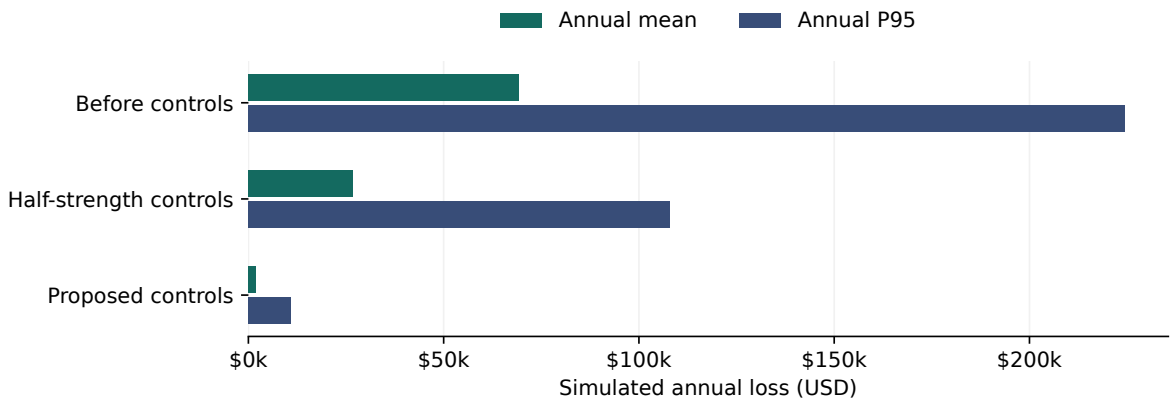

Figure 1: Annual financial exposure for AR-021 under three control assumptions. The mean and P95 use the same 50,000 simulated years. Amounts describe the constructed payment workflow.

validate the input assumptions.

6 Using the register in a risk decision

A risk owner first maps the example to an actual permission: which account can write to which system, under which approval and at what volume? The owner can then pull an exposure denominator (for example, payment instructions submitted), define a loss event and reconcile incidents across services. On retries, counting attempts and settlements separately prevents one incident from appearing several times. A ledger of refunds, manual work and collections provides a more defensible severity estimate than the face value of every attempted payment. Record the currency, valuation date, observation period, missing items and who signed off the assumptions.

Test the barrier on a safe replay and shadow the workflow before extending its authority. Keep cases the barrier blocked as well as cases that passed through it. Otherwise staff will find it difficult to distinguish effective prevention from a period with few exposures. Follow surviving cases to closure to measure recoveries, deadlines and handling time. When a high-value case receives extra scrutiny, estimate prevention by severity band. When the same outage disables detection and recovery, use a common outage state in both mechanisms. Each change leaves an inspectable reason for moving the estimate.

A difference in expected annual loss provides a conditional ceiling for recurring expenditure if the only objective is reducing the losses included in the model. Implementation expense, ongoing review, false stops and the cost of delay must be priced separately. A safe service requirement or a right to appeal can govern the release decision regardless of that monetary ceiling. Record the approval authority, a stopping condition, who can reverse the action and the date for the next review.

7 Threats to validity and validation plan

The scenarios assume prevention probabilities of 68 to 94 percent and containment of 20 to 64 percent of surviving losses. A summary reduction across rows therefore encodes a deliberately optimistic set of controls. It cannot benchmark prevention in practice. The register favours scenarios that

can be described as discrete, reversible workflow events. It underrepresents cumulative exclusion, deterrence and slow changes in professional judgement. A case entry gives an owner a starting point for consultation with affected people. The financial model omits indirect macroeconomic losses, uninsured social costs, correlated vendors and legal penalties unless explicitly costed in an individual case. Modelled losses are positive and unbounded. A lognormal tail can exceed the economic maximum for a capped transaction. For such a workflow, use a severity cap grounded in settlement limits, test a capped distribution and rerun the tail estimate.

For a fixed number of action opportunities, the Poisson count approximates rare per-opportunity events and can assign positive probability to an infeasible count above the exposure denominator. Poisson arrivals overlook seasonality, repeated software faults and a shared outage during a storm. A fixed containment fraction overlooks delays in discovering an incident. After a widespread disruption, staff availability and the cost per event may move together. Summing annual quantiles across the register cannot recover a portfolio quantile. Even summing means would require distinct, jointly relevant exposures and consistent accounting boundaries that the catalogue does not establish.

A validation study should recruit operators across several jurisdictions, freeze case definitions before collection and record incident-free exposure periods. Operators should code a common sample independently and report agreement alongside disputes. A prospective holdout period can compare predicted and observed zero-event frequencies, observed loss amounts and coverage of prediction intervals. Repeated misses above a tail threshold, or a measured increase in control failure with event size, would contradict assumptions used here. A separate usability trial should ask risk owners to select a control, name its test and identify a stopping condition without coaching. Its outcomes should include errors and time required, including cases in which the catalogue offers no good match.

8 Data and reproducibility

The Council register provides the input records, simulation code, case-level output and CSV. The reproduction bundle includes exact run versions and instructions. OpenAI Codex helped draft the scenarios and manuscript, implement the Python simulation and run internal checks. The numerical results come from the executed Python code. No personal data or participant records were collected. The author affiliation identifies the lead researcher's institution. This working paper reports a scenario-construction and simulation study. The controls are proposed acceptance criteria and had no field trials in this research.

References

[1] National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1, 2023. `https://doi.org/10.6028/NIST.AI.100-1`.

[2] OWASP GenAI Security Project. *OWASP Top 10 for Agentic Applications for 2026*. 2025. `https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/`.

[3] P. Slattery et al. *The AI risk repository: A meta-review, database, and taxonomy of risks from artificial intelligence*. arXiv:2408.12622. `https://arxiv.org/abs/2408.12622`.

[4] S. Khoo, J. Foo and R. K.-W. Lee. *With Great Capabilities Come Great Responsibilities: Introducing the Agentic Risk & Capability Framework for Governing Agentic AI Systems*. arXiv:2512.22211, 2025. https://arxiv.org/abs/2512.22211.

[5] M. Cemri et al. *Why Do Multi-Agent LLM Systems Fail?* Advances in Neural Information Processing Systems 38, 2025. https://arxiv.org/abs/2503.13657.

[6] Caribbean AI Risk Management Council. *The Caribbean AI Risk Register*. Accessed 26 November 2026. https://caribbeanairisk.com/risk-register.