

ORIGINAL RESEARCH

Open access

Trust Must Survive Deployment—A Continuous Assurance Lifecycle for Drift Detection, Explanation Monitoring, Incident Response, Model Restriction, and Retirement in Adaptive AI Systems

James Whitfield¹, Aisha Rahman^{2*}, Kenji Sato³

Abstract

Trustworthy artificial intelligence cannot be established permanently before deployment because deployed and adaptive systems operate within environments that continue to change. Data distributions, relationships between inputs and outcomes, user behaviour, organizational workflows, system dependencies, and decision consequences may diverge from the conditions represented during development. This article proposes a continuous trustworthiness lifecycle for organizing the evidence and governance activities needed after deployment. The approach begins with a versioned deployment context and risk baseline, followed by monitoring of data, concept, behavioural, performance, calibration, uncertainty, and explanation changes. It integrates technical monitoring with human interpretation, user and affected-person feedback, incident identification, corrective action, accountability review, model restriction, temporary suspension, retirement, and assurance updating. The framework distinguishes drift signals from confirmed degradation and treats explanation stability as a separate assurance concern rather than assuming that stable predictive performance preserves explanation quality. Incidents and complaints are positioned as lifecycle evidence that may trigger investigation, remediation, updating, rollback, restriction, suspension, or retirement review. These responses remain human-governed because monitoring indicators cannot independently determine safety, ethical acceptability, legal compliance, or appropriate end-of-life decisions. The main contribution is a conceptual lifecycle that connects post-deployment monitoring with decision records, responsibility allocation, residual-risk communication, and organizational learning. It supports structured reasoning about whether an adaptive AI system should continue operating unchanged, undergo correction, operate within a narrower scope, be suspended, or enter retirement review. The framework requires domain-specific evaluation and institutional governance and should not be interpreted as an implemented assurance process, certification model, or deployment-readiness determination.

Keywords Trustworthy AI, Continuous assurance, Adaptive AI, Drift detection, Explanation monitoring, Incident response

*Correspondence:

Aisha Rahman
aisha.rahman@lums.edu.pk

¹ Department of Trustworthy Systems Engineering, Warwick Business School, Coventry, United Kingdom

² Department of AI Governance and Policy, Lahore University of Management Sciences, Lahore, Pakistan

³ Department of Intelligent Systems and Informatics, Keio University, Tokyo, Japan

Introduction

Trustworthy AI is often evaluated through development-time evidence concerning predictive performance, robustness, fairness, transparency, explainability, and human oversight. Such evidence is indispensable, but it is temporally bounded:

it describes a particular model version, dataset, task definition, technical environment, and intended-use context. Machine-learning assurance must therefore extend across operation, system change, maintenance, and eventual withdrawal rather than conclude when a model passes pre-deployment evaluation [1].

The persistence of trustworthiness is not only a technical monitoring problem. High-level ethical principles cannot independently determine how operational responsibilities should be allocated, how emerging failures should be investigated, or when continued use becomes unjustified [2]. Post-deployment assurance consequently requires institutional mechanisms that connect observed evidence with accountable decisions, documented interventions, stakeholder communication, and review of unresolved risk.

Trustworthy-AI scholarship identifies recurring principles such as transparency, accountability, fairness, non-maleficence, privacy, and human control, while also showing that trustworthiness is a multidimensional property involving technical performance, explainability, robustness, governance, and social context [3, 4]. These dimensions can change independently. A model may preserve aggregate accuracy while becoming poorly calibrated for a changing population, produce unstable explanations, generate unequal errors, or influence users in ways that were not represented during initial evaluation.

This article develops a conceptual continuous trustworthiness lifecycle for adaptive and deployed AI. The lifecycle connects deployment-risk baselines, drift detection, performance and explanation monitoring, uncertainty assessment, feedback, incident response, correction, accountability, restriction, suspension, retirement, and assurance updating. It does not treat continuous monitoring as certification. Instead, it organizes lifecycle evidence that may support qualified and human-governed decisions while maintaining explicit boundaries against claims of permanent trustworthiness, automatic safety determination, legal sufficiency, ethical completeness, or deployment readiness.

Adaptive AI and the Limits of Pre-Deployment Evaluation

Adaptive AI includes systems whose effective behaviour may change through retraining, parameter updates, changing data pipelines, altered software dependencies, feedback loops, or evolving patterns of use. Even systems whose model parameters remain fixed can become functionally different when workflows, populations, interfaces, or decision contexts change. Real-world implementation therefore introduces generalizability, integration, safety, and maintenance problems that cannot be resolved by development performance alone [5].

Changing environments weaken the assumption that development data adequately represent future operation. Distribution shift may arise from new populations, collection practices, devices, institutions, incentives, or external conditions. Causal relationships can also change, meaning that apparently similar inputs no longer support the same conclusions. Deployment-time stability must therefore be investigated rather than inferred from historical validation [6].

Post-deployment evidence is required because performance may deteriorate gradually, episodically, or only within particular contexts. Continual quality monitoring can identify emerging discrepancies, while longitudinal evidence demonstrates that properties such as calibration may change after initial development [7, 8]. However, deterioration cannot always be attributed to a single source: data quality, prevalence, user behaviour, workflow change, software modification, feedback effects, or model updating may interact.

Pre-deployment evaluation is therefore necessary but not permanently sufficient. A continuous assurance approach should maintain versioned information about intended purpose, supported populations, decision context, known limitations, technical dependencies, monitoring coverage, and responsible actors. Material changes should trigger renewed review of whether existing evidence remains applicable. This framework supports that review process, but it does not determine whether a system is ready for deployment or whether its continued operation is legally, ethically, or operationally acceptable.

Concept drift concerns change in the relationship between observed inputs and the outcome or decision target. Such change can make previously learned patterns unreliable even when the incoming features appear familiar. Drift may be abrupt, gradual, incremental, recurring, or difficult to distinguish from noise, making continuing detection and adaptation important but inherently uncertain [9].

Recurring concept drift further complicates lifecycle assurance because earlier conditions can return. A model may encounter seasonal, institutional, or behavioural patterns that resemble previously observed states without being identical to them. Lifecycle learning should consequently retain information about earlier contexts, interventions, and model versions rather than assume that operational evolution follows a one-directional path [11].

Behavioral drift extends the analysis beyond statistical distributions. Users may over-rely on outputs, develop workarounds, alter documentation, contest recommendations, or change their behaviour in response to the system. Feedback loops may then affect future observations, labels, opportunities, and outcomes. Context shift, environmental change, data-quality drift, model drift, population shift, calibration drift, and user-behaviour change should therefore be treated as related but distinct assurance signals. **Table 1** summarises post-deployment drift and shift risks that can undermine trustworthiness after pre-deployment evaluation has been completed.

Table 1. Post-Deployment Drift and Shift Risks in Adaptive AI Systems: Data Drift, Concept Drift, Distribution Shift, Behavioral Drift, Model Drift, Data Quality Drift, Input Shift, Output Shift, Prediction Shift, Calibration Drift, Population Shift, Context Shift, Environmental Change, User Behavior Change, and Pre-Deployment Sufficiency Boundaries

Drift or shift type	Core monitoring question	How the change appears	Potential trustworthiness risk	Detection or evidence need	Failure or misinterpretation risk	Governance response need
Data drift	Have operational feature distributions changed?	Changed ranges, frequencies, correlations, or missingness	Reduced transferability or uneven errors	Baseline comparison plus contextual evidence	Benign change may be treated as harmful	Investigate significance and affect scope
Concept drift	Has the input–outcome relationship changed?	Changed errors or conditional relationships	Learned decision rules become unreliable	Recent labels, calibration, and domain review	Noise or delayed labels may be misclassified	Escalate model to domain assessors

Distribution shift	Has the joint data-generating distribution changed?	Combined input, outcome, or relationship change	Development evidence may no longer transfer	Distribution and task-performance evidence	One distance measure may be treated as conclusive	Transfer review
Behavioral drift	Have model, user, or workflow behaviours changed?	Changed reliance, overrides, recommendations, or workarounds	New sociotechnical failure pathways	Interaction logs and qualitative evidence	Change may be attributed to the wrong component	Human factors investigation
Model drift	Has effective model behaviour changed?	Changed versions, preprocessing, outputs, or dependencies	Deployed system differs from evaluated system	Versioned model and pipeline evidence	Upstream causes may be overlooked	Change control, renewal, reevaluation
Data quality drift	Have completeness, accuracy, timeliness, or provenance changed?	Missing, delayed, corrupted, or inconsistent inputs	Hidden degradation or inequitable errors	Quality and provenance checks	Stable distributions may conceal low quality	Data governance, intervention
Input shift	Are new values, categories, devices, or combinations appearing?	Movement beyond development coverage	Unsupported extrapolation	Feature-level and multivariate comparison	Interactions may be missed	Intended review
Output shift	Has the distribution of model outputs changed?	Changed score, class, ranking, or abstention rates	Altered allocation or intervention patterns	Output trends linked to context and outcomes	Legitimate prevalence change may be mistaken for defect	Decision impact review
Prediction shift	Are predictions changing without an established justification?	Score movement or reordered classifications	Changed treatment of cases or groups	Cohort and version comparisons	Output change may reflect valid context change	Context investigation
Calibration drift	Do predicted probabilities match observed frequencies?	Systematic overprediction or underprediction	Misleading confidence and threshold use	Recent outcome-linked calibration evidence	Discrimination metrics may conceal the problem	Recalibration or broader review
Label shift	Has outcome prevalence changed under the required assumptions?	Changed class proportions	Biased probabilities or decisions	Outcome prevalence and assumption checks	Shift may be claimed without verifying assumptions	Domain performance review

Covariate shift	Have inputs changed while the outcome mechanism is assumed stable?	New feature frequencies or regions	Poor transfer to operational cases	Input evidence and stability-assumption review	The assumption may be unverifiable	Transfer assessment
Population shift	Has the composition of affected populations changed?	New demographic, geographic, clinical, or behavioural groups	Emerging disparity or exclusion	Representation and subgroup evidence	Population change may be reduced to one attribute	Fairness scope re
Context shift	Have purpose, workflow, stakes, or institutional conditions changed?	Outputs acquire different meanings or consequences	Prior evidence becomes contextually invalid	Workflow and impact documentation	Identical data may be assumed to imply identical use	Governance and scope review
Environmental change	Have external operating conditions changed?	New sites, devices, practices, threats, or constraints	Invalid assumptions or unexpected failure modes	Dependency and environmental evidence	External causes may be omitted	System-investigation
User behavior change	Have operators or affected users changed how they engage with the system?	Automation bias, avoidance, gaming, or workarounds	Overreliance, exclusion, or feedback distortion	Usage and qualitative evidence	Behavioural records may be treated as self-explanatory	Human-centered review
Pre-deployment sufficiency boundary	Can initial evaluation permanently establish trustworthiness?	Evidence becomes stale as operation changes	False confidence in historical validation	Continuing lifecycle evidence	Approval may be treated as permanent	Schedule and event trigger reassessment
Drift-detection guarantee boundary	Does an alert or absence of an alert determine trustworthiness?	Detector warning or silence	False reassurance or unnecessary action	Corroborating technical and human evidence	Imperfect signals become automated verdicts	Human governance interpretation

Explanation Stability and Performance Monitoring

Performance monitoring should examine more than whether aggregate accuracy remains close to an initial benchmark. Continuing assurance may require calibration, uncertainty, error severity, subgroup performance, decision utility, abstention behaviour, and the consequences of outputs in current workflows. Explanation monitoring adds a distinct question: whether the system continues to provide explanations that are faithful, sufficiently stable, contextually appropriate, and useful for the stakeholders expected to interpret them [13].

Explanation stability concerns whether similar cases or justified perturbations produce reasonably consistent explanatory accounts. Explanation consistency compares repeated runs or methods, while fidelity concerns whether an explanation accurately represents the model behaviour it claims to describe. Quantitative evaluation can examine fidelity, robustness, complexity, sensitivity, and related properties, but evaluation choices must remain aligned with the explanation method and intended purpose [14].

Explanation instability may arise from model updates, changes in data relationships, explanation-method sensitivity, or operation near decision boundaries. Feature-attribution rankings may shift, counterfactual suggestions may become inconsistent, and explanations may remain visually plausible despite weak fidelity. Current evaluation practices are fragmented, so no single stability, robustness, or uncertainty metric can establish that explanations remain adequate [15].

Human interpretation is equally important because explanation usefulness depends on the stakeholder, task, decision context, and reason for requesting an explanation [16]. Monitoring should therefore integrate performance signals, drift signals, explanation stability, uncertainty, user complaints, affected-person feedback, incidents, and evidence about how explanations influence reliance and contestation. **Figure 1** illustrates the continuous monitoring logic for adaptive AI, showing how drift, performance change, explanation instability, uncertainty, and human feedback can trigger review, correction, restriction, or escalation.

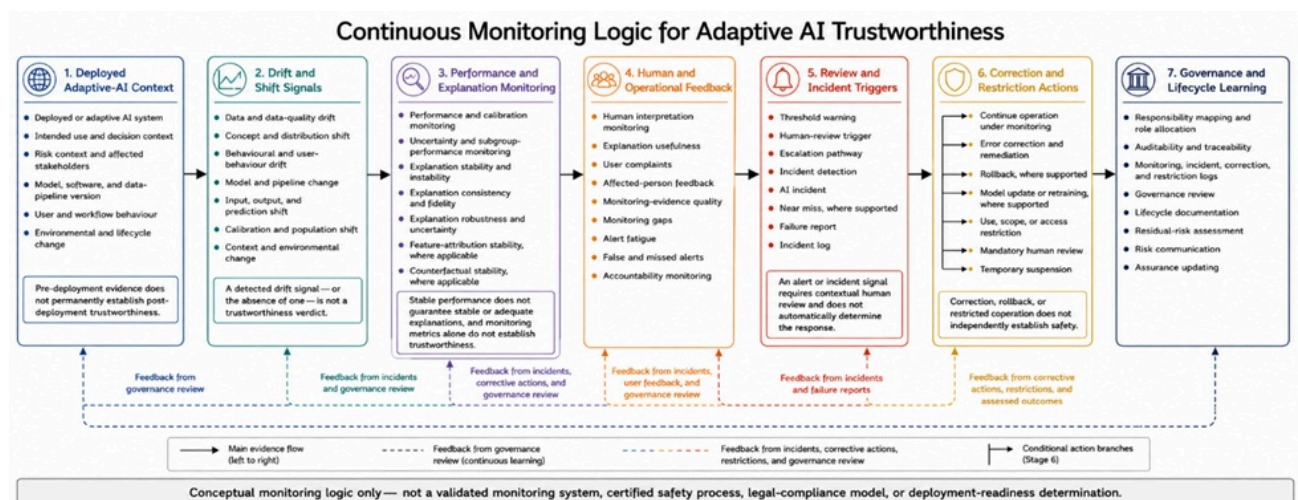


Figure 1. Continuous Monitoring Logic for Adaptive AI Trustworthiness

Alt-text description:

A conceptual monitoring pathway showing drift signals, performance changes, explanation instability, uncertainty changes, human feedback, incident signals, review triggers, correction, restriction, and governance review.

The monitoring process must itself be evaluated. Missing or delayed outcomes, incomplete subgroup information, unreported incidents, weak provenance, and inaccessible feedback channels can produce monitoring gaps. Excessive alerts may reduce reviewer attention, whereas false alerts and missed alerts can distort intervention. Continuous assurance must therefore assess the coverage, timeliness, representativeness, traceability, and decision relevance of monitoring evidence. Even well-designed monitoring cannot independently establish operational safety, accountability, legal compliance, ethical sufficiency, or deployment readiness.

Incident Response, Correction, and Accountability

Post-deployment assurance requires a clear distinction between an observed anomaly, a confirmed failure, an AI incident, and a near miss. An incident may involve technical malfunction, harmful output, inappropriate reliance, workflow disruption, unfair impact, or a failure of organizational controls. Responsibility should remain traceable across developers, deployers, operators, reviewers, and institutional decision-makers rather than being assigned exclusively to either the model or its immediate user [17].

Accountability begins before an incident occurs. System design, data selection, interface choices, deployment scope, monitoring coverage, and escalation arrangements all embody decisions for which organizational actors remain responsible [18]. Responsibility mapping should therefore identify who owns monitoring, investigation, communication, correction, restriction, suspension, and closure. Role allocation should also clarify decision authority so that shared responsibility does not become an accountability gap.

Transparency and explanation may support investigation, but they cannot restore accountability by themselves [19]. Useful incident evidence may include the relevant model and software versions, input provenance, outputs, explanations, uncertainty information, user actions, overrides, affected-person feedback, and decision records. Traceability connects these elements, while auditability enables structured review of whether the evidence, process, and response were adequate.

Incident pathways should accept signals from monitoring systems, failure reports, user complaints, affected-person feedback, audits, and human review. Translating ethical commitments into operational practice requires mechanisms for triage, escalation, remediation, communication, and documented follow-up [20]. Accessible complaint and contestation channels are particularly important because harms may be visible to affected persons before they appear in aggregate performance indicators.

Correction may include data repair, workflow modification, explanation redesign, rollback, model updating, or retraining where the evidence supports those actions. Every intervention should be recorded in a correction log and followed by reassessment because resolving an observed failure does not demonstrate that systemic causes or residual risks have disappeared. Auditing can support evidence review and institutional accountability, but its scope, independence, methods, and limitations must remain explicit [21]. Incident closure therefore represents a qualified governance decision rather than proof that the system has become safe or ethically sufficient.

Model Restriction, Suspension, and Retirement

When evidence no longer supports unchanged operation, a proportionate response may involve model restriction rather than immediate correction or complete withdrawal. Restriction can narrow permitted decisions, populations, locations, access rights, autonomy, or operating conditions. Responsible deployment requires safeguards that remain responsive to changing evidence, including mechanisms for reconsidering use when performance or contextual validity deteriorates [22].

Use restriction identifies tasks for which the model should no longer be applied, while scope restriction limits operation to supported populations or contexts. Access restriction can confine use to authorized personnel, and mandatory human review can apply to high-uncertainty or high-consequence outputs. Continued effectiveness should be demonstrated within the remaining scope rather than inferred from historical development results [23].

Temporary suspension is a stronger intervention that pauses all or substantial use while investigation, correction, or replacement review proceeds. A degraded mode may sometimes reduce model autonomy or functionality, but it should not be presumed safer without evaluation of the alternative workflow and its burdens. Operational lifecycle research indicates that conventional AI lifecycle models often underrepresent maintenance, organizational feedback, changing responsibilities, and end-of-life decisions [24].

Technical practices such as version control, observability, controlled release, rollback, and traceable dependencies can support restriction and suspension processes [25]. These mechanisms are enabling infrastructure rather than evidence of trustworthiness. A rollback may restore an earlier model version, for example, but that version may also be unsuitable for the current population or context. Post-restriction monitoring must consequently test whether the control is followed, whether it addresses the identified concern, and whether risk has shifted elsewhere.

Model retirement becomes relevant when the original purpose is obsolete, essential dependencies are unsupported, persistent failures cannot be adequately controlled, or continued operation is unjustified relative to replacement or non-use. Corrective updating can improve a drifting model in some settings, but an update creates a new candidate system requiring renewed evaluation [26]. Retirement and decommissioning therefore require human-governed review of alternatives, residual risk, affected parties, retained records, hidden integrations, and post-retirement consequences.

Table 2 maps explanation stability monitoring, performance monitoring, incident response, correction, accountability, restriction, suspension, and retirement requirements for continuous AI assurance.

Table 2. Continuous Assurance Requirements for Explanation Stability, Performance Monitoring, Incident Response, Correction, Accountability, Model Restriction, Suspension, and Retirement: Monitoring Evidence, Drift Warnings, Explanation Drift, Human Review Triggers, Incident Logs, Correction Logs, Restriction Logs, Retirement Logs, Residual Risk, and Lifecycle Boundaries

Assurance requirement	Core lifecycle question	Required evidence or trigger	Assurance or risk-control function	Potential output	Failure or residual risk	Review, documentation or governance need
Performance monitoring	Does the model remain effective for current tasks, contexts, and groups?	Calibration, error, utility, subgroup, abstention, and outcome evidence	Identify meaningful degradation	Continue, investigate, correct, or restrict	Aggregate measures may conceal consequential failure	Versioned monitoring record and authorized review
Explanation stability	Are explanations appropriately stable, faithful, robust, and useful?	Longitudinal explanations, perturbation tests, fidelity checks, and stakeholder evidence	Detect explanation change or inadequacy	Investigation, redesign, warning, or withdrawal	Stable explanations may remain misleading	Explanation log and method-specific review
Monitoring evidence quality	Is evidence timely, representative, traceable, and decision-relevant?	Provenance, coverage, latency, completeness, and validation information	Qualify assurance inputs	Accepted evidence or documented evidence gap	Poor evidence may produce confident but invalid conclusions	Monitoring-quality review
Drift warning	Does a detected change warrant further investigation?	Data, concept, behavioural, model, calibration, or	Initiate contextual assessment	Observe, investigate, or escalate	False alerts and missed alerts remain possible	Detector limitations and rationale log

		explanation signal				
Human review trigger	Does uncertainty, harm, contestation, or system change require human judgment?	Severe alert, incident, complaint, uncertainty, or unsupported context	Prevent automated resolution of contested risk	Reviewed continuation or escalation	Review may be superficial or under-resourced	Competence, authority, and workload controls
AI incident	Has AI operation caused or contributed to material harm or unacceptable risk?	Monitoring, outcome, complaint, audit, or operational evidence	Establish severity, scope, and contributing factors	Incident response and control decision	Narrow definitions may exclude sociotechnical harm	Incident log and multidisciplinary investigation
Near miss	Was material harm narrowly avoided?	Override, abstention, manual interception, or report	Support preventive learning	Preventive correction or monitoring revision	Near misses may be normalized or unreported	Non-punitive reporting and trend review
User or affected-person feedback	What operational or impact problem has been experienced?	Complaint, appeal, contestation, survey, or interview	Reveal issues absent from telemetry	Investigation, support, correction, or scope review	Inaccessible channels can exclude highly affected groups	Traceable feedback and closure process
Incident response	What immediate and longer-term action is proportionate?	Credible or confirmed incident	Contain risk and preserve evidence	Correction, restriction, suspension, or review	Premature closure can leave systemic causes	Escalation record and decision log
Error correction	Can a verified error be addressed without introducing unsupported risk?	Root-cause analysis and correction options	Address specific failure	Data, model, explanation, or workflow correction	Local correction may not address structural causes	Correction log and follow-up review
Rollback	Is return to an earlier controlled state justified?	Failed release, harmful update, or severe degradation	Provide reversible containment	Restored prior version	Earlier version may also be unsuitable	Version provenance and authorization

Model update or retraining	Is adaptive modification justified by evidence?	Drift, performance, incident, data, and alternative-response evidence	Develop a revised candidate system	Update, retraining, rejection, or non-model response	New bias, instability, or evidence invalidation	Change control and renewed evaluation
Responsibility mapping	Who owns monitoring, response, communication, and approval?	System and organizational role analysis	Prevent accountability diffusion	Named responsibilities and decision rights	Shared roles may become unowned obligations	Periodic role review
Auditability and traceability	Can evidence and decisions be independently reconstructed?	Logs, provenance, versions, tests, and role records	Enable review and learning	Audit findings and corrective recommendations	Records may be incomplete or ceremonial	Evidence-integrity and independence controls
Model restriction	Should permitted use be narrowed?	Context-specific degradation, incident, uncertainty, or evidence gap	Reduce exposure proportionately	Use, scope, access, or autonomy restriction	Remaining use may retain significant risk	Restriction log and post-restriction monitoring
Temporary suspension	Should operation pause while evidence is developed?	Severe or unresolved concern	Limit exposure during review	Suspension with reinstatement conditions	Suspension may shift risk to alternatives	Authorized decision and contingency review
Model retirement	Is continued operation no longer justified?	Persistent risk, obsolete purpose, invalidated evidence, or unsupported dependencies	End reliance on the system	Retirement and decommissioning	Hidden integrations and legacy effects may remain	Retirement log and post-retirement review
Residual-risk communication	What unresolved uncertainty must be communicated, and to whom?	Integrated assurance and governance evidence	Support informed decision-making	Audience-specific risk statement	Generic warnings may obscure responsibility	Communication-quality and comprehension review
Restriction-safety boundary	Does restricted use establish safety?	Evidence concerning the remaining scope	Prevent overclaiming	Conditional restricted operation	Restricted contexts may remain	Continuing assurance

					unsafe or unsupported	
Retirement-automation boundary	Can monitoring automatically determine retirement?	Alerts, incidents, alternatives, and impact evidence	Preserve contextual human judgment	Reasoned governance decision	Automated thresholds may ignore rights and dependencies	Multidisciplinary authorization

Proposed Continuous Trustworthiness Lifecycle

The proposed lifecycle begins with deployment-context and risk-baseline definition. This stage records the system version, intended purpose, supported population, operational environment, decision stakes, stakeholder roles, technical dependencies, known limitations, and available evidence. An end-to-end AI lifecycle should connect conception and development with production, monitoring, feedback, maintenance, and subsequent change [27]. The baseline is therefore a versioned reference point rather than a permanent declaration of trustworthiness.

The second stage develops a post-deployment monitoring plan covering data, concept, behavioural, model, performance, calibration, uncertainty, explanation, robustness, and fairness dimensions where relevant. Monitoring choices should reflect the harms, obligations, and affected interests associated with the system, not merely the signals that are easiest to measure. Ethical considerations such as non-maleficence, autonomy, justice, beneficence, and explicability provide broader interpretive obligations but require context-specific operationalization [28].

The third and fourth stages connect monitoring with human and organizational evidence. Drift warnings, performance changes, explanation instability, uncertainty, incidents, complaints, and affected-person feedback enter a review process that examines evidential quality, monitoring gaps, causal hypotheses, and affected contexts. Human-centered AI requires technical reliability to be combined with competent human oversight, organizational controls, and clearly allocated authority [29].

The fifth and sixth stages organize corrective and risk-control decisions. Correction, remediation, rollback, updating, and retraining are considered alongside continued monitoring, scope restriction, access restriction, mandatory human review, temporary suspension, replacement, or retirement. Fairness may change with populations, data, labels, and operational conditions, while post-market evidence can support renewed auditing and governance assessment [30, 31]. These signals inform but do not automate decisions because their interpretation depends on context, alternatives, residual risk, and the interests of affected parties.

The final stage updates lifecycle evidence through monitoring logs, incident logs, correction logs, restriction records, retirement documentation, governance decisions, and residual-risk communication. Every material intervention returns the system to a revised baseline and monitoring plan. Table 3 presents the proposed continuous trustworthiness lifecycle linking deployment context, monitoring, drift detection, explanation stability, incident response, correction, restriction, suspension, retirement, documentation, governance review, and assurance updating.

Table 3. Proposed Continuous Trustworthiness Lifecycle for Adaptive AI Systems: Deployment Context, Risk Baseline, Data Drift Monitoring, Concept Drift Monitoring, Behavioral Drift Monitoring, Performance Monitoring, Explanation Stability Monitoring, Incident Detection, Correction, Accountability, Model Restriction, Suspension, Retirement, Residual-Risk Communication, Documentation, Governance Review, and Assurance Updating

Lifecycle stage	Core purpose	Required input	Assurance function	Potential output	Validation or governance	Failure risk
-----------------	--------------	----------------	--------------------	------------------	--------------------------	--------------

					need	
Deployment context and risk baseline	Define the system, purpose, environment, stakeholders, and evidence scope	Intended use, version, population, dependencies, limitations, and risk record	Establish a reference state	Versioned baseline	Domain and institutional review	Vague baseline prevents meaningful comparison
Monitoring plan	Specify signals, coverage, frequency, responsibility, and evidence limitations	Baseline, risk pathways, available outcomes, and stakeholder needs	Align observation with material questions	Monitoring and coverage plan	Monitoring-quality review	Easily measured indicators may crowd out important harms
Data and quality monitoring	Detect material changes in inputs, collection, and provenance	Baseline and current feature and quality evidence	Generate qualified warnings	Continue or investigate	Context and outcome interpretation	Benign shift or missed harmful change
Concept and behavioural monitoring	Assess changing relationships and sociotechnical behaviour	Labels, errors, interactions, reliance, and workflow evidence	Identify changed system conditions	Model or human-factors review	Domain and mixed-method assessment	Noise or behavioural ambiguity
Performance, calibration, and uncertainty monitoring	Determine whether outputs remain supported and useful	Outcomes, errors, utility, subgroup, calibration, and uncertainty evidence	Detect degradation and unsupported confidence	Continue, investigate, correct, or restrict	Domain-specific criteria	Aggregate metrics or weak labels mislead
Explanation stability monitoring	Evaluate fidelity, robustness, consistency, uncertainty, and usefulness	Versioned explanations, perturbations, predictions, and stakeholder evidence	Detect explanation-related change	Investigation, redesign, or warning	Method- and stakeholder-specific validation	Stable but misleading explanations
Incident and feedback review	Capture failures, near misses, complaints, and	Monitoring, reports, appeals, audits, and affected-	Activate response and reveal blind spots	Incident record and escalation	Accessible reporting and multidisciplinary triage	Underreporting or narrow classification

	experienced impact	person evidence					
Correction and remediation	Address verified technical and organizational causes	Root-cause analysis, alternatives, and affected-party evidence	Reduce recurrence or impact	Correction, rollback, workflow change, or update	Post-correction evaluation	Local fixes leave systemic causes	
Accountability review	Allocate responsibility and preserve answerability	Roles, decisions, logs, evidence, and challenges	Ensure ownership and reviewability	Responsibility and decision record	Authority and independence review	Documentation without consequences	
Restriction pathway	Narrow use while uncertainty or context-specific failure is addressed	Risk evidence, affected scope, alternatives, and residual risk	Reduce exposure	Use, scope, access, or autonomy restriction	Authorized and enforceable decision	Risk persists or moves elsewhere	
Suspension pathway	Pause operation during severe or unresolved concern	Severity, uncertainty, alternatives, and dependency evidence	Provide temporary containment	Suspension with review conditions	Contingency and reinstatement review	Alternative processes create new risk	
Retirement pathway	End use when continued operation is unjustified	Persistent risk, obsolescence, alternatives, dependencies, and impacts	Govern withdrawal	Retirement and decommissioning	Multidisciplinary authorization	Legacy outputs or hidden integrations remain	
Residual-risk communication	Communicate remaining uncertainty and obligations	Integrated lifecycle evidence	Support informed governance and use	Audience-specific statement	Comprehension and actionability review	Generic disclosure obscures responsibility	
Documentation and auditability	Preserve reconstructable evidence across versions and decisions	Monitoring, incident, correction, restriction, and retirement records	Enable independent examination and learning	Lifecycle assurance record	Integrity, retention, and access controls	Incomplete or ceremonial documentation	
Governance review and assurance updating	Reassess whether current evidence	Integrated technical, human, impact, and	Produce a qualified lifecycle decision	Continue, correct, restrict, suspend, retire, or gather evidence	Legitimate multidisciplinary authority	Routine approval without meaningful challenge	

	supports the operating state	contextual evidence				
--	---------------------------------	------------------------	--	--	--	--

Operational and Research Implications

Continuous assurance research should investigate how monitoring evidence remains linked to the precise system version, deployment context, population, and decision process to which it applies. Data-drift research should move beyond detector comparisons toward evaluation of whether warnings predict meaningful changes in performance, calibration, fairness, or impact. Concept-drift studies should address delayed labels, recurring contexts, causal ambiguity, and the possibility that adaptation itself creates new instability [9].

Behavioral-drift research requires combined technical, organizational, and human-centered methods. Interaction logs alone cannot explain changes in reliance, workarounds, contestation, or avoidance. Explanation-stability research similarly needs longitudinal evaluation of fidelity, robustness, uncertainty, counterfactual feasibility, attribution stability, and stakeholder usefulness in operational contexts rather than isolated demonstrations [14].

Incident-response and accountability research should examine reporting accessibility, near-miss recognition, escalation quality, response timeliness, role allocation, correction effectiveness, and recurrence. Greater attention is also needed to affected-person feedback because conventional monitoring may overlook burdens experienced outside the deploying organization. Auditing should be evaluated as one component of lifecycle governance rather than treated as an independent guarantee of trustworthy conduct [21].

Restriction, suspension, and retirement remain comparatively underdeveloped research domains. Studies should examine proportionality, enforcement, displaced risk, reinstatement, replacement, hidden dependencies, decommissioning, and post-retirement effects. These are conceptual and research implications, not legal advice, compliance guidance, procurement advice, audit instruction, MLOps instruction, tool-selection advice, threshold-selection advice, platform-selection advice, suspension or retirement advice, safety certification, or deployment instruction. Operational adoption requires separate domain-specific evaluation and legitimate institutional authority [24].

Conclusion

This article has proposed a continuous trustworthiness lifecycle for adaptive AI systems that links versioned deployment-risk baselines with data, concept, and behavioral drift detection; performance, calibration, uncertainty, and explanation monitoring; user and affected-person feedback; incident response; correction; accountability; model restriction; suspension; retirement; governance review; and assurance updating. Its central claim is that trustworthiness must survive deployment through continuing evidence and accountable lifecycle decisions rather than being assumed from pre-deployment evaluation, benchmark performance, initial explanations, or static governance commitments. Monitoring signals remain incomplete and fallible, corrective actions may introduce new risks, restrictions do not certify safety, and retirement decisions cannot be automated from thresholds alone. The framework therefore supports structured, cautious reasoning about post-deployment change while preserving the need for domain-specific validation, human-centered review, institutional governance, residual-risk management, and independent legal or safety assessment.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 02 Jul 2025

Revised: 24 Jul 2025

Accepted: 20 Sep 2025

Published online: 15 March 2026

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Ashmore R, Calinescu R, Paterson C. Assuring the machine learning lifecycle: desiderata, methods, and challenges. *ACM Comput Surv.* 2021;54(5):1-39.
<https://doi.org/10.1145/3453444>.
- Mittelstadt B. Principles alone cannot guarantee ethical AI. *Nat Mach Intell.* 2019;1(11):501-7.
<https://doi.org/10.1038/s42256-019-0114-4>.
- Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell.* 2019;1(9):389-99.
<https://doi.org/10.1038/s42256-019-0088-2>.
- Kaur D, Uslu S, Rittichier KJ, Durresi A. Trustworthy artificial intelligence: a review. *ACM Comput Surv.* 2022;55(2):1-38.
<https://doi.org/10.1145/3491209>.
- Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 2019;17(1):195.
<https://doi.org/10.1186/s12916-019-1426-2>.
- Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics.* 2020;21(2):345-52.
- Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. *NPJ Digit Med.* 2022;5(1):66.
<https://doi.org/10.1038/s41746-022-00611-y>.
- Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc.* 2017;24(6):1052-61.

Lu J, Liu A, Dong F, Gu F, Gama J, Zhang G. Learning under concept drift: a review. *IEEE Trans Knowl Data Eng*. 2019;31(12):2346-63.
<https://doi.org/10.1109/TKDE.2018.2876857>.

Bayram F, Ahmed BS, Kassler A. From concept drift to model degradation: an overview on performance-aware drift detectors. *Knowl Based Syst*. 2022;245:108632.
<https://doi.org/10.1016/j.knosys.2022.108632>.

Suárez-Cetrulo AL, Quintana D, Cervantes A. A survey on machine learning for recurring concept drifting data streams. *Expert Syst Appl*. 2023;213:118934.
<https://doi.org/10.1016/j.eswa.2022.118934>.

de Barros RSM, de Carvalho Santos SGT. An overview and comprehensive comparison of ensembles for concept drift. *Inf Fusion*. 2019;52:213-44.
<https://doi.org/10.1016/j.inffus.2019.03.006>.

Vilone G, Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf Fusion*. 2021;76:89-106.
<https://doi.org/10.1016/j.inffus.2021.05.009>.

Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, et al. From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. *ACM Comput Surv*. 2023;55(13s):1-42.
<https://doi.org/10.1145/3583558>.

Saeed W, Omlin CW. Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities. *Knowl Based Syst*. 2023;263:110273.
<https://doi.org/10.1016/j.knosys.2023.110273>.

Langer M, Oster D, Speith T, Hermanns H, Kästner L, Schmidt E, et al. What do we want from explainable artificial intelligence (XAI)?—a stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artif Intell*. 2021;296:103473.
<https://doi.org/10.1016/j.artint.2021.103473>.

Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. *Bull World Health Organ*. 2020;98(4):251-6.
<https://doi.org/10.2471/BLT.19.237487>.

Martin K. Ethical implications and accountability of algorithms. *J Bus Ethics*. 2019;160(4):835-50.
<https://doi.org/10.1007/s10551-018-3921-3>.

de Laat PB. Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? *Philos Technol*. 2018;31(4):525-41.

Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Sci Eng Ethics*. 2020;26(4):2141-68.
<https://doi.org/10.1007/s11948-019-00165-5>.

Mökander J, Morley J, Taddeo M, Floridi L. Ethics-based auditing of automated decision-making systems: nature, scope, and limitations. *Sci Eng Ethics*. 2021;27(4):44.
<https://doi.org/10.1007/s11948-021-00319-4>.

Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med*. 2019;25(9):1337-40.
<https://doi.org/10.1038/s41591-019-0548-6>.

Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ*. 2020;368.
<https://doi.org/10.1136/bmj.l6927>.

Haakman M, Cruz L, Huijgens HC, van Deursen A. AI lifecycle models need to be revised: an exploratory study in fintech. *Empir Softw Eng*. 2021;26(5):95.
<https://doi.org/10.1007/s10664-021-09993-1>.

Kreuzberger D, Kühl N, Hirschl S. Machine learning operations: overview, definition, and architecture. *IEEE Access*. 2023;11:31866-79.
<https://doi.org/10.1109/ACCESS.2023.3262138>.

Davis SE, Greevy RA Jr, Fonnesbeck C, Lasko TA, Walsh CG, Matheny ME. A nonparametric updating method to correct clinical prediction model drift. *J Am Med Inform Assoc*. 2019;26(12):1448-57.

De Silva D, Alahakoon D. An artificial intelligence life cycle: from conception to production. *Patterns (N Y)*. 2022;3(6):100489.
<https://doi.org/10.1016/j.patter.2022.100489>.

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach*. 2018;28(4):689-707.
<https://doi.org/10.1007/s11023-018-9482-5>.

Shneiderman B. Human-centered artificial intelligence: reliable, safe and trustworthy. *Int J Hum Comput Interact*. 2020;36(6):495-504.
<https://doi.org/10.1080/10447318.2020.1741118>.

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv*. 2021;54(6):1-35.
<https://doi.org/10.1145/3457607>.

Mökander J, Axente M, Casolari F, Floridi L. Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation. *Minds Mach*. 2022;32(2):241-68.
<https://doi.org/10.1007/s11023-021-09577-4>.