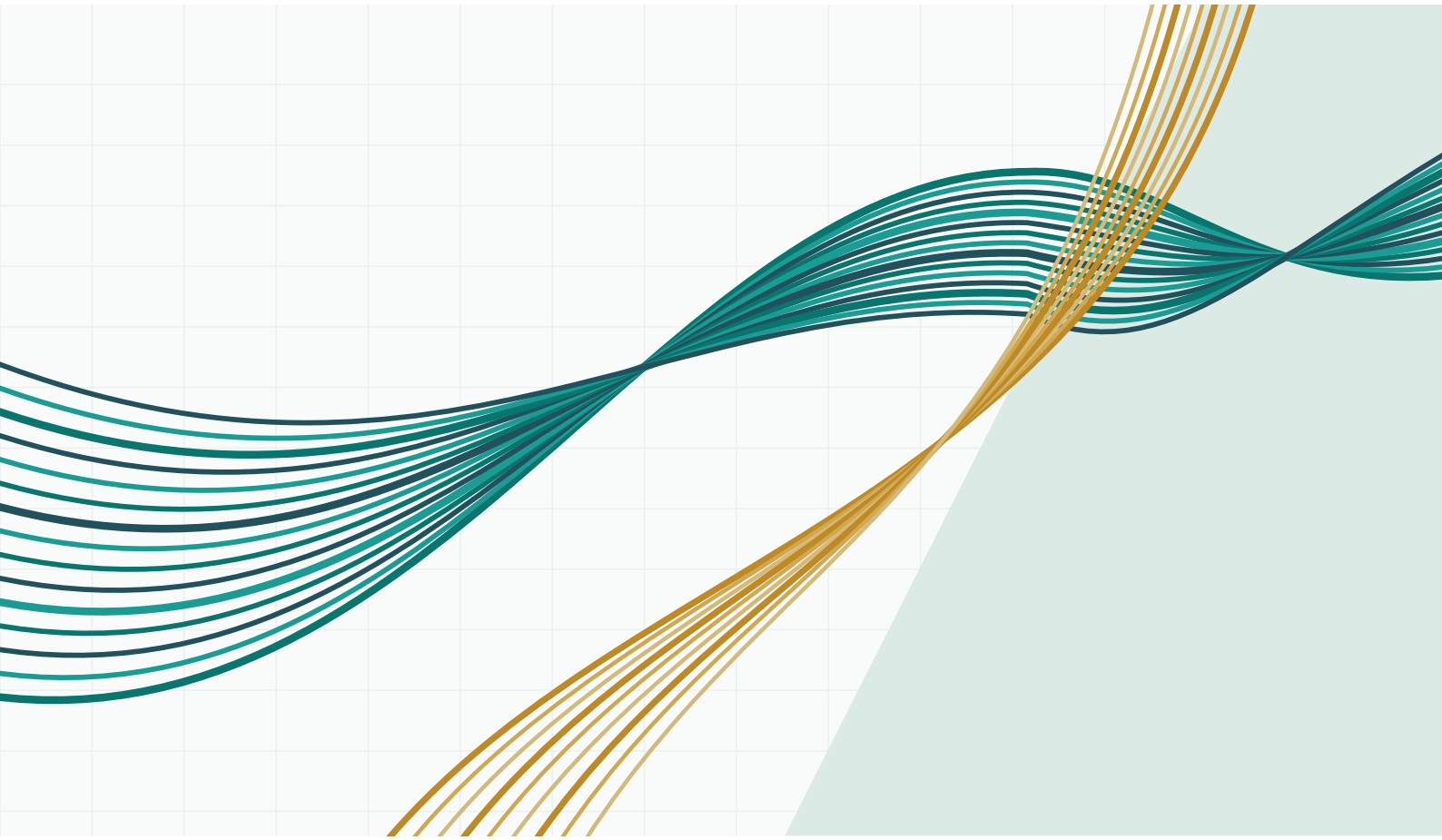
CAIA
CARIBBEAN ARTIFICIAL INTELLIGENCE ASSOCIATION
CAIA

RESEARCH & POLICY SERIES

When is an AI system safe enough to use?

Evidence for Caribbean deployment decisions

Lead author Adrian Dunkley



NOVEMBER 2026 EDITION

23 November 2026

OVERVIEW

Abstract

Lead author Adrian Dunkley, Caribbean AI Association

Research & Policy Series · November 2026 edition · 23 November 2026

An accurate transcript can still send an urgent call to the back of a queue. Caribbean institutions deciding whether to use artificial intelligence (AI) need evidence about what happens to the caller after the system produces its answer. This whitepaper proposes a way to make that decision. It draws on regional policy, selected national laws and regulatory guidance, Caribbean language research and technical literature on AI risk. The proposed method asks service owners to specify the outcome people should receive, test the mistakes that would put it at risk and show how a person can obtain a correction. Three worked examples examine credit applications, public warnings and school admissions. A field protocol specifies how to compare the service with its existing process, including cases that return to staff. Field studies are needed to establish whether the method reduces harm in practice.

Executive decision

Authorise a defined use only when the institution can explain its benefit, establish its authority to use the data and produce local evidence that the service works under the proposed conditions. Someone on duty must be able to stop an unsafe action and put the affected person’s case right. If the evidence applies only to one task, language or group of users, restrict approval accordingly and maintain a staffed route for everyone else. An unresolved legal question or a failed safety check calls for a hold.

Reading guide

Introduction 03
Method and regional evidence 04
Define the decision and its owner 05
Evaluate local performance and set release limits 06
Three Caribbean deployment decisions 07
Implementation, field validation and limitations 08
Conclusion and release record 09
References 10

THE DECISION

Introduction

Consider a housing agency that asks its staff to close more applications each week. It introduces a system that flags incomplete files for return. Each time a family resubmits a form and receives it back, the agency records another closure. Its report looks better. The family is still waiting.

This example illustrates a familiar problem in AI safety research. A measure that stands in for a goal, known as a *proxy*, can reward conduct that defeats the goal [1]. The agency can address that problem before buying or training a model. It can follow each eligible household to a correct decision and count return visits, including cases that remain open. The point of the service then becomes visible to both staff and applicants.

Safety concerns the harm a system could cause in use and the safeguards that prevent or limit it. **Alignment**, as used here, asks whether the system's actions serve an agreed purpose and stay within the authority granted for that service. A housing officer may value a shorter queue while an applicant values a fair chance to complete the file. Where those interests conflict, an authorised person should hear the people affected and record the decision on priorities.

The unit of approval is the whole service. A ranking model, the data it uses, the permissions on its account and the people who act on its recommendations can each change what happens to a resident. Adding permission to send an email changes what an assistant can do even when the model stays the same. Caribbean institutions using a common supplier may still have different legal duties, working languages and routes for appeal. The method applies to predictive and generative systems deployed in those settings. Assessing the full capabilities of a general-purpose model is a separate scientific undertaking [2].

The question for a future field study is whether this method reduces consequential errors that reach people while preserving their access to timely service.

EVIDENCE AND SCOPE

Method and regional evidence

The paper uses a targeted review of material available by 23 November 2026. It gives priority to original research and official sources. Each type of evidence answers a different question. A statute can establish a duty. A regulator can explain how it expects supervised institutions to act. A language study can establish what researchers measured on a particular task. CAIA used these distinctions to design the release procedure and worked examples. The review included no surveys or evaluations of operating services and yields no estimate of how common unsafe systems are.

The Caribbean AI Policy Roadmap identifies culture, governance, education and resilience as regional priorities. UNESCO describes consultations across Dutch- and English-speaking countries [3]. A service used by speakers of Haitian Creole or Spanish needs evidence from those communities too. The 2026 Santo Domingo summit adopted a Latin American and Caribbean roadmap for cooperation among authorities and sectors [4]. Both roadmaps inform policy. CAIA's model AI Act is a separate consultation proposal [5].

For a legal decision, identify the service, its operator, the applicable territory and the provisions in force. Saint Lucia's revised laws show why this matters. The commencement note names the parts of its Data Protection Act that a January 2023 instrument brought into force. Read that note alongside the relevant section and check for later instruments [17]. Applicable domestic and cross-border rules determine the obligations in each case.

The US National Institute of Standards and Technology (NIST) provides a voluntary AI risk management framework. Its companion profile covers risks from generative systems, including unsupported answers and dependence on outside suppliers [6, 7]. UNESCO's ethics recommendation addresses human oversight and public participation [8]. CAIA draws on these sources in asking an institution to identify who can challenge its decision and what evidence that person can inspect.

Armstrong and colleagues developed a Jamaican Patois benchmark that tests whether one statement follows from another [9]. CreoleVal covers up to 28 Creole languages across eight language-processing tasks, with different languages represented in each task [10]. These studies offer starting material for technical evaluation. An emergency office still needs to find out whether a local reader understands where a warning applies, what action to take and when the warning expires.

PERMISSIONS

Define the decision and its owner

Begin with the change a person will experience. A patient who receives a corrected appointment notice after the clinic closes has still missed the appointment. Ask a frontline worker and a person who uses the service what a mistake would cost them. Use their answers to define the acceptable outcome, who may change the decision and where someone can challenge it.

Match permission to the action

Proposed action	Authority to grant	Evidence required
Draft a case summary	Read selected records and write in a private draft area	Staff can trace each material statement to its source. The institution removes drafts on schedule.
Recommend priority	Suggest a position in a queue	A reviewer can inspect the source file, change the priority and find cases approaching the service time limit.
Carry out a routine action	Act on an approved destination within a stated limit	A permission check outside the model blocks other actions. An authorised person can stop, trace and reverse the change.
Affect eligibility, safety, money or legal rights	CAIA recommends reserving the final decision for a qualified person	The reviewer has time and discretion. An affected person can reach an officer with authority to correct the outcome.

A request to exercise restraint offers little protection if an assistant's account can change a balance or publish a notice. Set permissions separately for reading, drafting, recommending and taking action. In a rehearsal, give the system a document containing misleading directions and check that the document cannot expand those permissions. Then give a reviewer a plausible but wrong recommendation and access to the correct source. Observe whether the reviewer catches it during a normal shift and can reject it without penalty.

Trace one request through the supplier

Follow an actual type of record from collection to deletion, including its attachments, prompts, logs and backups. Identify any subcontractor or remote support worker who can inspect it. At each destination, record which fields go there, the reason for sending them, the applicable legal basis, who has access and when the copy is due for deletion.

Find out whether the supplier may reuse inputs, which settings it can change without notice and how the institution can recover its records when the contract ends. Keep live personal records out of a trial while a material answer is missing. If the supplier can update a model without giving customers the option to keep an approved version, assign an officer to watch for updates and suspend sensitive functions until the new version passes its tests.

EVALUATION

Evaluate local performance and set release limits

Set aside cases the development team has not used. Obtain permission before testing with real records. As well as routine requests, include conflicting documents, a change of language, an unavailable supplier and a customer asking to correct a record. Define a safe response for each case before testing. Faced with two different appointment times, for example, the system should flag the disagreement for confirmation before telling someone to travel.

Count the consequences

A missed urgent case can leave someone waiting for help. Too many routine cases marked urgent can delay everyone already in that queue. Measure both errors against the existing service using the same eligibility rules. Report time to a correct decision, unresolved cases and people who abandon each channel. A rate needs a numerator, a denominator and a time period. Say whose outcomes are unknown. Compare results across languages and channels when collection is lawful and the groups are large enough to protect privacy. An authorised reviewer can inspect a problem that is unsafe to report publicly.

Suppose reviewers find no critical failures in a random sample of 100 independent cases. With a constant failure probability and complete detection, the exact one-sided 95% binomial upper confidence limit is about 2.95%. If p is the probability that one case fails, the chance that all 100 pass is $(1 - p)$ to the power of 100. Setting that chance equal to 0.05 gives the upper limit. This sample alone cannot establish that the failure rate is below one in 100. As James Hanley and Abby Lippman-Hand observed, “a zero numerator does not necessarily mean ‘no risk’” [11].

Difficult cases chosen deliberately can reveal a broken safeguard. Estimating the failure rate across the service requires a sample that represents its users. Repeating one case, or counting decisions made during the same outage as independent observations, can give a misleading impression of precision.

Agree when to stop

Before examining the results, name a failure serious enough to stop the affected function. Examples include a payment made without approval, a public warning issued for the wrong district or another person’s file appearing on screen. Name the officer and deputy who can disable the function and the person who will take over unfinished cases.

Test recovery under an ordinary operating constraint, such as the supplier being unreachable. On reconnecting, staff should match each pending request to its acknowledgement before resending it. Check that each person receives the action once and that no case disappears. A team with too little evidence can continue a lower-risk trial. A wider release needs evidence for the additional use.

WORKED APPLICATIONS

Three Caribbean deployment decisions

The following scenarios apply the method to proposed services. The cited laws, regulator guidance and technical standards are the sources an institution would examine.

A Jamaican lender reviews applicants without payslips

A lender proposes using a model to route applications from people who cannot provide a conventional payslip. Its risk officer wants to know whether those applicants end up in a queue that seldom receives attention. Jamaica's Data Protection Act 2020, section 12, addresses significant decisions based solely on automated processing of personal data to evaluate a person, subject to exemptions. It provides for written notices and reconsideration in specified circumstances. Jamaica's Information Commissioner gives guidance on requests and complaints [12, 13].

A credit officer can examine other evidence, change the decision and respond to an applicant who contacts the lender outside the model's interface. Compare waiting times and reversals alongside approval rates. If staff routinely accept the recommendation without opening the file, suspend it and examine who makes the decision in practice. Route notices covered by section 12 to staff who can meet the applicable steps and deadlines. A lender extending the service to Barbados needs a separate local review. The Central Bank of Barbados's market conduct guidance addresses complaints and access for vulnerable groups among the entities it regulates [14].

An emergency office translates a public warning

An official bulletin names two neighbouring districts and gives an expiry time. The Common Alerting Protocol provides a standard format for alerts, and the World Meteorological Organization maintains a register of recognised alerting authorities [15]. The local issuing authority decides which bulletin is current and who may publish it.

Limit the tool to drafting. A bilingual reviewer compares the district, instruction, time and cancellation status with the original. A speaker of the target language variety then explains, in their own words, where the warning applies, what to do and when it ends. If the reviewers disagree or a reviewer cannot be reached, use the existing authorised warning procedure. Continue issuing the official notice through established channels. Include a correction in the rehearsal so staff can tell whether people who received an earlier translation also receive the update.

A school helps parents complete admissions

A parent wants to supply a missing document. An assistant can quote the published instructions and direct an exception to a member of staff. It can do that without access to private pupil files. Show the date of the instructions and give parents a way to contact the school directly.

Invite families to complete the task on a phone with a slow connection, using a keyboard or using a screen reader. WCAG 2.2 gives testable criteria for the web interface [16]. Arrange assistance for children or caregivers and collect only the information needed for this evaluation. Ask each participant to submit a test document and find the receipt. A correct answer in the chat cannot recover a submission lost on the next screen. Keep the staffed route open while the school fixes the form.

VALIDATION

Implementation, field validation and limitations

A small office can begin with one reversible task, a service owner and a second person authorised to challenge the evidence. A larger institution should involve operations, legal or compliance staff, the person responsible for data protection and people who use the service. Where a supplier handles part of a case, name the person on each side who can correct it. Ask the duty team to contact both outside office hours.

How to test the proposed method

Start in shadow mode. The proposed system handles test cases alongside the existing process without changing the live outcome. Define eligibility, the follow-up period and the maximum tolerable increase in harm before the pilot. Use the same service standard for the existing and proposed workflows.

For a study using random allocation, document the applicable ethics or institutional oversight decision, legal authority and requirements for participant notice or consent. Exclude cases where assigning a route or delaying action could compromise rights or safety. If both routes are acceptable and enough cases are available, assign comparable cases or shifts at random. Where that is impractical, compare matched cases or shifts and record differences in staffing, season, demand and case difficulty. Account for dependence between observations where the same staff or cases contribute repeatedly.

Choose one primary outcome. For example, count the proportion of eligible cases in which a consequential error remains uncorrected at the end of follow-up. Report unresolved cases and incomplete follow-up separately. Where feasible, conceal from the reviewer which workflow produced each decision. Ask a second reviewer to examine disputed cases and report disagreements alongside the final classifications.

Keep complaints and correction available to both groups. Count the time the person spends on the process, including return visits. Retain cases in their original group if staff take over, and record what happened to them. If the apparent improvement disappears when these cases are counted, hold expansion and investigate why staff intervened.

Limits of the evidence

Jurisdiction pages give a dated account of selected instruments. Before authorisation, obtain a local legal interpretation where a material issue of territorial reach, commencement, amendment or sector regulation remains open. A study across several agencies would also need permission at each site, common outcomes and safeguards against identifying participants. A favourable result at one site gives a second institution a reason to test the method in its own setting.

Serious events can be too infrequent for a field study to estimate their probability with useful precision. In that situation, retain enforceable permission limits and rehearse the fallback. Name the person who accepts the remaining risk and the evidence that would require a fresh decision.

FIELD WORKSHEET

Conclusion and release record

An institution can approve a service when the evidence covers the people, decisions and conditions within that approval. Narrow the release when the evidence covers only part of the proposed use. Hold it when an important legal question or safety test remains open. Stop an operating function when its agreed limit is breached, and restore a reliable route for the people who depend on it.

Record the decision

Record	What the next officer needs to find
Use and limits	Service, jurisdiction, languages, population, system version, permitted actions and actions reserved for people.
Legal basis	Applicable provision and source, commencement or scope question, sector regulator and officer responsible for resolving uncertainty.
Benefit and harm	Existing service, intended outcome, people exposed to a credible failure and how they can obtain a correction.
Evaluation	Case selection, comparison service, counts and denominators, access or language gaps, serious failures and unresolved cases.
Decision	Approved use or reason to narrow, hold or stop. Record dissent, owner, deputy, expiry and conditions that require another test.
Recovery	How to disable the function, who covers unfinished cases, notice duties, remedy and evidence required to restart.

When a failure occurs, restrict access to the evidence needed to establish what happened. Record the version, permissions and affected period under the applicable retention rules. Give each person a case reference and an officer authorised to provide a remedy. Confirm that downstream recipients acted on the correction. In an accessible form, tell the person what changed and how to request further review. Check whether they can now obtain the service, including if their case had already been closed.

Legal and operational leads should establish any notification duties and deadlines in each affected jurisdiction. Before restarting, repeat the failed case and related cases, exercise the fallback and obtain fresh approval.

Cite and adapt this paper

Dunkley, A. (2026). *When is an AI system safe enough to use? Evidence for Caribbean deployment decisions.* Caribbean AI Association, Research & Policy Series, November 2026 edition. Canonical publication.

Credit Adrian Dunkley as lead author and the Caribbean AI Association as publisher. For an adaptation, link the original edition and identify the adapter, changes and date. Retain third-party attribution and obtain any permissions the reuse requires. The adapter takes responsibility for its interpretation. The regional guidance and jurisdiction directory identify local sources for the release record.

SOURCES

References

01 Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. and Mané, D. (2016). Concrete Problems in AI Safety. Research preprint.

02 International AI Safety Report (2026). International AI Safety Report 2026, executive summary.

03 UNESCO (2024). Caribbean Artificial Intelligence Policy Roadmap.

04 UNESCO (2026). Latin America and the Caribbean adopt a regional roadmap for 2026–2027. Report of the Santo Domingo summit, 2 July.

05 Dunkley, A. (2026). A model AI Act for CARICOM. Caribbean AI Association consultation draft v0.1, 21 October.

06 Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1.

07 Autio, C. and colleagues (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1.

08 UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence.

09 Armstrong, R.-A., Hewitt, J. and Manning, C. (2022). JamPatoisNLI: A Jamaican Patois Natural Language Inference Dataset. Findings of EMNLP.

10 Lent, H. and colleagues (2024). CreoleVal: Multilingual Multitask Benchmarks for Creoles. Transactions of the Association for Computational Linguistics.

11 Hanley, J. A. and Lippman-Hand, A. (1983). If nothing goes wrong, is everything all right? Interpreting zero numerators. JAMA.

12 Jamaica (2020). The Data Protection Act, 2020, section 12.

13 Office of the Information Commissioner, Jamaica (2025). Responding to access requests and complaints from data subjects.

14 Central Bank of Barbados (2024). Market conduct guideline for regulated financial institutions.

15 World Meteorological Organization. Common Alerting Protocol.

16 World Wide Web Consortium (2024). Web Content Accessibility Guidelines 2.2.

17 Attorney General's Chambers, Saint Lucia. Data Protection Act, commencement and arrangement of sections. Revised Laws of Saint Lucia, 2023.