

Responsible AI Governance for Enterprise and Government Digital Platforms in Saudi Arabia: Integrating AI Innovation, Data Privacy and Cyber Resilience

Wamiq Rafi Syed^{1*} 

¹Veer Bahadur Singh Purvanchal University: Jaunpur, Uttar Pradesh, IN

*Corresponding Author

Wamiq Rafi Syed

Veer Bahadur Singh
Purvanchal University:
Jaunpur, Uttar Pradesh, IN

Article History

Received: 13.08.2026

Accepted: 05.10.2026

Published: 08.10.2026

Abstract: In Saudi Arabia, artificial intelligence is being introduced into public services, business operations and the digital platforms that deal with citizens, thereby increasing the demand for a kind of governance that allows innovation without sacrificing privacy, accountability or cyber resilience. This review brings together the scholarly work published between 2020 and 2025 in order to examine how responsible AI can be put into practice across the whole platform lifecycle rather than simply remaining as a set of general ethical principles. The analysis is grounded in four areas of evidence: organisational AI governance, data practices that ensure privacy, the use of trustworthy and explainable AI, and adversarial cybersecurity. It argues that these different areas are generally handled through separate compliance pathways, even though a failure in one can affect the others poor data provenance can result in unfair or insecure models; opaque models can make it hard to carry out audits or to investigate incidents; and a cyber breach can wipe out the privacy and accountability safeguards. A structured integrative review of 30 scholarly sources, verified by their DOI numbers, is used to develop a governance framework tailored to the Saudi context for application in both corporate and government platforms. The proposed framework connects the six lifecycle stages use-case intake, data stewardship, model assurance, secure release, continuous monitoring and controlled retirement with cross-cutting elements such as accountability, human oversight, third-party assurance and evidence retention. The review concludes that responsible AI is most credible when governance is expressed in the form of testable controls, clearly defined decision rights and continuously updated evidence of assurance. For organisations in Saudi Arabia, the resulting approach offers a practical means of achieving fast digital transformation while incorporating privacy-by-design and resilience-by-design, and at the same time remaining adaptable to sectoral risks and changing national requirements.

Keywords: Artificial Intelligence Governance; Responsible AI; Data Privacy; Explainable AI; Cybersecurity; AI Assurance; Privacy by Design.

Copyright © 2026 The Author(s): This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY-NC 4.0) which permits unrestricted use, distribution, and reproduction in any medium for non-commercial use provided the original author and source are credited.

1. INTRODUCTION

Artificial intelligence has moved on from being employed in single analytics projects to now being an integral part of the core activities of digital

organisations. Nowadays, various systems including recommendation engines, fraud detection tools, automated triage procedures, virtual assistants, predictive maintenance systems, identity services

Citation: Syed, W. R. (2026). Responsible AI Governance for Enterprise and Government Digital Platforms in Saudi Arabia: Integrating AI Innovation, Data Privacy and Cyber Resilience; *Glob Acad J Econ Buss*, 8(5), 1608-1619.

<https://doi.org/10.36348/gajeb.2026.v08i05.048>

and generative interfaces all contribute to the way in which organisations allocate their attention, customise their services and manage operational risk. While this expansion brings about both economic and service benefits, it also changes the nature of governance. The question is no longer just whether an algorithm complies with a given set of ethical principles at a particular time; governance now has to determine who is permitted to authorise a use case, which data may be used, how the models are tested, when human intervention is needed, how third-party components are managed, and what evidence is kept when there are changes in performance, threats or legal requirements. Repeated reviews of AI ethics have pointed out the gap between general principles and real-world practice, and research into organisational auditing indicates that accountability relies on having traceable processes that cover the entire system lifecycle (Hagendorff, 2020; Morley *et al.*, 2020; Raji *et al.*, 2020).

This is particularly significant for the digital platforms that are used by businesses and governments. Such platforms generally comprise a variety of different components; they involve identity stores, application programming interfaces, cloud services, old databases, workflow engines, analytics pipelines and other external software. As AI is included in these platforms, it takes on the risks that are inherent in the environment and may in fact exacerbate them. A biased dataset can have an impact on a number of the services that follow it; if a model is made available through an insecure interface then it becomes a target for attack; and if a model is supplied by a vendor then it can be modified without the organisation which has deployed it having control over the data used to train the model or the vendor's internal architecture.

Saudi Arabia provides a suitable setting in which to look at this issue of governance. As a result of the nation's digital transformation, data and artificial intelligence have assumed a more strategic role in public administration and in areas such as finance, health, telecommunications, energy and smart infrastructure. Research carried out in the region shows that governance is progressing alongside bold initiatives, since there is growing attention being paid to institutional responsibility, cultural appropriateness, privacy, cybersecurity and cooperation between different sectors (Ibrahim, 2024; Trigui *et al.*, 2024). Studies led by Saudi Arabia on privacy and data sharing also demonstrate that AI governance cannot be separated from the proper management of sensitive data and from the development of interoperable digital infrastructure (Alsaigh *et al.*, 2024). Moreover, national cybersecurity capability has become a key

component of digital transformation, emphasizing the need to treat AI security as part of the overall enterprise risk rather than as a separate issue that is addressed later (AlDaajeh *et al.*, 2022).

The literature remains fragmented: in the case of responsible AI, the focus is generally on fairness, transparency or accountability; in the area of privacy research, attention is given to data minimisation, federated learning and technical protection; and cybersecurity research looks at adversarial examples, poisoning, intrusion or attack response. While these areas are useful from an analytical standpoint, they are operationally interdependent. Privacy controls can have an impact on model utility, explanation methods might reveal information, security monitoring may require sensitive telemetry, and modifications made to a model in order to address a vulnerability could render previous fairness or performance tests invalid. As a result, a governance system that adjusts each of these aspects separately ends up creating unmanaged trade-offs.

The article addresses this issue by treating innovation, privacy and cyber resilience as interrelated design constraints. It is based on evidence from the period 2020 to 2025 and offers a lifecycle-oriented framework that is appropriate for the real conditions experienced by enterprises and government platforms in Saudi Arabia. This method is consistent with research that aims to move responsible AI from simple declarations to operational controls and verifiable organisational practices (Mökander and Floridi, 2021; Camilleri, 2024; Papagiannidis *et al.*, 2025).

2. Saudi Digital Platform Governance Context

Saudi digital platforms are at the junction where public value, commercial innovation and national infrastructure come together. Today, government services are increasingly dependent on the sharing of data and on digital identity, while companies use AI to automate their decisions, detect anomalies and tailor their interactions. As a result, faults can go beyond the boundaries of the team that developed a particular model; a defective decision rule could affect citizen access, the way customers are treated or the continuity of operations; if the training data comes to light then there could be a breach of privacy; and if a model is compromised it might be exploited as a way of gaining access to other connected services. Governance should therefore adopt a platform approach and consider the dependencies, interfaces and future reuse of the services rather than viewing each model as an independent entity.

At the regional level, the literature points out three features that show the importance of this perspective. First, state institutions have a major role to play in shaping AI strategy and in encouraging the use of AI in the public sector, and it follows that organisational governance should connect technical assurance with formal decision-making power (Ibrahim, 2024; Trigui *et al.*, 2024). Second, as there are expectations concerning data protection and data sharing, it is necessary to know at what stage personal or sensitive information enters the AI lifecycle, how it is transformed and which external parties are able to access it. The academic work carried out in Saudi Arabia on governance and privacy stresses the need to balance innovation with protection rather than seeing data access and privacy as opposing objectives (Alsaigh *et al.*, 2024). Third, cyber resilience is closely connected with service continuity. Research into the national cybersecurity strategy shows that institutional capacity, education and coordinated controls actually have an effect on the resilience of digital ecosystems (AlDaajeh *et al.*, 2022).

The fact that a risk-proportionate approach is appropriate means that an internal assistant with a low impact need not have to satisfy the same level of assurance as an AI system which affects eligibility, security operations or essential services. Yet proportionality does not allow the requirements to be informal. Each system must have a clearly defined owner, a formally approved purpose, a record of the sources of its data, baseline security controls and criteria for monitoring. The additional requirements should then increase in accordance with the potential for harm, opacity, autonomy, the sensitivity of the

data, exposure to hostile input and dependence on third parties. Another factor that must be considered is organisational diversity. Governments have to consider the need for public accountability, the consistency of administrative treatment and the length of their service periods. Businesses, by contrast, may place greater importance on competitive speed, customer experience and commercial confidentiality. In both cases, however, there must be reliable records explaining why the AI system exists, what constraints apply and who has the power to stop or alter it.

3. Aim, Objectives and Review Questions

The aim of this review is to develop an integrated, lifecycle-oriented model of responsible AI governance for digital platforms used by Saudi businesses and the government, a model which supports beneficial AI innovation while taking account of data privacy and cyber resilience. The analysis is based on four objectives: to identify the governance capabilities that transform responsible-AI principles into operational controls; to look at how privacy obligations should affect the data and model lifecycles; to incorporate adversarial and operational security into AI assurance; and to turn the findings thus obtained into a platform governance structure suitable for the various risk levels in the public and enterprise sectors. The relevant review questions and analytical lenses are set out in Table 1. The focus is deliberately both organisational and technical since responsible outcomes are just as dependent on decision-making authority, evidence, and escalation procedures as they are on model performance metrics.

Table 1: Review objectives, research questions and analytic lenses

Objective	Research question	Primary analytic lens
Operationalise responsible-AI principles	RQ1. Which organisational capabilities translate responsible-AI principles into testable controls?	Decision rights, lifecycle gates, audit evidence
Integrate privacy into the AI lifecycle	RQ2. How should privacy shape data acquisition, model development, logging and downstream reuse?	Data provenance, minimisation, privacy-enhancing methods
Embed cyber resilience in assurance	RQ3. How can adversarial and operational security be integrated with model governance?	Threat modelling, secure release, monitoring, incident response
Contextualise governance for Saudi platforms	RQ4. How can a common governance architecture accommodate public- and enterprise-sector differences?	Risk proportionality, institutional accountability, third-party governance

4. REVIEW METHODOLOGY

An integrative review based on a structured approach was chosen since the research question involves information systems, AI ethics, privacy engineering, machine learning assurance and cybersecurity. As these areas make use of different constructs and evaluation techniques, statistical

aggregation is not suitable. Instead, integrative synthesis is more appropriate for identifying the relationships between governance mechanisms and for transforming them into an operational architecture. The review was limited to scholarly work published between 2020 and 2025 in order that the synthesis might reflect the current shift from

ethical principles to lifecycle governance, privacy-preserving computation and risks associated with adversarial AI.

The search for literature employed combinations of terms relating to responsible AI, AI governance, algorithmic auditing, trustworthy AI, explainability, fairness, dataset documentation, privacy-preserving machine learning, federated learning, adversarial machine learning, AI cybersecurity, public-sector AI governance, Saudi Arabia and the wider Middle East. In order to find foundational and scholarship relevant to the region, citation chaining was applied to the reference papers, and the bibliographic information and DOI records were then verified using publisher or scholarly metadata. A preference was placed on peer-reviewed journal and conference publications that had a direct connection with governance mechanisms, data stewardship, system assurance, organisational accountability or AI-specific security. Publications that were entirely technical and lacked any governance implications, non-scholarly commentary, sources outside the date range, and items without a verifiable DOI were excluded from the analytic corpus.

The resulting corpus consists of 30 sources which have been verified by means of their DOIs. The sources were selected on the basis of a deliberate purpose rather than with the aim of estimating the prevalence of AI topics throughout the entire field of AI literature. Each source was evaluated along five dimensions: governance level (whether it referred to a principle, an organisational process, lifecycle control or a technical mechanism); risk area (fairness/accountability, privacy, security/resilience or cross-cutting); lifecycle stage; the type of assurance evidence provided; and its applicability to public or enterprise platforms. In addition, the sources focusing on Saudi Arabia and the MENA region were coded with respect to their institutional context and the regional governance implications.

The process took place in three stages. In the first stage, recurring governance mechanisms were identified, these comprising impact assessment, dataset documentation, explainability, human oversight, audit, privacy-enhancing computation, adversarial testing, monitoring and incident response. In the second stage, the mechanisms were linked to various points in the AI lifecycle in order to highlight the gaps that arise when controls are seen as one-off checks carried out before deployment. In the third stage, the mechanisms were incorporated into a governance architecture that connects decision-making authority with technical evidence. The quality assessment concentrated on conceptual clarity, methodological transparency, relevance to

operational governance and the ability to trace claims. For the purpose of establishing general frameworks, review and framework papers were used, while technical studies provided the information needed to define the specific control mechanisms. Special care was taken to avoid equating explainability with accountability or treating privacy technology as a full legal solution. The review also makes a distinction between model-level security and broader platform resilience. This multi-layered approach to appraisal is in line with the growing view in the literature that trustworthy AI results from the interaction of organisational and technical safeguards rather than from any single measure (Chamola *et al.*, 2023; Batool *et al.*, 2025).

5. Thematic Synthesis

5.1 From principles to governable organisational capability

In the field of responsible AI, it is consistently found that principles only gain meaning when they affect the way decisions are made. Although fairness, transparency and accountability are often supported, the use of general terms fails to clarify who is responsible, what kind of evidence is adequate and what takes place in the event of a conflict between values. Hagendorff (2020) points out this weakness in ethical guidelines, while Morley *et al.* (2020) develop practical tools in relation to the various stages of the machine-learning pipeline. Raji *et al.* (2020) take this a step further by viewing internal algorithmic auditing as an end-to-end organisational process which produces documentation during development. The above studies suggest that the main unit of governance should be the accountable decision, not the principle statement.

The fact that lifecycle harm frameworks reach this conclusion is clear. Suresh and Gutttag (2021) have shown that harm can arise at any stage of the process, including during the formulation of the problem, data collection, model development, evaluation and deployment, which means that a final validation test will not be able to correct all the errors that occurred earlier on. Dataset documentation helps to deal with this issue by making the provenance, composition, intended use and any assumptions regarding maintenance visible to those who use the dataset afterwards (Geburu *et al.*, 2021). Likewise, research into fairness proves that bias has many origins and cannot be reduced to a single mathematical criterion (Mehrabi *et al.*, 2021). These findings support the idea that platform governance should include clear checkpoints at which owners must justify the purpose, the data, the performance thresholds and the residual risk.

Auditing should also be continuous. Mökander and Floridi (2021) claim that ethics-based

auditing is most effective when it is an ongoing, system-wide process rather than a one-off certification event. Papagiannidis et al. (2025) integrate responsible-AI governance into structural, relational and procedural practices, pointing out that governance relies on formal arrangements, the relationships with stakeholders and repeatable processes. Camilleri (2024) makes a similar connection between AI governance and an organisation's social responsibility as well as stakeholder expectations. Taken together, then, this body of literature indicates that a Saudi organisation should regard responsible AI as a management capability: policies should set out the minimum expectations, committees or accountable executives should be given the authority, lifecycle procedures should generate evidence, and monitoring should check whether the earlier assumptions are still valid.

Human oversight is thus a control function and not just a slogan; it only proves effective when an individual has access to useful information, has sufficient time, possesses the ability to question the output and has the authority to take action. Explainability can help to achieve this function, but it must be appropriate to the particular decision context. Barredo Arrieta et al. (2020) identify various types of explanation and stress the importance of considering who the explanation is intended for. Rudin et al. (2022) warn that in certain high-stakes situations interpretable models should be chosen rather than automatically relying on post-hoc explanations of opaque systems. The question then becomes whether the explanation allows for a specific action to be taken, such as approval, review, contestation, debugging or investigation of an incident.

5.2 Privacy as lifecycle data governance

The privacy risk associated with AI arises long before the inference stage. The training data might include identifiers, behavioural traces or commercially sensitive records; the labels could disclose protected information; feature engineering might result in new inferences; and the logs produced after deployment can become an additional repository for personal data. For this reason, responsible governance must ensure traceability from the original data through to the derived datasets, models, prompts, embeddings, logs and the subsequent outputs. Kaissis et al. (2020) show that secure and federated learning can reduce exposure of centralised data while still allowing collaborative model development. Kairouz et al. (2021) similarly indicate that federated learning may change where the computation takes place but does not remove the privacy, security or systems issues.

Technical privacy mechanisms ought therefore to be seen as part of a broader decision-making process. Radanliev et al. (2024) examine various methods such as differential privacy, homomorphic encryption and federated learning, stressing that privacy protection demands both algorithmic and governance measures. Alsaigh et al. (2024) incorporate privacy and data sharing into the area of AI governance and advocate for harmonisation rather than making a binary decision between access and protection. The practical result for Saudi digital platforms is that data purpose, sensitivity, retention, transfer, access and transformation should be made explicit at every stage of its lifecycle. A privacy review carried out only at the start of a project is not enough since activities such as fine-tuning, retrieval augmentation, new integrations or monitoring data could materially change the information boundary.

Privacy also has a bearing on explainability and accountability. Although detailed explanations might reveal sensitive attributes or training information, thorough audit logs can enhance traceability at the cost of increasing the amount of personal data that is stored. Governance therefore needs to define the minimum amount of evidence required for accountability and ensure that this evidence is protected in accordance with its sensitivity. This issue is a design challenge, not a justification for giving up either of the two objectives. Data minimisation can be combined with the structured retention of evidence by keeping metadata that is relevant to decisions, version identifiers, approvals and test results while at the same time avoiding unnecessary duplication of the raw personal data.

Another worry is reliance on third parties. Both enterprise and government platforms are now using external models, externally hosted application programming interfaces and pre-trained components. Governance should therefore include requirements for vendor due diligence, notification of contractual changes, restrictions on data use, obligations relating to security, and an internal record of the approved uses of the external component. Since the internal details of the model are not available, the need for scenario testing and monitoring at the point of deployment actually increases.

5.3 Cyber resilience as an AI governance requirement

The security features introduced by AI systems go beyond what traditional application controls are able to cover. Research into adversarial machine learning has shown that there are attacks such as poisoning, evasion, and privacy breaches

which focus on the statistical behaviour of the models and not just on software vulnerabilities (Xi, 2020). Reviews of existing intrusion and malware situations reveal that machine-learning-based defences can themselves be influenced by adversarial inputs, leading to a recursive security issue in which the defensive model becomes part of the attack surface (Martins *et al.*, 2020). Moreover, large language models bring to light new kinds of misuse, information hazards, and risks related to human-computer interaction that go beyond the standard categories of confidentiality, integrity and availability (Weidinger *et al.*, 2022).

Because of these risks, resilience has to be considered a property that spans the entire lifecycle. In order to achieve this, security teams need to have visibility regarding the model's provenance, the inputs used during training and fine-tuning, the interfaces that are exposed, any actions carried out with privileged access, the external tools employed, and the monitoring thresholds. Although pre-deployment red-team exercises are helpful, they cannot predict every possible shift in distribution or malicious strategy. The production telemetry therefore has to be able to identify abnormal input patterns, performance degradation, attempts to extract the model, manipulation of the prompt, unusual tool calls, and other signs that are relevant to the architecture. Governance then needs to specify which individuals are authorised to isolate a model, roll back a version, disable an integration, or trigger a manual process when those indicators go beyond the established tolerance levels.

This is particularly crucial in the case of integrated platforms, since an AI component might have access to other systems. The severity of the possible consequences of a model error grows when the output can cause payments to be made, alter records, send out communications or invoke privileged services. Measures such as applying the principle of least privilege, using segregated credentials, setting rate limits, introducing approval procedures and incorporating fail-safe mechanisms can help reduce the extent of this impact. In this respect, responsible AI and cybersecurity come together in that both examine whether the system will still remain within authorised behaviour when faced with foreseeable stress. Trustworthy-AI evaluations that incorporate robustness and security together with fairness and explainability form the basis for this integrated perspective (Chamola *et al.*, 2023).

The fact that Saudi Arabia is undergoing digital transformation also highlights the importance of institutional cyber capabilities. As AlDaajeh *et al.* (2022) demonstrate, national cybersecurity

strategies can enhance cybersecurity education and capability, showing that resilience depends not only on technical tools but also on people and institutions. Accordingly, at the organisational level, responsible AI governance should link together the owners of the models, the privacy function, cybersecurity operations, procurement and business leadership. An incident in which a model's behaviour is manipulated can at the same time be a security event, a privacy event and a service-quality event. Having separate escalation channels can result in the combined risk being recognised more slowly.

5.4 Risk-proportionate assurance and accountability

The consensus in the relevant literature is that governance ought to be proportional to the impact but should maintain a consistent structure. Taeihagh (2021) refers to AI governance as a problem involving multiple actors and emphasizes the need for adaptive policy tools, whereas de Almeida *et al.* (2021) suggest a governance framework that connects ethical issues with institutional arrangements. Djeflal *et al.* (2022) demonstrate that different states adopt varying combinations of entrepreneurial, regulatory and responsibility-based approaches. Wirtz *et al.* (2022) combine risk categories with governance guidelines, thus underlining the advantages of risk-based controls over a uniform bureaucracy. This is relevant to Saudi organisations in that it justifies tiering use cases according to consequence, autonomy, sensitivity, reach, reversibility and exposure to hostile conditions.

The level of evidence needed should be determined by the risk tier rather than by whether evidence exists. All AI systems must have a purpose statement, an owner, a data record, a security baseline, performance criteria and a monitoring plan; more highly risky systems should also include independent challenge, evaluation at the subgroup level, stronger interpretability or a justifiable reason for opacity, adversarial testing, a human override mechanism, resilience exercises and senior approval (Hacker, 2023). This ensures that low-risk experimentation can proceed quickly while avoiding a prototype becoming a high-impact production service without being noticed.

Accountability also requires versioned evidence. Models, datasets, prompts, policies and external services are all subject to change. Approving one version does not mean that approval of version three is automatically justified. The governance record should connect each release to the data version, the model configuration, the evaluation results, the known limitations, the approvals and the monitoring thresholds that were responsible for the

deployment. If a significant change takes place, the organisation can decide which of the assurances are still valid and which have to be repeated. The situation in Saudi Arabia and the wider MENA region gives this approach additional contextual relevance. Trigui et al. (2024) point out the uneven level of governance maturity in the region and stress the importance of balancing innovation with ethical and regulatory concerns. Ibrahim (2024) shows that the use of AI in Saudi governance gives rise to both institutional and questions of legitimacy as well as technical issues. These views warn against treating the framework as a fixed compliance checklist. Governance has to be locally accountable, understandable to the relevant stakeholders and able to keep up with developments in national institutions, sector rules and technology.

6. Integrated Responsible AI Governance Framework

The synthesis presents a six-gate governance structure as illustrated in Figure 1 and implements it as a continuous assurance loop, as shown in Figure 2. The first gate, which is the use-case intake, defines the business or public purpose, identifies the stakeholders affected, specifies the prohibited uses, designates the risk level and names the accountable owner. A project should not proceed to model development just because data and technology are available; the owner had to explain which decision is being improved and why it is appropriate to use AI for that purpose.

With regard to Gate 2, issues such as data stewardship, records provenance, lawful organisational authority, sensitivity, quality, representativeness, retention and transfer conditions, the concepts of dataset documentation are useful since they require teams to state the intended uses and known limitations rather than regard the data as neutral input (Geburu *et al.*, 2021). Instead of being used as generic labels, privacy-enhancing methods can then be chosen on the basis of the threat model and the collaboration pattern.

The assessment of the model at Gate 3 looks at performance, robustness, fairness, explainability and failure modes in relation to the risk level. This gate must examine significant operating conditions, not just average accuracy. In the case of decisions with high consequences, the requirements should involve subgroup analysis, review by independent assessors and a clear justification for the model's complexity. Evidence from risk taxonomies and from research into interpretable AI are complementary in

that the former expands the range of harms that are taken into account and the latter establishes whether the decision logic can be understood adequately for the purpose of oversight (Mehrabi *et al.*, 2021; Rudin *et al.*, 2022).

The model is viewed as a component of the production attack surface when considering Gate 4 and secure release. The controls available are those relating to least privilege, interface hardening, verification of dependencies, management of secrets, testing for abuse, the ability to roll back, and a well-defined set of automated actions. Adversarial testing should be based on realistic threat scenarios not on isolated benchmark attacks. The key question regarding approval is whether the remaining risk to the model and the platform is acceptable for the degree of autonomy intended.

At Gate 5, continuous monitoring involves comparing the actual production behaviour with the assumptions which were used as the basis for the release. The monitoring should include assessments of model quality, detection of drift, fairness indicators where applicable, privacy-related logging, security anomalies, user complaints and operational incidents. When trigger thresholds are reached, there should be clearly defined actions taken such as an investigation, using a human-only fallback, rolling back the model, retraining it or suspending it. Regarding Gate 6, controlled retirement, data access must be closed, the minimum assurance record must be archived, the credentials must be disabled and it must be confirmed how dependent services will move on. Retirement is important since abandoned models, endpoints and datasets may continue to pose security and privacy risks.

At all the gates there are four mechanisms that ensure coherence. The first of these is the assignment of accountable decision-making powers, which specifies who has the authority to approve, challenge, suspend and retire a system. The second involves designing human oversight in such a way that it allows for actual intervention. The third is evidence management, which produces a versioned record of assurance that links up claims with tests and approvals. The fourth consists of third-party governance imposing the same standards on external models and services by means of due diligence, contractual controls and boundary testing. Table 2 converts these mechanisms into an actionable lifecycle control matrix.

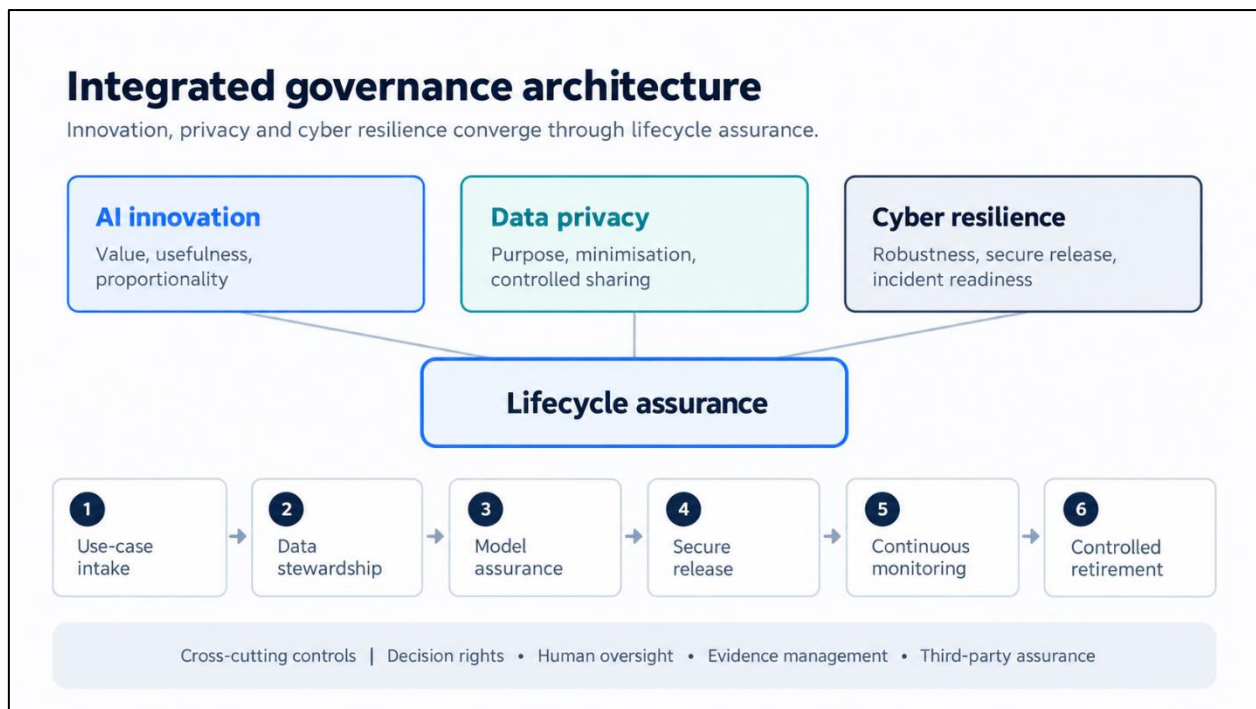


Figure 1: Integrated governance architecture linking innovation, privacy and cyber resilience to six lifecycle gates. Source: authors' synthesis

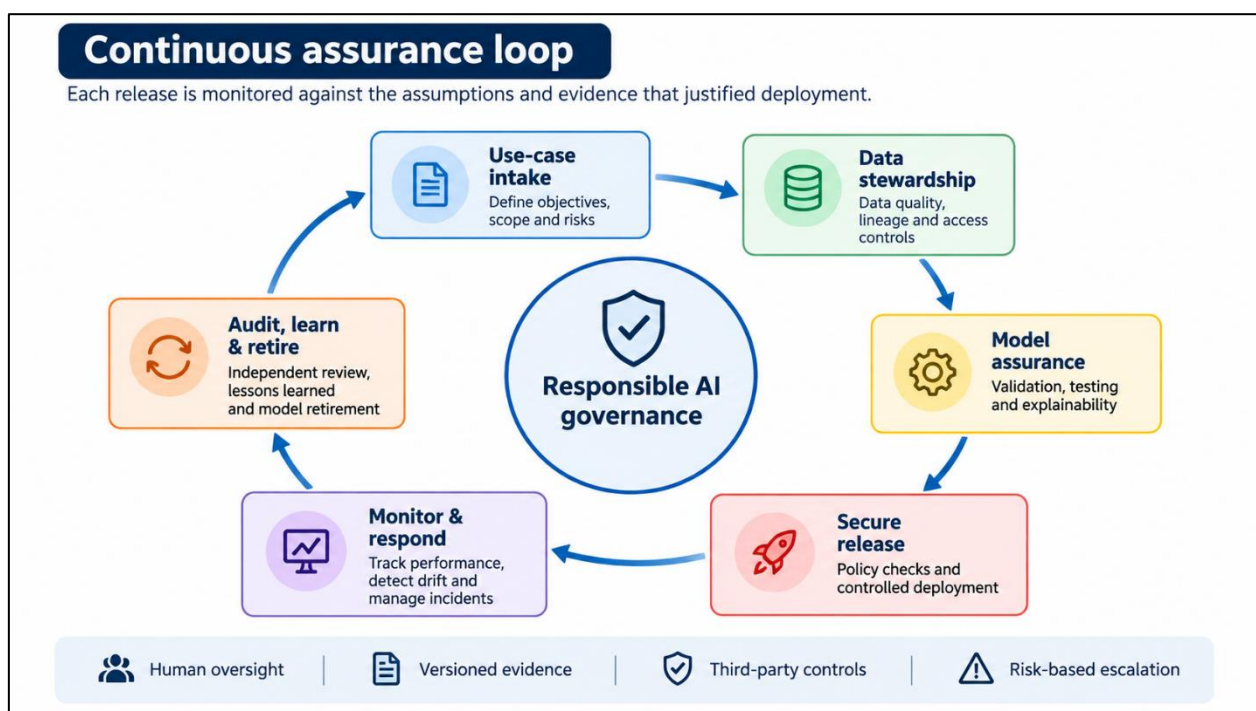


Figure 2: Continuous assurance loop for enterprise and government AI platforms. Source: authors' synthesis

Table 2: Lifecycle control matrix for integrated responsible AI assurance

Gate	Core governance decision	Integrated privacy and resilience controls	Minimum assurance evidence
1. Use-case intake	Is AI necessary, proportionate and owned at the correct risk tier?	Purpose limitation; affected-stakeholder analysis; misuse and dependency scenarios	Use-case charter; accountable owner; risk tier; prohibited uses

Gate	Core governance decision	Integrated privacy and resilience controls	Minimum assurance evidence
2. Data stewardship	Are data sources fit, permitted, representative and controlled?	Provenance; minimisation; access control; retention; secure transfer; privacy-enhancing methods where justified	Data inventory; provenance record; quality and sensitivity assessment
3. Model assurance	Does the model meet performance, fairness, robustness and oversight requirements?	Subgroup testing; interpretability; failure-mode analysis; adversarial evaluation	Evaluation report; limitations; test results; review sign-off
4. Secure release	Can the model operate safely within the production platform and its privileges?	Least privilege; interface hardening; dependency verification; abuse testing; rollback and fail-safe modes	Release approval; threat model; security test record; rollback plan
5. Continuous monitoring	Do production conditions remain within approved assumptions and tolerances?	Drift and quality indicators; security telemetry; privacy-aware logging; complaint and incident signals	Monitoring dashboard; thresholds; incident records; change decisions
6. Controlled retirement	Can the service exit without leaving unmanaged data, credentials or dependencies?	Credential revocation; data disposition; archival limits; dependent-service transition	Retirement approval; access closure; retained assurance record

7. Implications for Enterprise and Government Platforms

With regard to government platforms, the framework places a strong emphasis on public accountability, consistency, contestability and continuity. Where systems have an impact on access to services or on the way administrative decisions are made, there must be clear procedures allowing for human review and records that are sufficient to reconstruct the main decisions. Explanations should be tailored to the kind of stakeholder who has to act on them: an operator requires detailed diagnostic information, an oversight body needs evidence of assurance, and a person affected by the decision might need a brief reason together with a route for review. Assuming that one technical explanation is adequate for all groups of stakeholders risks achieving formal transparency without ensuring practical accountability.

For enterprises, the same architecture promotes innovation by lowering the level of uncertainty when moving from experimentation to production; risk tiering ensures that each prototype does not draw in heavy-duty governance, and the standard intake and evidence requirements make it easier to scale up later on. Procurement is a key area for exerting influence; when external AI is obtained, the contractual obligations relating to data use, notification of updates, security testing, cooperation in the event of an incident, and service exit can in part take the place of the visibility that would otherwise be available through internal development.

In both sectors the key point for central management is that governance functions which are

normally kept separate should be integrated. Privacy teams should be informed about any changes to models and logging; cybersecurity teams should become aware of the data and model provenance; AI owners must know which controls are required before a release; and internal audit should be able to trace the evidence without having to reconstruct the project afterwards. Responsible AI can be considered operationally credible in practice once these different functions share a common lifecycle vocabulary and have a single understanding of system ownership.

The framework also changes the way in which innovation is viewed. Although governance is usually seen as creating obstacles, it can in fact speed up adoption since teams will know in advance what evidence is needed for a release. Privacy-preserving techniques, interpretable models, limited autonomy and a phased approach to deployment can make previously unacceptable applications amenable to governance. The appropriate balance is thus not between innovation and control, but between unmanaged innovation and evidence-based innovation. This distinction is important in the quickly digitising environment of Saudi Arabia since trust and resilience are prerequisites for sustainable scaling rather than being extra costs incurred after value has been created (Vinuesa *et al.*, 2020; Radanliev *et al.*, 2024).

8. Research Gaps, Limitations and Future Work

The review points out a number of research gaps. Firstly, there is a need for more empirical studies concerning how Saudi organisations in fact distribute AI decision-making powers among the areas of technology, law, privacy, cyber security and

business. Secondly, the metrics used for continuous assurance are still in an immature stage since organisations find it easier to measure model accuracy than the effectiveness of human supervision, the ability to challenge decisions or the governance escalation process. Thirdly, when training data and information about the internal models are not available, there is a requirement for more robust methods of ensuring the governance of third-party and foundation models. Fourthly, research should look into how the compromises between privacy, robustness and explainability vary in Arabic-language and culturally specific applications rather than assuming that the results from other contexts remain unchanged.

The review does have some limitations; the corpus it uses consists of 30 sources which has been deliberately selected with a focus on the concept rather than being comprehensive, and only scholarly work that can be verified by DOI numbers published between 2020 and 2025 has been included. Because of the fast pace of regulatory and technical developments, guidance for implementation will keep on changing. Moreover, the framework is normative in that it combines the evidence to propose a control structure, but it has not yet been tested using comparative case studies or assessed in relation to actual incident outcomes. Future research should therefore carry out a pilot study of the six-gate model in government and enterprise environments in Saudi Arabia, compare low-risk and high-risk use cases, and examine whether the quality of the evidence, the time taken for the review, incident detection, and stakeholder confidence improve as the governance cycles progress.

9. CONCLUSION

The concept of responsible AI governance for digital platforms in Saudi Arabia must be seen as a process of continuous assurance covering the entire socio-technical lifecycle. The literature published between 2020 and 2025 makes it clear that principles on their own are not enough since harms can arise from the way problems are formulated and from the data used, privacy controls interact with evidence of accountability, adversaries are able to target model behaviour, and platform dependencies can magnify local failures. Governance that is to be effective therefore goes beyond the establishment of an ethics committee or a simple final model review.

The six-gate design includes use-case authorisation, data stewardship, model assurance, secure release, continuous monitoring and controlled retirement; its advantage is due to the relationships between these gates and the presence of accountable decision-making rights, human oversight, versioned evidence and third-party control. The architecture is

intentionally risk-proportionate in order that efficient progress may be made with regard to low-impact experimentation while systems which have greater consequence, autonomy, sensitivity or are exposed to hostile environments are given more thorough assurance.

For enterprise and government organisations in Saudi Arabia the real aim is not to delay the adoption of AI but to ensure that its scaling can be properly justified. Innovation, privacy and cyber resilience are not separate activities which can be settled at the end of a project; rather they are design constraints which should influence the organisation's purpose, the data used, the architecture, the release process and monitoring from the very start. Only when governance puts forward testable statements, assigns specific owners and makes evidence readily available can responsible AI become an operational capability capable of adjusting to new models, new threats and changing expectations.

REFERENCES

- AlDaajeh, S., Saleous, H., Alrabae, S., Barka, E., Breiting, F., & Choo, K.-K. R. (2022). The role of national cybersecurity strategies on the improvement of cybersecurity education. *Computers & Security*, 119, 102754. <https://doi.org/10.1016/j.cose.2022.102754>
- Alsaigh, R., Mehmood, R., Katib, I., Liang, X., Alshantiti, A., Corchado, J. M., & See, S. (2024). Harmonizing AI governance regulations and neuroinformatics: Perspectives on privacy and data sharing. *Frontiers in Neuroinformatics*, 18, 1472653. <https://doi.org/10.3389/fninf.2024.1472653>
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bannetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Batool, A., Zowghi, D., & Bano, M. (2025). AI governance: A systematic literature review. *AI and Ethics*, 5, 3265–3279. <https://doi.org/10.1007/s43681-024-00653-w>
- Camilleri, M. A. (2024). Artificial intelligence governance: Ethical considerations and implications for social responsibility. *Expert Systems*, 41(7), e13406. <https://doi.org/10.1111/exsy.13406>
- Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., & Sikdar, B. (2023). A review of trustworthy and explainable artificial intelligence (XAI). *IEEE Access*, 11, 78994–79015.

- <https://doi.org/10.1109/ACCESS.2023.3294569>
- de Almeida, P. G. R., dos Santos, C. D., & Farias, J. S. (2021). Artificial intelligence regulation: A framework for governance. *Ethics and Information Technology*, 23, 505–525. <https://doi.org/10.1007/s10676-021-09593-z>
 - Djeflal, C., Siewert, M. B., & Wurster, S. (2022). Role of the state and responsibility in governing artificial intelligence: A comparative analysis of AI strategies. *Journal of European Public Policy*, 29(11), 1799–1821. <https://doi.org/10.1080/13501763.2022.2094987>
 - Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>
 - Hacker, P. (2023). The European AI liability directives – Critique of a half-hearted approach and lessons for the future. *Computer Law & Security Review*, 51, 105871. <https://doi.org/10.1016/j.clsr.2023.105871>
 - Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30, 99–120. <https://doi.org/10.1007/s11023-020-09517-8>
 - Ibrahim, N. M. H. (2024). Artificial Intelligence (AI) and Saudi Arabia's governance. *Journal of Developing Societies*, 40(4), 500–530. <https://doi.org/10.1177/0169796X241288590>
 - Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., ... Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2), 1–210. <https://doi.org/10.1561/22000000083>
 - Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2, 305–311. <https://doi.org/10.1038/s42256-020-0186-1>
 - Martins, N., Cruz, J. M., Cruz, T., & Abreu, P. H. (2020). Adversarial machine learning applied to intrusion and malware scenarios: A systematic review. *IEEE Access*, 8, 35403–35419. <https://doi.org/10.1109/ACCESS.2020.2974752>
 - Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), Article 115, 1–35. <https://doi.org/10.1145/3457607>
 - Mökander, J., & Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. *Minds and Machines*, 31, 323–327. <https://doi.org/10.1007/s11023-021-09557-8>
 - Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, 26, 2141–2168. <https://doi.org/10.1007/s11948-019-00165-5>
 - Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, 34(2), 101885. <https://doi.org/10.1016/j.jsis.2024.101885>
 - Radanliev, P., Santos, O., Brandon-Jones, A., & Joinson, A. (2024). Ethics and responsible AI deployment. *Frontiers in Artificial Intelligence*, 7, 1377011. <https://doi.org/10.3389/frai.2024.1377011>
 - Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. <https://doi.org/10.1145/3351095.3372873>
 - Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16, 1–85. <https://doi.org/10.1214/21-SS133>
 - Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, 26, 2749–2767. <https://doi.org/10.1007/s11948-020-00228-y>
 - Suresh, H., & Gutttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, Article 17, 1–9. <https://doi.org/10.1145/3465416.3483305>
 - Taeihagh, A. (2021). Governance of artificial intelligence. *Policy and Society*, 40(2), 137–157. <https://doi.org/10.1080/14494035.2021.1928377>
 - Trigui, H., Guerfali, F., Harigua-Souiai, E., Qasrawi, R., Atri, C., Sokhn, E. S., El Morr, C., Hammami, K., Souiai, O., & Wu, J. (2024). Exploring AI governance in the Middle East and North Africa (MENA) region: Gaps, efforts, and initiatives. *Data & Policy*, 6, e83. <https://doi.org/10.1017/dap.2024.85>
 - Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The

role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, 11, 233. <https://doi.org/10.1038/s41467-019-14108-y>

- Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., ... Gabriel, I. (2022). Taxonomy of risks posed by language models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229. <https://doi.org/10.1145/3531146.3533088>
- Wirtz, B. W., Weyerer, J. C., & Kehl, I. (2022). Governance of artificial intelligence: A risk and guideline-based integrative framework. *Government Information Quarterly*, 39(4), 101685. <https://doi.org/10.1016/j.giq.2022.101685>
- Xi, B. (2020). Adversarial machine learning for cybersecurity and computer vision: Current developments and challenges. *WIREs Computational Statistics*, 12(5), e1511. <https://doi.org/10.1002/wics.1511>