

EXPLAINABLE AI

Explainable AI and Real-Time Decisioning for Automated Government Case Management in Saudi Arabia

Arshad Pasha Mohammed

Published October 2026

Abstract

Automated government case management is moving from digitised workflow toward intelligent triage, eligibility assessment, risk prioritisation and service routing. In Saudi Arabia, this transition is especially consequential because digital government programmes increasingly connect high-volume public services with data-intensive decision processes. The central challenge is not simply whether artificial intelligence can accelerate case handling, but whether decisions can remain timely, reviewable, understandable and contestable when models operate at operational speed. This review examines how explainable artificial intelligence and real-time decisioning can be combined for automated government case management in Saudi Arabia. It synthesises thirty DOI-verified studies published from 2020 to 2025 across explainability, human-AI decision support, public-sector algorithmic governance, Saudi digital government and streaming analytics. The review develops a socio-technical framework in which case intake, data quality, machine reasoning, rule validation, explanation generation, human oversight, notification, appeal and continuous monitoring form one auditable decision loop. The analysis finds that explainability is valuable only when it is matched to the decision recipient and linked to operational controls; generic feature-attribution displays alone do not guarantee justified or trustworthy outcomes. Real-time systems also require explicit latency budgets, uncertainty handling, concept-drift monitoring and escalation rules so that speed does not suppress due process. For Saudi public agencies, successful adoption depends on combining interoperable data, secure governance, accountable ownership and Arabic-capable explanation interfaces with measurable human override and appeal mechanisms. The study proposes evaluation criteria spanning predictive validity, explanation fidelity, decision latency, fairness, reviewability and institutional accountability. It concludes that explainable real-time case management should be designed as governed decision infrastructure rather than as an isolated prediction model.

Keywords: explainable AI; real-time decisioning; automated case management; digital government; public-sector AI; Saudi Arabia; algorithmic accountability; human oversight

1 Introduction

Digital government has progressed from online forms and electronic records toward integrated service environments in which data are shared across systems, workflows are automated and decisions can be supported by machine learning. This progression creates a distinct opportunity for government case management. Applications for benefits, permits, registrations, inspections, compliance actions, grants, licensing, complaints and other administrative services often follow repetitive but information-rich pathways. Artificial intelligence can classify documents, estimate risk, identify missing information, recommend next actions and prioritise cases. Yet public-sector decisions differ from ordinary commercial recommendations because they can affect rights, obligations, access to services and the distribution of public resources. A technically accurate prediction is

therefore insufficient if the reasoning cannot be inspected, challenged or connected to the legal and procedural basis of the case.

Explainable artificial intelligence (XAI) responds to this problem by making model behaviour or individual outputs more intelligible to human users [1–6]. The literature distinguishes intrinsically interpretable models from post-hoc explanation methods and stresses that an explanation should be evaluated for fidelity, clarity, stability and usefulness rather than treated as a decorative interface [2–7]. Human-centred studies also show that more explanation does not automatically produce better decisions. Practitioners may over-trust interpretability tools [8], explanation methods can be manipulated [9], and explanations can increase acceptance of machine advice even when the advice is wrong [11–14]. These findings are important for public administration, where an explanation must support calibrated reliance instead of merely reassuring the user.

A second requirement is timeliness. Government case-management platforms increasingly operate as event-driven systems: a document is uploaded, an identity attribute is verified, a payment is received, a deadline expires or an external registry changes. Decisions may therefore need to be updated immediately rather than after a periodic batch process. Real-time decisioning can reduce backlogs and shorten service cycles, but it introduces engineering and governance risks. Data distributions can change, model confidence can deteriorate, explanations can add computational delay and an automated action can propagate before a human reviewer notices an error. Research on streaming data shows that concept drift can quietly reduce classifier performance when the environment evolves [30]. Consequently, real-time public decisioning must be designed around monitoring, fallback states and controlled escalation, not only low latency.

Saudi Arabia provides a significant setting for this discussion. Studies of Saudi public organisations show movement toward AI assimilation, smart-government adoption and digital service expansion, while also identifying organisational attention, security, trust, managerial readiness and infrastructure as decisive conditions [25–29]. The smart-city XAI survey by Javed et al. [1] is a useful foundation because it frames explainability as essential to trustworthy smart-city decision making and explicitly identifies smart governance, real-time systems, privacy, security and regulatory compliance as connected research concerns. However, a general smart-city lens does not resolve the more specific design problem of high-volume administrative cases, where each recommendation must remain linked to evidence, policy rules, human authority and an appeal path.

This review addresses that gap. It treats automated case management as a socio-technical decision system rather than a single algorithm. The contribution is threefold: first, it synthesises recent XAI and public-governance evidence around the operational lifecycle of a government case; second, it develops an architecture for real-time, explainable and human-supervised decisioning; and third, it proposes evaluation criteria and a research agenda tailored to Saudi digital-government conditions. The resulting framework emphasises that performance, explanation, fairness, security and procedural review must be engineered together.

2 Aim and objectives

The aim of this review is to develop an evidence-based framework for integrating explainable AI with real-time decisioning in automated government case-management systems in Saudi Arabia. The emphasis is on administrative decision support and controlled automation rather than unrestricted replacement of human judgement.

Four objectives guide the analysis. First, the review identifies which forms of explanation are operationally useful at different stages of a case, including triage, eligibility assessment, prioritisation, exception handling and appeal. Second, it examines how low-latency inference can coexist with data validation, uncertainty estimation, rule checks and human escalation. Third, it analyses governance requirements involving accountability, fairness, privacy, security, auditability and organisational ownership. Fourth, it converts these findings into a practical assurance model and research agenda for Saudi public-sector deployment. These objectives deliberately separate the value of prediction from the value of a defensible administrative decision: the former is a model property, whereas the latter depends on the complete institutional process in which the model is embedded.

3 Review methodology

A structured integrative review was used because the research problem spans computer science, human-computer interaction, public administration, law, digital government and data-stream analytics. The search scope was limited to peer-reviewed work published from 2020 through 2025 so that the synthesis reflects contemporary AI deployment practices. Candidate studies were identified through scholarly search and publisher-indexed records using combinations of terms including “explainable artificial intelligence”, “automated decision making”, “public sector”, “government”, “case management”, “human-AI decision making”, “algorithmic auditing”, “Saudi Arabia”, “digital government”, “smart government”, “real-time decisioning” and “concept drift”.

Inclusion required direct relevance to at least one of five analytical domains: XAI methods and evaluation; human use of explanations; public-sector AI governance and accountability; Saudi digital-government adoption; or real-time and streaming decision reliability. Studies were excluded when they were outside the 2020-2025 period, lacked a verifiable DOI, were purely promotional, or did not contribute to the decision lifecycle examined here. Thirty sources were retained. DOI records were cross-checked against publisher or scholarly-index metadata, and the final evidence base deliberately combines broad reviews with empirical studies so that conceptual claims are balanced by observations of actual human or organisational behaviour.

The synthesis followed a thematic matrix rather than a vote-counting approach because the included studies use heterogeneous methods and outcomes. Evidence was coded against six themes: explanation purpose, human reliance, decision speed, governance control, Saudi implementation conditions and lifecycle assurance. Findings were then mapped onto a generic government case pathway from intake to post-decision monitoring. This method does not claim statistical meta-analysis, nor does it infer effectiveness where comparable effect sizes are unavailable. Its purpose is analytical integration: identifying recurring design principles, tensions and gaps that can inform a coherent architecture for explainable real-time case management.

4 Explainability as a decision control

XAI is often described as a bridge between complex models and human understanding, but that description is too broad for administrative systems. In case management, an explanation must perform a specific control function. It can help a case officer verify whether the model used relevant evidence, help a supervisor investigate an exception, help an auditor reconstruct a decision, or help an applicant understand what affected an outcome. These users do not need identical information. Reviews of XAI consistently show that explanation quality is multidimensional and that different techniques expose different aspects of model behaviour [2–7]. Feature attributions can identify influ-

ential variables, counterfactuals can show what would have changed the result, and interpretable rule models can expose decision structure directly.

The operational lesson is that explanation should be role-conditioned. A case officer may need concise reasons, confidence and policy references; a data scientist may need sensitivity diagnostics and feature behaviour; an auditor may need model version, evidence lineage and a replayable decision trace; and a citizen may need plain-language reasons and actionable information about review or appeal. Providing the same technical chart to all users can create apparent transparency without practical understanding. Vilone and Longo [4] and Nauta et al. [7] emphasise the importance of evaluating explanation properties, while human-subject research shows that explanations can fail to improve uncertainty awareness or trust calibration [12,13]. Explanation fidelity is equally important. Post-hoc methods approximate or probe model behaviour, so they can produce unstable or misleading representations. Slack et al. [9] demonstrate that popular explanation methods can be fooled under adversarial conditions. Kaur et al. [8] show that even expert users can misunderstand interpretability tools. These findings imply that a government platform should not present explanations as unquestionable evidence. Each explanation should carry provenance: model identifier, explanation method, input snapshot, timestamp, confidence indicators and known limitations. Where a decision can materially affect an individual, the explanation should be supplemented by policy-grounded reasons and a human-review mechanism.

Counterfactual explanations are especially attractive in case management because they can turn a score into an actionable statement: which feasible change in evidence or circumstance could alter the outcome [10]. However, public administration imposes constraints on actionability. A counterfactual must not suggest changing immutable attributes, manipulating a protected characteristic or supplying information irrelevant to the governing rule. Feasible counterfactuals should therefore be generated inside a policy-constrained space. The system must distinguish between factual correction, additional evidence and genuinely changeable conditions. This makes explanation generation part of decision governance, not merely a visualisation step.

5 Real-time decisioning in government case work

Real-time decisioning means that the system can evaluate new case information and produce or update a recommendation within an operationally defined time window. That definition is more useful than treating “real time” as a universal millisecond target. The correct latency target is therefore risk-sensitive. Public agencies should define service-level budgets that allocate time across data retrieval, identity checks, model inference, explanation generation, rule validation, logging and, where required, human review.

This decomposition exposes an important trade-off. The fastest model is not necessarily the fastest trustworthy decision. Real-time architecture should therefore favour bounded, observable stages. Each stage should either return a validated result, a confidence state or an escalation signal. This principle aligns with broader XAI literature that connects trustworthy use to transparency, security and monitoring [1–6].

Streaming environments also create non-stationarity. Concept-drift research shows that models trained on historical distributions may degrade as incoming data evolve [30]. For government case management, drift monitoring should extend beyond predictive accuracy. Agencies should watch changes in input distributions, missing-data rates, explanation patterns, override rates, appeal

outcomes and subgroup error rates. A rise in human overrides or successful appeals may signal a problem before aggregate accuracy metrics do.

A robust real-time design also requires uncertainty-aware routing. Cases near a decision threshold, cases with missing critical evidence, cases outside the training distribution and cases triggering conflicting rules should not be forced through the same automated path as routine cases. Instead, the decision engine should assign an automation state: auto-process, request information, hold for verification, or escalate for human review. This architecture preserves speed where confidence is high while reserving human attention for ambiguity. It also prevents a common failure mode in which automation converts model uncertainty into false administrative certainty.

6 Automated case management as a socio-technical system

Public-sector research increasingly treats AI adoption as an organisational and governance problem rather than a software procurement problem [15–24]. Automated case management illustrates why. A model may be developed by one team, integrated by another, operated by case workers, monitored by a digital office and governed by legal, audit or policy functions. Responsibility can become fragmented precisely when the system appears most automated. Busuioc [19] argues that algorithmic systems create distinctive accountability challenges, while Raji et al. [18] and Mökander et al. [22] show the importance of lifecycle auditing rather than one-time review. The appropriate unit of assurance is therefore the decision service, including data, model, rules, interface, staff practices and post-decision remedies.

Figure 1 represents this service as a closed operational loop. Digital intake begins the case, but intake data should not flow directly into a model. Identity and evidence validation are separate controls because erroneous source data cannot be repaired by explainability. Data integration then assembles authorised records from internal or external systems with timestamps and provenance. The decision engine combines model outputs with deterministic rules where appropriate. An explanation layer translates model reasoning into role-specific reasons, while a policy and compliance layer checks whether the proposed action is procedurally permissible. High-risk or uncertain cases are routed to human review. The final decision is communicated together with applicable reasons and review options, while audit logging captures the evidence, model version, rules, explanation and human actions. Monitoring closes the loop by detecting drift, anomalies, recurrent overrides and appeal patterns.

This architecture deliberately separates prediction from administrative action. A risk score can inform a decision without becoming the decision. That distinction is important because the legal or policy basis of an action may depend on explicit conditions that should not be inferred indirectly from statistical correlation. It also reduces the danger of “automation bias”, where staff accept model outputs without sufficient scrutiny. Bansal et al. [11] and Alufaisan et al. [12] show that explanations do not automatically improve human-AI team performance. Effective oversight therefore requires interface design, training and authority to disagree with the model.



Figure 1. End-to-end framework for explainable AI and real-time decisioning in government case management.

7 Saudi Arabia: adoption conditions and design priorities

Saudi public-sector studies provide strong evidence that organisational conditions will shape the effectiveness of any explainable case-management programme. Alshahrani, Dennehy and Mäntymäki [25] show that AI assimilation in Saudi public organisations depends on how managerial attention is directed toward technological and organisational challenges. Alajmi et al. [27] identify security concerns, strategy, managerial support, incentives, trust and cultural values among determinants of smart-government adoption. Majrashi [28] further reports that perceived usefulness, perceived impact and ethical dimensions such as transparency, privacy and security influence managers' intentions and trust. These studies suggest that deployment cannot be sustained through technical performance alone.

Data governance is a first priority. Government cases may draw on identity, employment, licensing, financial, geographic or service-history information. Integration can improve accuracy and reduce repeated requests to applicants, but it also expands the consequences of stale, incorrect or over-broad data access. Research on Saudi e-government security highlights the importance of protecting confidentiality, integrity and availability [26]. In an explainable system, data lineage should therefore be visible at decision time: which source supplied a field, when it was retrieved, whether it was verified and whether it was transformed before inference. A second priority is institutional interoperability. Real-time case management often requires multiple agencies or registries to exchange data. The architecture should avoid assuming that all sources have identical update cycles or quality. Each source needs freshness rules, fallback behaviour and ownership. If a critical registry is unavailable,

the system should degrade safely rather than substitute unverified inference. A third priority is human-centred communication in the local service environment. Explanations should be available in clear Arabic and, where service design requires, English, with consistent terminology across digital channels and case-worker interfaces. Technical feature names should be mapped to administratively meaningful concepts. For example, a model variable representing a composite history score should not be displayed as a raw internal code. The interface should distinguish facts, model inferences, policy rules and human judgements so that users can see what kind of reason they are reading. This distinction supports both citizen understanding and internal accountability.

Finally, Saudi digital-government development is continuing to expand into emerging technologies. A 2025 multi-case study of Saudi government sectors identifies drivers, enablers and challenges associated with adopting AI and other emerging technologies, including ethical concerns, resources and implementation conditions [29]. This reinforces the need for reusable governance patterns. Rather than building separate assurance logic for every ministry or case type, public institutions can standardise model documentation, explanation records, human-override coding, audit events and monitoring indicators while allowing policy-specific rules to remain local.

8 Explanation design for case officers, citizens and auditors

A practical explanation framework should answer different questions at different levels. For the case officer, the core questions are: what outcome is recommended, which evidence materially influenced it, how confident is the system, which rule or threshold is implicated, and what should trigger escalation? For the citizen, the questions are: what was decided, what information was considered, what were the main reasons, is any information missing or incorrect, and how can the decision be reviewed? For the auditor, the questions are: which model and data versions were used, whether the explanation faithfully represented the model, who approved or overrode the recommendation, and whether similar cases were treated consistently.

The literature suggests that feature contribution explanations can improve subjective understanding in some contexts, but their effectiveness depends on user expertise and model complexity [13]. Shin [14] links explainability and causability to trust and acceptance, while public-sector work highlights the relationship between transparency and perceived legitimacy [16]. These findings support a layered explanation interface. The first layer should present a concise outcome and two or three dominant reasons in plain language. A second layer can show supporting evidence, confidence, counterfactual options and policy references. A third technical layer can expose model diagnostics to authorised specialists. Progressive disclosure prevents cognitive overload while preserving depth for review.

Explanations should also be contrastive where possible. People commonly ask why one outcome occurred instead of another. This is more useful than listing ten weighted features. Counterfactual methods can support such reasoning [10], but constraints must be encoded so that suggested changes are lawful, feasible and relevant. The system should never imply that an applicant must change an immutable characteristic to obtain a different outcome.

Crucially, explanation quality should be tested with task-based measures. Nauta et al. [7] identify numerous properties for evaluating XAI, and human-centred studies warn that perceived clarity can diverge from actual decision support [8,11–13]. Government evaluation should therefore measure whether case workers detect model errors, whether citizens correctly understand the main reason, whether auditors can reproduce a decision and whether explanations remain stable under immateri-

al input changes. A high satisfaction score alone is not evidence of a trustworthy explanation. Table 1 summarises common explanation patterns and their operational fit for government case management.

Table 1. Explanation patterns and their operational fit for government case management.

Technique	Primary output	Typical fidelity	Latency impact	Strongest use	Key limitation
Interpretable rules / trees	Logic is directly inspectable	High when model itself is shown	Very low	Stable eligibility and routing logic	May lose performance if relationships are highly nonlinear
Feature contribution	Shows variables pushing a prediction up or down	Medium; depends on method and model	Low to medium	Officer diagnostics and case-level reasons	Can be unstable or misunderstood; requires fidelity testing
Counterfactual explanation	Shows feasible changes associated with a different outcome	Medium to high for actionability	Medium	Citizen-facing review and remediation	Must respect policy, causality, feasibility and immutable attributes
Example-based explanation	Shows similar or representative cases	Variable	Low to medium	Training, comparison and anomaly review	Privacy leakage and misleading similarity are risks
Surrogate / extracted rules	Approximates a complex model with simpler local logic	Variable; must be measured	Medium	Audit investigation and model debugging	Approximation can diverge from the underlying model
Uncertainty + abstention	Explains when the model should not decide	Not a causal explanation, but critical assurance	Very low	Escalation and human-review routing	Requires calibrated confidence and clear fallback policy

9 Governance, fairness, security and procedural safeguards

Public decision systems require governance controls before, during and after deployment. The literature on public-sector AI repeatedly identifies accountability, transparency and the distribution of responsibility as unresolved challenges [15,17–24]. Kuziemski and Misuraca [15] show how AI governance intersects with data governance and regulatory practices. Zuiderwijk et al. [20] find that public-governance research contains significant conceptual and practice-driven work but still needs deeper empirical evidence. Mahroof et al. [24] demonstrate that algorithmic decision-making can also shift organisational power among managers and analysts. These findings indicate that a governance framework must specify not only model quality but decision rights.

Every automated case service should therefore have named owners for data, model performance, policy logic, operational workflow, security and final administrative authority. Ownership should be recorded in deployment documentation and linked to change control. No model update should silently alter the decision policy. Material changes in features, thresholds or training data should trigger validation, approval and versioning. Likewise, explanation templates should be tested whenever the model changes, because an explanation that was faithful to one model version may become misleading after retraining.

Fairness requires both statistical and procedural analysis. Mehrabi et al. [23] catalogue multiple sources and definitions of bias in machine learning, illustrating that fairness cannot be reduced to a single metric. Government agencies should select fairness measures that fit the case context, protected interests and operational objective, then examine error rates and outcomes across relevant groups where lawful and appropriate. At the same time, procedural fairness requires consistent evidence standards, comparable escalation rules and meaningful opportunities to correct data or seek review. A model can satisfy one statistical parity measure while still producing a process that citizens cannot understand or challenge.

Security and privacy must also extend to explanations. An explanation can unintentionally reveal sensitive features, internal risk logic or information about other records. Access should therefore be role-based, and citizen-facing explanations should disclose enough to support understanding without exposing protected system details. Internal audit records can retain deeper technical information under controlled access. The Saudi e-government security literature reinforces the importance of data safeguards [26], while broader XAI work shows that security and transparency should be considered together rather than as competing afterthoughts [1,2].

Procedural safeguards are especially important when automation is consequential. Suksi [21] argues that automated public decisions can challenge established administrative due-process mechanisms. In practical system design, this means preserving human authority for defined categories of high-impact decisions, recording when a human accepts or overrides a model, and providing an appeal or reconsideration path. Automation should never make responsibility disappear; it should make responsibility more traceable. Figure 2 summarises these safeguards as a layered assurance model.



Figure 2. Layered assurance model for responsible AI in public-sector case management.

10 Evaluating a real-time explainable case-management system

Traditional model metrics such as accuracy, precision, recall and area under the curve remain necessary, but they are not sufficient for a government case service. Evaluation should be multidimensional. Predictive validity measures whether the model estimates the intended outcome accurately. Calibration measures whether confidence scores correspond to observed correctness. Robustness tests whether small, irrelevant input changes alter results unexpectedly. Fairness analysis examines distributional error patterns. Explanation fidelity measures whether stated reasons reflect the model. Human-factors evaluation measures whether users understand the explanation and make better calibrated decisions. Operational evaluation measures latency, uptime, queue behaviour and safe degradation. Governance evaluation measures audit completeness, override traceability, appeal resolution and change-control compliance.

Table 2 translates these dimensions into the case lifecycle. The central point is that evidence should be collected where risk arises. Data quality is measured at intake and integration, not only after model failure. Explanation performance is measured at the point of human use, not only in a laboratory. Appeal outcomes and override reasons become feedback signals for monitoring. Latency should also be treated as a distribution rather than an average. Agencies should track percentile latency for each component, define maximum tolerated delay and specify fallback behaviour. Explanation generation must be included in the budget. If a complex post-hoc method is too slow for a particular pathway, the design can use an interpretable model, pre-computed global diagnostics or a two-stage process in which a fast decision is held from execution until the required explanation is available.

Monitoring should connect technical and administrative signals. Drift detectors can identify changing data patterns [30], but human review and appeals contain equally valuable evidence. Rising override rates may indicate model misalignment; concentrated appeals may indicate unclear explanations; repeated data corrections may reveal upstream quality problems. The monitoring dashboard should therefore integrate model metrics with workflow metrics and citizen-resolution indicators. This approach turns operational experience into governed learning rather than uncontrolled self-modification.

Table 2. Lifecycle controls for explainable real-time government case management.

Case stage	AI / automation role	Explanation & assurance evidence	Real-time control	Example metrics
Intake and evidence capture	Document classification, completeness checks	Source provenance; missing evidence; confidence	Validate format and identity before inference	missing-data rate; extraction accuracy; intake latency
Data integration	Record matching and feature assembly	Source, timestamp and transformation lineage	Freshness rules; fail closed or hold when critical sources unavailable	source availability; stale-record rate; match error
Decisioning	Risk, eligibility or routing recommendation	Key reasons; confidence; threshold position	Risk-sensitive latency budget; abstain on uncertainty	accuracy; calibration; p95 latency; abstention rate
Policy control	Deterministic rules and procedural constraints	Applicable rule and conflict status	Block impermissible actions; escalate rule-model conflicts	rule violations; conflict frequency

Case stage	AI / automation role	Explanation & assurance evidence	Real-time control	Example metrics
Human review	Officer assessment and override	Model reason plus evidence trace	Prioritise high-risk, novel and low-confidence cases	override rate; reviewer agreement; handling time
Notification and appeal	Decision communication and case reopening	Plain-language reasons; correction and review path	Do not execute consequential action before required explanation is available	comprehension; appeal rate; upheld/reversed decisions
Monitoring	Drift, fairness, operational and audit analytics	Versioned history of decision and explanation	Alert on drift, subgroup errors, unusual overrides or failures	drift indicators; subgroup error; audit completeness; uptime

11 Implementation roadmap and research agenda

A staged implementation strategy is preferable to immediate end-to-end automation. The first stage should select a bounded case type with stable policy rules, adequate historical data and measurable outcomes. The objective should be decision support, not maximum automation. Baseline measures should capture current processing time, error patterns, escalation rates and appeal outcomes before AI is introduced. The second stage should establish the assurance foundation: data lineage, access controls, model documentation, approved explanation methods, latency targets, fairness tests, override codes and audit logging. Staff should be trained not only to read explanations but to recognise uncertainty and challenge the model. Human-AI studies show that explanations can increase reliance without improving correctness [11–13], so training should include deliberately incorrect recommendations and ambiguous cases. Supervisors should review patterns of acceptance and override to detect automation bias.

The third stage can introduce controlled automation for low-risk cases that satisfy explicit confidence and data-quality criteria. High-risk, novel or conflicting cases should remain human-reviewed. Automation thresholds should be adjustable through governed configuration rather than hard-coded application logic. As evidence accumulates, agencies can expand the automated boundary while retaining monitoring and rollback capability. This incremental approach aligns with Saudi evidence that organisational readiness, managerial support, trust and security conditions shape adoption [25,27–29].

Research should now move beyond demonstrations of explanation techniques toward field evaluation of complete administrative systems. Five questions are especially important. First, which explanation formats best improve error detection by Saudi case officers without causing information overload? Second, how should Arabic administrative language represent uncertainty, counterfactuals and model limitations in ways citizens understand? Third, what combination of drift indicators, override patterns and appeal outcomes provides the earliest warning of model deterioration? Fourth, how should latency budgets change with case risk, especially when external registries are unavailable? Fifth, how can agencies measure whether explainability improves procedural fairness rather than merely perceived transparency?

A further research priority is institutional learning across agencies. Shared case-management platforms may enable common technical components, but policy contexts differ. Comparative studies should distinguish reusable infrastructure from domain-specific decision logic. Standardised audit

events and explanation schemas could make cross-agency assurance more efficient, while preserving local authority over substantive rules.

12 Conclusion

Explainable AI and real-time decisioning can strengthen automated government case management when they are treated as complementary parts of a governed decision process. Explainability without reliable data, human authority or appeal mechanisms can create an illusion of transparency. Real-time inference without monitoring and uncertainty controls can accelerate error. For Saudi Arabia, the evidence points toward a layered architecture that combines validated data integration, risk-sensitive decision engines, role-specific explanations, policy checks, human escalation, secure logging and continuous monitoring. The framework should preserve a clear distinction between statistical recommendation and administrative action, with named responsibility for each. Evaluation should span accuracy, calibration, explanation fidelity, fairness, latency, override behaviour, appeal outcomes and auditability.

The central implication is that trustworthy government automation is not achieved by selecting the most accurate model or adding an explanation panel after deployment. It is achieved by engineering an accountable service in which every consequential decision can be traced from evidence to model, from model to rule, from rule to human authority and from outcome to review. This is the design principle most likely to reconcile faster public services with transparency, procedural protection and sustainable institutional trust.

References

- [1] Javed AR, Ahmed W, Pandya S, Maddikunta PKR, Alazab M, Gadekallu TR. A survey of explainable artificial intelligence for smart cities. *Electronics*. 2023;12(4):1020. <https://doi.org/10.3390/electronics12041020>
- [2] Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*. 2020;58:82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- [3] Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller KR. Explaining deep neural networks and beyond: a review of methods and applications. *Proceedings of the IEEE*. 2021;109(3):247–278. <https://doi.org/10.1109/JPROC.2021.3060483>
- [4] Vilone G, Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*. 2021;76:89–106. <https://doi.org/10.1016/j.inffus.2021.05.009>
- [5] Angelov PP, Soares EA, Jiang R, Arnold NI, Atkinson PM. Explainable artificial intelligence: an analytical review. *WIREs Data Mining and Knowledge Discovery*. 2021;11(5):e1424. <https://doi.org/10.1002/widm.1424>
- [6] Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, et al. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*. 2023;99:101805. <https://doi.org/10.1016/j.inffus.2023.101805>
- [7] 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 Computing Surveys*. 2023;55(13s):295. <https://doi.org/10.1145/3583558>
- [8] Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Wortman Vaughan J. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020. p. 1–14. <https://doi.org/10.1145/3313831.3376219>
- [9] Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*. 2020. p. 180–186. <https://doi.org/10.1145/3375627.3375830>

- [10] Mothilal RK, Sharma A, Tan C. Explaining machine learning classifiers through diverse counterfactual explanations. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020. p. 607–617. <https://doi.org/10.1145/3351095.3372850>
- [11] Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, et al. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021. Article 81. <https://doi.org/10.1145/3411764.3445717>
- [12] Alufaisan Y, Marusich LR, Bakdash JZ, Zhou Y, Kantarcioglu M. Does explainable artificial intelligence improve human decision-making? *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021;35(8):6618–6626. <https://doi.org/10.1609/aaai.v35i8.16819>
- [13] Wang X, Yin M. Effects of explanations in AI-assisted decision making: principles and comparisons. *ACM Transactions on Interactive Intelligent Systems*. 2022;12(4):27. <https://doi.org/10.1145/3519266>
- [14] Shin D. The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. *International Journal of Human-Computer Studies*. 2021;146:102551. <https://doi.org/10.1016/j.ijhcs.2020.102551>
- [15] Kuziemski M, Misuraca G. AI governance in the public sector: three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy*. 2020;44(6):101976. <https://doi.org/10.1016/j.telpol.2020.101976>
- [16] de Fine Licht K, de Fine Licht J. Artificial intelligence, transparency, and public decision-making. *AI & Society*. 2020;35:917–926. <https://doi.org/10.1007/s00146-020-00960-w>
- [17] Valle-Cruz D, Criado JI, Sandoval-Almazán R, Ruvalcaba-Gómez EA. Assessing the public policy-cycle framework in the age of artificial intelligence: from agenda-setting to policy evaluation. *Government Information Quarterly*. 2020;37(4):101509. <https://doi.org/10.1016/j.giq.2020.101509>
- [18] Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020. p. 33–44. <https://doi.org/10.1145/3351095.3372873>
- [19] Busuioc M. Accountable artificial intelligence: holding algorithms to account. *Public Administration Review*. 2021;81(5):825–836. <https://doi.org/10.1111/puar.13293>
- [20] Zuiderwijk A, Chen YC, Salem F. Implications of the use of artificial intelligence in public governance: a systematic literature review and a research agenda. *Government Information Quarterly*. 2021;38(3):101577. <https://doi.org/10.1016/j.giq.2021.101577>
- [21] Suksi M. Administrative due process when using automated decision-making in public administration: some notes from a Finnish perspective. *Artificial Intelligence and Law*. 2021;29:87–110. <https://doi.org/10.1007/s10506-020-09269-x>
- [22] Mökander J, Morley J, Taddeo M, Floridi L. Ethics-based auditing of automated decision-making systems: nature, scope, and limitations. *Science and Engineering Ethics*. 2021;27:44. <https://doi.org/10.1007/s11948-021-00319-4>
- [23] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Computing Surveys*. 2021;54(6):115. <https://doi.org/10.1145/3457607>
- [24] Mahroof K, Weerakkody V, Hussain Z, Sivarajah U. Navigating power dynamics in the public sector through AI-driven algorithmic decision-making. *Government Information Quarterly*. 2025;42(3):102053. <https://doi.org/10.1016/j.giq.2025.102053>
- [25] Alshahrani A, Dennehy D, Mäntymäki M. An attention-based view of AI assimilation in public sector organizations: the case of Saudi Arabia. *Government Information Quarterly*. 2022;39(4):101617. <https://doi.org/10.1016/j.giq.2021.101617>
- [26] Alharbi AS, Halikias G, Rajarajan M, Yamin M. A review of effectiveness of Saudi E-government data security management. *International Journal of Information Technology*. 2021;13:573–579. <https://doi.org/10.1007/s41870-021-00611-3>
- [27] Alajmi M, Mohammadian M, Talukder M. The determinants of smart government systems adoption by public sector organizations in Saudi Arabia. *Heliyon*. 2023;9(10):e20394. <https://doi.org/10.1016/j.heliyon.2023.e20394>
- [28] Majrashi K. Determinants of public sector managers' intentions to adopt AI in the workplace. *International Journal of Public Administration in the Digital Age*. 2024;11(1):1–26. <https://doi.org/10.4018/IJPADA.342849>

-
- [29] Alshahrani A. Adopting emerging technologies in digital government: a multi-case analysis of drivers, enablers, and challenges in Saudi Arabia. *Digital Government: Research and Practice*. 2025;6(3):36. <https://doi.org/10.1145/3719297>
- [30] Agrahari S, Singh AK. Concept drift detection in data stream mining: a literature review. *Journal of King Saud University - Computer and Information Sciences*. 2022;34(10B):9523–9540. <https://doi.org/10.1016/j.jksuci.2021.11.006>