4iGov

AI GOVERNANCE PRACTITIONER SERIES

ARTEFACT 3

AI Agent Capability Mapping and Risk Assessment

Assessment Framework for Automated and Agentic AI Systems in Regulated Environments

A structured pass/fail assessment tool for evaluating AI and automated systems against governance readiness criteria before and after deployment.

Published by 4iGov | 4igov.cloud | 2026

1. PURPOSE AND HOW TO USE THIS ASSESSMENT

This assessment provides a structured framework for evaluating AI and automated systems against seven governance capability dimensions. It produces a pass, conditional pass, or fail rating for each dimension, a total score, and an overall deployment readiness verdict.

The framework applies to three system types: automated decision systems (rule-based, no learning), AI-assisted systems (human in the loop), and agentic AI systems (autonomous, multi-step decision authority). The Apple Card dispute routing system is assessed in Section 4 as a worked example of an automated system. Agentic AI extensions are mapped in Section 5.

Rating Definitions

- PASS — The control or capability is in place, documented, and evidenced. No material gap identified.
- CONDITIONAL PASS — The capability exists but has gaps that must be remediated within a defined timeframe. Deployment may proceed with documented remediation plan.
- FAIL — The capability is absent or materially deficient. Deployment must not proceed until the gap is remediated and re-assessed.

When to Use This Assessment

- Before first deployment of any AI or automated system in a regulated environment
- Before any material update to an existing system — new model version, workflow change, UI update affecting regulated processes
- As part of the annual governance review for all active AI use cases
- Following any incident or regulatory inquiry involving an AI system
- When onboarding a third party AI provider whose systems will process regulated obligations

2. CAPABILITY DIMENSIONS AND ASSESSMENT CRITERIA

The seven dimensions below define what governance-ready AI looks like across the full lifecycle. Each dimension is broken into assessment criteria with scoring guidance. Maximum score per dimension is 4. Overall maximum is 28.

	Dimension	What Is Being Assessed	Pass Threshold	Fail Condition
	Governance Accountability	A named owner exists for the AI system and for every integration boundary it operates across. Accountability is documented in a register, not implied by org structure.	Owner named in writing. Integration boundary accountability matrix complete.	No named owner. Integration boundaries unowned.
	Risk Classification and Regulatory Mapping	The system has been formally classified under EU AI Act. All regulatory obligations triggered by the classification are identified and documented.	Risk tier documented. Regulatory obligations list approved by compliance.	No risk classification. Regulatory obligations not mapped.
	Bias and Fairness Controls	For systems affecting consumer outcomes: bias assessment conducted, proxy variable analysis completed, disparate impact tested before deployment.	Bias assessment documented. Proxy variable analysis on record. Disparate impact results reviewed.	No bias assessment. No proxy variable analysis. System deployed without fairness testing.
	Transparency and Explainability	Decisions or outputs produced by the system can be explained to the affected consumer and to regulators in plain language. No black box operation for regulated decisions.	Decision explanation available at consumer and regulator level. Customer service can articulate basis for any automated decision.	Decisions cannot be explained. Black box operation. Customer service unable to respond to consumer challenges.
	Human Oversight and Intervention	Effective human oversight is possible. Override mechanisms exist. Escalation paths are defined and tested for edge cases, failure states, and regulatory threshold breaches.	Override mechanism documented and tested. Escalation path defined for all failure states. Regulatory breach escalation tested.	No override mechanism. No escalation path. System operates without human intervention capability.
	Monitoring and Incident Detection	Real-time or near-real-time monitoring is active on the system. SLA breach alerting is configured and tested. Processing failures are detected within defined timeframes tied to regulatory obligations.	Monitoring active. SLA breach alerts tested. Processing failures detected within 24 hours.	No monitoring. No alerting. Processing failures detectable only through consumer complaints or regulatory inquiry.
	Audit Trail and Evidencing	Every decision, action, and state transition produced by the system is logged in an auditable, tamper-evident record. Regulators can reconstruct the full processing history for any consumer interaction.	Complete audit trail. Logs retained per regulatory retention requirements. Reconstructable for any transaction on request.	No audit trail. Logs incomplete or absent. Processing history cannot be reconstructed.

3. SCORING GUIDE AND DEPLOYMENT VERDICTS

Score each dimension from 0 to 4 using the criteria below. Sum the scores for an overall rating. Apply the deployment verdict based on the total score and any individual FAIL ratings.

Score	Rating	Criteria
4	Full Pass	Control is in place, documented, independently evidenced, and has been tested or validated within the past 12 months.
3	Conditional Pass — Minor	Control is in place and documented but evidence is incomplete or testing has not been conducted within 12 months. Remediation required within 90 days.
2	Conditional Pass — Material	Control partially exists but has a material gap that could result in a regulatory breach under adverse conditions. Remediation required within 30 days. Deployment may proceed with risk acceptance.
1	Fail — Remediable	Control is absent but a remediation plan exists and is in progress. Deployment must not proceed until score reaches 3 or above.
0	Fail — Critical	Control is absent with no remediation plan. Deployment must not proceed. Escalation to governance committee required.

Total Score	Deployment Verdict	Conditions
25 to 28	APPROVED	All dimensions at 3 or above. No individual FAIL ratings. Proceed to deployment with standard monitoring.
18 to 24	CONDITIONAL APPROVAL	No dimension below 2. Remediation plan required for any dimension below 4. Board or governance committee sign-off required.
10 to 17	DEFERRED	One or more dimensions at 1. Deployment deferred until all dimensions reach 2 or above. Escalation to governance committee required.
0 to 9 or any dimension at 0	NOT APPROVED	Critical control absent. Deployment must not proceed. Full remediation programme required before re-assessment.

4. WORKED ASSESSMENT — APPLE CARD AUTOMATED SYSTEMS

The assessment below applies the framework to two Apple Card automated systems: the 2019 credit limit algorithm and the 2024 dispute routing system. Both are assessed at the point of their respective deployment failures — not as they may exist today.

This assessment is based entirely on publicly available information: the CFPB consent order (October 2024), the NYDFS investigation findings (2019 to 2020), and publicly reported technical and regulatory details. All ratings represent inferences drawn from documented outcomes and regulatory findings — not verified internal assessments of Apple or Goldman Sachs. Neither 4iGov nor the authors have access to internal documentation, system architecture records, or governance materials belonging to either organisation. Ratings reflect what the public record evidences as absent or deficient, not a definitive determination of internal governance posture. This document is a governance learning exercise and practitioner reference tool, not a legal or regulatory finding.

4.1 System Profile

Important Disclaimer — Basis of Assessment Ratings

All ratings below are inferences drawn from publicly available sources only: the CFPB consent order (October 2024), NYDFS investigation findings (2019 to 2020), and publicly reported technical details. 4iGov has no access to internal documentation, system architecture records, or governance materials belonging to Apple Inc. or Goldman Sachs. Ratings reflect governance gaps inferred from documented regulatory outcomes, not a verified or audited determination of internal controls. This assessment is a practitioner learning exercise, not a legal, regulatory, or audit finding.

Attribute	System A — Credit Limit Algorithm (2019)	System B — Dispute Routing System (2024)
System Type	AI-driven — machine learning model determining consumer credit limits	Automated — rule-based workflow routing consumer dispute messages to Goldman Sachs backend
EU AI Act Classification	High Risk — Annex III financial services, consumer credit decisions	High Risk — Annex III financial services, automated processing of consumer financial obligations
Affected Population	All Apple Card applicants and existing cardholders at credit review	All Apple Card consumers submitting billing disputes via Apple Wallet
Regulatory Obligations Triggered	Equal Credit Opportunity Act (ECOA), Fair Housing Act (FHA), TILA, NYDFS fair lending requirements	TILA Regulation Z — Billing Error Notice investigation and acknowledgment requirements, CFPB consumer protection standards
Third Party Dependency	Goldman Sachs as issuing bank; Apple as model deployer	Apple as frontend owner; Goldman Sachs as backend processor and regulated entity

4.2 Dimension-by-Dimension Assessment

	Dimension	Sys A Score	Sys A Rating	Evidence A	Sys B Score	Sys B Rating	Assessment Notes
	Governance Accountability	1	FAIL	No evidence in public record of an accountability matrix for the Apple-Goldman	0	FAIL	No owner for dispute transmission boundary. Neither Apple nor Goldman accountable for the dead state between systems.

	Dimension	Sys A Score	Sys A Rating	Evidence A	Sys B Score	Sys B Rating	Assessment Notes
				integration boundary. CFPB consent order identifies Goldman Sachs as responsible for dispute investigation but does not evidence a defined cross-party accountability framework. Integration boundary ownership inferred as absent based on documented failure.			
	Risk Classification	0	FAIL	EU AI Act was not in force at the time of either failure. Ratings reflect what would have been required under current standards. No public evidence of a formal risk classification process or documented regulatory obligations mapping for either system at deployment.	0	FAIL	No risk classification for updated workflow. TILA obligations not mapped to technical system behaviour.
	Bias and Fairness	0	FAIL	NYDFS investigation and public reporting indicate no	N/A	N/A	Not applicable to dispute routing workflow — no model-based decision affecting protected characteristics.

	Dimension	Sys A Score	Sys A Rating	Evidence A	Sys B Score	Sys B Rating	Assessment Notes
				pre-launch disparate impact testing was conducted. Apple and Goldman Sachs stated gender was not a direct model input but no proxy variable analysis was publicly evidenced. Bias assessment inferred as absent based on regulatory findings and public statements.			
	Transparency	0	FAIL	Public reporting confirms customer service representatives were unable to explain credit limit decisions to consumers. CFPB consent order documents that consumers received no notification when disputes were not transmitted. Both failures are consistent with absence of transparency controls and are inferred	0	FAIL	Silent dead state. No output to consumer on incomplete submission. No visibility to Goldman Sachs that disputes existed.

	Dimension	Sys A Score	Sys A Rating	Evidence A	Sys B Score	Sys B Rating	Assessment Notes
				accordingly.			
	Human Oversight	1	FAIL	No public evidence of a human escalation mechanism for potentially discriminatory credit decisions (System A) or for accumulated unprocessed disputes (System B). The 12-month duration of the System B failure without human detection is consistent with absence of oversight controls and is inferred accordingly.	0	FAIL	No escalation path for unprocessed disputes. No override mechanism. System operated without human intervention capability.
	Monitoring	1	FAIL	The 12-month undetected dispute processing failure in System B is the primary evidence for absent monitoring. No monitoring framework capable of detecting a systemic failure at this scale would allow a 12-month gap. Inferred as absent based on documented	0	FAIL	No monitoring on dispute transmission pipeline. Processing failure ran undetected for 12+ months.

	Dimension	Sys A Score	Sys A Rating	Evidence A	Sys B Score	Sys B Rating	Assessment Notes
				regulatory outcome. System A post-launch bias monitoring inferred as absent given no corrective action was identified until public reporting.			
	Audit Trail	2	COND	Transaction and message logs are presumed to exist given the scale of operation. Scored conditional rather than fail as the public record does not confirm complete absence of logging. The inability to reconstruct end-to-end dispute processing history across both systems is inferred from the documented failure to identify or remediate the processing gap over 12 months.	2	COND	Dispute message logs existed in Apple system but were not transmitted to Goldman Sachs. No end-to-end audit trail across both systems.

4.3 Score Summary and Deployment Verdict

Dimension	Sys A Score	Sys B Score	Key Gap
D1 — Governance Accountability	1/0	**No**	

Dimension	Sys A Score	Sys B Score	Key Gap
		accountability matrix. No integration boundary owner.	
D2 — Risk Classification	0/0	No EU AI Act classification. Regulatory obligations not mapped.	
D3 — Bias and Fairness	0/N/A	No bias assessment. No proxy variable analysis (System A only).	
D4 — Transparency and Explainability	0/0	Black box operation. Silent failure state.	
D5 — Human Oversight	1/0	No override mechanism. No escalation path.	
D6 — Monitoring and Detection	1/0	No monitoring. No alerting. Failures undetected.	
D7 — Audit Trail	2/2	Incomplete. Not reconstructable end-to-end.	
TOTAL SCORE (out of 28)	**5 / 28**	**2 / 28**	**VERDICT: NOT APPROVED — Both systems had critical control failures across 5 of 7 dimensions. Neither should have been deployed under this framework.**

5. AGENTIC AI SYSTEM EXTENSIONS

Agentic AI systems — those with autonomous, multi-step decision authority operating across tools, APIs, or workflows without human approval at each step — require additional governance controls beyond the seven core dimensions. This section defines the additional assessment criteria that apply when the system being evaluated is agentic rather than automated or AI-assisted.

What Makes a System Agentic

- The system can independently initiate actions across multiple tools, APIs, or external systems without human approval at each step
- The system maintains state or memory across interactions and uses prior context to inform subsequent decisions
- The system can delegate sub-tasks to other AI agents or automated processes
- The system has the ability to modify its own workflow or tool selection based on intermediate outcomes

	Additional Dimension	What Is Being Assessed	Pass Threshold	Fail Condition
	Autonomy Boundary Definition	The system has clearly defined boundaries on what actions it can take autonomously versus what requires human approval. Boundaries are documented and technically enforced.	Autonomy boundaries documented. Technical guardrails in place and tested.	No defined boundaries. System can take any action within its tool access without restriction.
	Goal Alignment and Constraint Verification	The system operates within defined objectives and does not pursue goals that conflict with regulatory obligations or consumer protection standards, even when optimising for efficiency.	Goal constraints documented. Adversarial testing conducted to verify constraint adherence under edge conditions.	No goal constraints. System can optimise for efficiency in ways that conflict with regulatory obligations.
	Multi-Agent Oversight	Where the agentic system delegates to sub-agents or other automated processes, governance accountability extends to those sub-systems. No sub-agent operates outside the governance framework.	Sub-agent inventory maintained. Governance controls applied to all sub-agents. No unaccounted delegation paths.	Sub-agents not inventoried. Delegation paths outside governance scope.
	Reversibility and Rollback	Actions taken by the agentic system can be reversed or rolled back where regulatory or consumer harm is identified. Rollback procedures are documented and tested.	Rollback procedures documented. Tested for all regulated action types.	Actions irreversible. No rollback capability for consumer-affecting decisions.
	Prompt and Input Security	The system is protected against prompt injection, adversarial inputs, and manipulation attempts that could cause it to act outside its defined governance boundaries.	Adversarial input testing conducted. Injection protections documented and active.	No adversarial testing. System vulnerable to inputs that bypass governance constraints.
	Regulatory Obligation Awareness	The agentic system has explicit awareness of the regulatory obligations relevant to its domain and refuses or escalates actions that would breach those obligations.	Regulatory obligation constraints embedded and tested. Escalation on constraint breach confirmed active.	No regulatory awareness. System can complete actions that breach obligations without detection or escalation.

Apple Card agentic context: The Apple Card dispute routing system was automated, not agentic — it executed a fixed workflow without autonomous decision authority. However, as financial services organisations deploy agentic AI in dispute resolution, customer service, credit review, and fraud management, these six additional dimensions become mandatory assessment criteria. The governance failures in the Apple Card case demonstrate the minimum baseline — agentic systems introduce additional risk vectors that require additional controls.

6. REMEDIATION REQUIREMENTS FOR FAILED DIMENSIONS

Any dimension rated FAIL or scoring 0 requires a remediation plan before the system proceeds to deployment or continues operation. The table below defines the minimum remediation requirements for each dimension.

	Dimension	Minimum Remediation Required	Owner	Target Timeline
	Governance Accountability	Define and document accountability matrix for system and all integration boundaries. Assign named owners. Present to governance committee.	Governance Lead	30 days
	Risk Classification	Complete EU AI Act risk tier classification. Document all regulatory obligations triggered. Obtain compliance sign-off.	Risk Officer	30 days
	Bias and Fairness	Conduct bias assessment and proxy variable analysis. Test for disparate impact across protected characteristics. Document results and mitigations.	Data Science Lead + Risk Officer	60 days
	Transparency	Implement decision explanation capability at consumer and regulator level. Train customer service on explanation provision. Test explanation quality.	Product Owner + Compliance	45 days
	Human Oversight	Define escalation paths for all failure and edge case states. Implement override mechanism. Test escalation under simulated adverse conditions.	Technology Lead + Governance Lead	45 days
	Monitoring	Implement real-time monitoring on all regulated workflow states. Configure SLA breach alerting tied to regulatory timeframes. Conduct 48-hour alert test before go-live.	Technology Lead	30 days
	Audit Trail	Implement end-to-end logging covering all system states and transitions. Confirm logs are tamper-evident and retained per regulatory requirements. Test full transaction reconstruction.	Technology Lead	45 days