How to cite this article:
Masoumi Godarzi, M., & Poyanrad, A. (2027). Responsible AI Governance in Iran: A Program-Theory Evaluation of the National Artificial Intelligence Strategy. *Journal of Historical Research, Law and Policy, 5*(5), 1-22. https://doi.org/10.61838/jhrlp.389

Article history:
Original Research

Dates:
Submission Date: 06 April 2026
Revision Date: 04 July 2026
Acceptance Date: 13 July 2026
First Publication Date: 15 July 2026
Final Pubication Date: 01 September 2027

Responsible AI Governance in Iran: A Program-Theory Evaluation of the National Artificial Intelligence Strategy

1. Mohammad. Masoumi Godarzi [1*] : MS Degree in Science, Technology and Innovation Policy, Faculty of Governance, University of Tehran, Tehran, Iran
2. Amin. Poyanrad [2]: MS Degree in Media Management, Faculty of Communications, IRIB University, Tehran, Iran

*corresponding author's email: Mohammad.masoumi@ut.ac.ir

ABSTRACT

Artificial intelligence has become a strategic technology whose development requires coordinated institutional, regulatory, ethical, and infrastructural governance. This study evaluated the capacity of Iran's National Artificial Intelligence Strategy to support responsible AI governance by reconstructing its underlying program theory and examining the coherence of its implementation architecture. The research adopted a qualitative, theory-based case-study design. Data were obtained through documentary analysis of the National Artificial Intelligence Strategy and related policy materials, together with semi-structured interviews with 12 experts from academic, research, policy, and industrial backgrounds. The data were coded and thematically analysed using MAXQDA, after which an integrated theory of change and theory of action was developed and applied to the national strategy. The findings indicated that responsible AI governance in Iran depends on sustained political commitment, a legally empowered and agile coordinating institution, equitable access to high-quality data and computational infrastructure, interdisciplinary human capital, effective university–industry collaboration, risk-based regulation, and continuous policy evaluation. The strategy provides an important normative foundation by emphasizing human dignity, national values, scientific development, priority sectors, and the participation of knowledge-based firms. Nevertheless, its implementation architecture remains insufficiently specified. Major weaknesses include ambiguity regarding institutional authority, sustainable financing, data classification, access to computational resources, legal liability, regulatory instruments, measurable indicators, and mechanisms for periodic revision. The analysis also identified major policy dualities involving open versus sensitive data, centralized versus networked governance, strict versus adaptive regulation, domestic versus international standards, state versus private investment, and problem-oriented versus technology-oriented development. The study concludes that the effectiveness of the strategy depends on translating broad objectives into an explicit, funded, accountable, and adaptive program of action. Responsible AI governance requires not merely ethical declarations, but a coherent causal chain connecting institutions, resources, activities, outputs, societal outcomes, and mechanisms of accountability.

Keywords: *Responsible AI governance; National Artificial Intelligence Strategy; program theory; theory of change; theory of action; AI policy; Iran; policy evaluation*

Introduction

Artificial intelligence has moved from a specialized field of computer science to a general-purpose technology capable of reshaping production, public administration, healthcare, education, communications, security, culture,

and everyday social interaction. Its transformative capacity has frequently been compared with that of electricity because its influence is not confined to a single industry but extends across economic and institutional systems (1). This breadth distinguishes artificial intelligence from many earlier technological innovations. AI systems can automate routine tasks, support complex decisions, generate text and visual content, identify patterns in large datasets, optimize industrial processes, and expand the analytical capacity of public institutions. At the same time, these systems can reproduce discrimination, obscure responsibility, intensify surveillance, concentrate informational power, displace forms of employment, and amplify false or manipulated content. The development of AI therefore presents governments with a dual imperative: they must create the conditions for innovation and technological capacity while ensuring that development remains consistent with public values, social security, human dignity, and long-term national interests. This dual imperative has transformed artificial intelligence into a central issue of contemporary governance rather than merely a matter of scientific or industrial policy. The international proliferation of national AI strategies demonstrates that states increasingly regard AI capability as an element of economic competitiveness, administrative effectiveness, geopolitical influence, and strategic autonomy (2). Nevertheless, the existence of a national strategy does not itself guarantee effective governance. The decisive question is whether the strategy establishes a plausible and operational relationship among its normative principles, institutional responsibilities, available resources, proposed actions, measurable outputs, and intended societal effects.

The governance problem is intensified by the distinctive characteristics of AI. Conventional regulation usually responds to relatively stable technologies, identifiable producers, and observable harms, whereas AI systems change through data, interaction, adaptation, retraining, and deployment in different organizational contexts. Generative AI has made these challenges particularly visible because responsibility for a harmful output may be distributed among data providers, model developers, application designers, deploying organizations, professional users, and end users. A responsible-governance framework must therefore determine not only whether an outcome is undesirable, but also which actor possessed the knowledge, authority, and practical capacity to prevent or correct it (3). Furthermore, AI-related harms may emerge gradually rather than through a single identifiable incident. Persistent bias in administrative decision-making, unequal access to computational infrastructure, erosion of informational trust, or dependence on foreign technological platforms may become visible only after they have affected institutions and social practices. The adaptive nature of AI consequently requires governance arrangements capable of continuous learning, monitoring, and revision rather than one-time legal intervention. Conceptualizing advanced AI as part of a complex adaptive system highlights the interaction among technical models, organizational routines, human behaviour, market incentives, regulation, and political institutions (4). Responsible governance must thus operate across this entire system instead of focusing exclusively on the technical properties of algorithms.

The emergence of AI ethics guidelines represented an important response to these challenges. International frameworks have repeatedly emphasized principles such as transparency, fairness, accountability, privacy, non-maleficence, human control, and respect for autonomy (5). However, the growth of ethical principles has also exposed the distance between normative agreement and institutional implementation. Organizations may endorse transparency without specifying what information must be disclosed, to whom it should be disclosed, or how disclosure should be verified. They may support accountability without identifying the accountable actor or the consequences of failing to fulfil an obligation. Similarly, a commitment to fairness may remain symbolic when no

procedure exists for testing discriminatory outcomes, receiving complaints, correcting models, or compensating affected individuals. AI governance must therefore move beyond the declaratory level of ethics and translate principles into decision rights, organizational responsibilities, regulatory standards, technical procedures, audit mechanisms, and forms of remedy. The central challenge is not the absence of desirable values but the absence of operational pathways through which those values become routine characteristics of AI development and deployment. Reviews of responsible AI governance have accordingly emphasized that ethics, organizational governance, regulation, stakeholder participation, and technical controls must be treated as interconnected components (6). A national strategy is responsible only when its values are embedded in the institutions and processes through which technological development is funded, authorized, monitored, and corrected.

The need for this institutional translation is particularly significant in countries confronting constraints in finance, infrastructure, international technological access, and administrative coordination. The global AI landscape is characterized by substantial asymmetries. A small number of corporations and countries control advanced semiconductor supply chains, large-scale computational facilities, foundational models, cloud platforms, and strategically valuable datasets. AI development therefore intersects with geopolitical competition and can reproduce dependency between technological centres and peripheral markets (7). Countries with restricted access to advanced processors or international services must decide whether to pursue costly national infrastructure, depend on external platforms, concentrate on efficient and smaller models, or prioritize a limited number of strategic applications. These choices are not purely technical. They influence national autonomy, market structure, access to knowledge, cybersecurity, and the distribution of public resources. The regulatory landscape is also fragmented and rapidly changing, requiring national policymakers to balance compatibility with international standards against local legal, cultural, and developmental priorities (8). A strategy designed without attention to these external conditions may establish unrealistic targets, while a strategy focused exclusively on constraints may suppress experimentation and innovation. Responsible governance consequently requires a realistic assessment of national capacities combined with mechanisms for adaptation as technological and international conditions change.

Iran's National Artificial Intelligence Strategy emerged within this broader environment of technological opportunity, geopolitical uncertainty, and domestic institutional fragmentation. The strategy seeks to articulate national values, establish a long-term vision, define macro-level objectives, identify priority applications, promote scientific and technological capacity, and create mechanisms for implementation and supervision. Its adoption represents an important recognition that AI development cannot be left to isolated universities, ministries, firms, or research centres. The document attempts to provide a common direction for a field involving data governance, digital infrastructure, education, industrial development, public services, ethics, regulation, culture, and security. Previous analyses of Iran's AI ecosystem have pointed to the importance of institutional mapping and a clear national division of labour because overlapping mandates can lead to duplication, conflict, and ineffective allocation of resources (9). Studies of the country's prospective regulatory system have similarly emphasized the need to determine the position, jurisdiction, and instruments of AI governance institutions (10). These concerns are especially relevant because the effectiveness of a national strategy depends less on the number of objectives it contains than on whether organizations possess the authority, funding, expertise, and incentives necessary to implement those objectives.

The National AI Strategy contains elements associated with responsible governance, including attention to human dignity, cultural values, privacy, scientific advancement, national priorities, public welfare, and participation

by universities and knowledge-based firms. Yet the presence of these components does not resolve questions about the strategy's internal logic. It remains necessary to determine how broad objectives are expected to produce concrete changes, which intermediate outcomes must be achieved, which institution is responsible for each stage, and which assumptions underlie the expected sequence of change. The strategy may, for example, call for the development of national AI capability without clarifying the respective responsibilities of government and private industry in financing computational infrastructure. It may promote data accessibility while leaving unresolved the classification of sensitive and non-sensitive public data. It may recognize the importance of responsibility without allocating liability among developers and deploying organizations. It may establish national rankings or numerical objectives without identifying composite measures capable of evaluating responsible development. Such gaps are not minor drafting problems; they constitute uncertainties in the program logic of the strategy. Evaluation must therefore investigate whether the document provides a coherent explanation of how its interventions are supposed to generate responsible technological and social outcomes.

Program theory offers an appropriate analytical approach to this problem because it reconstructs the reasoning that connects a policy intervention to its anticipated consequences. It distinguishes between a theory of change, which explains why a sequence of outcomes is expected to occur, and a theory of action, which explains how institutions, resources, and activities will produce that sequence. This distinction is particularly useful for AI governance. A strategy may possess an attractive theory of change, such as the proposition that investment in research, data, and human capital will improve national AI capability, but lack a credible theory of action specifying who will provide the investment, how resources will be distributed, which organizations will coordinate implementation, and how progress will be monitored. Conversely, a strategy may contain numerous activities without explaining why they are expected to produce responsible development. Theory-based evaluation makes it possible to distinguish a failure of implementation from a failure in policy design. If appropriate activities were not performed, the principal problem concerns implementation; if activities were performed but could not plausibly generate the intended outcomes, the underlying program theory is deficient. Evaluating this distinction is more informative than determining whether isolated indicators have been achieved.

Responsible AI governance also requires explicit consideration of competing policy directions. AI strategies are shaped by choices between open data and the protection of sensitive information, flexible regulation and strict precaution, national values and international norms, centralized authority and distributed participation, government investment and private initiative, technological capacity-building and immediate problem-solving, or self-regulation and independent oversight. These alternatives cannot be resolved through general support for all desirable objectives. Each choice creates benefits, costs, and implementation consequences. Agile governance has been proposed as a response to technologies that change more rapidly than traditional legislative processes, but agility must not become a justification for weak accountability (11). Similarly, collaborative regulatory models can encourage participation and learning, but their effectiveness depends on clear authority and enforceable boundaries (12). A national strategy must therefore identify where it stands on major policy dualities and explain how it will manage unavoidable tensions.

The present study evaluates the capacity of Iran's National Artificial Intelligence Strategy to support responsible AI governance by reconstructing its program theory and comparing that theory with the institutional and operational content of the strategy. The study addresses three questions. First, what theory of change and theory of action are required for responsible AI governance in Iran? Second, to what extent does the National AI Strategy contain the

institutions, resources, regulatory mechanisms, and evaluation arrangements necessary to realize that program theory? Third, how does the strategy position itself across the principal policy dualities involved in AI governance, and what implementation gaps result from unclear or incomplete choices? The study contributes to the literature by applying theory-based evaluation to a national AI strategy, integrating documentary evidence with expert interpretations from academia, policy research, and industry, and identifying context-sensitive requirements for responsible governance in a resource-constrained environment. Rather than assessing the strategy solely by the principles it declares, the analysis examines whether those principles are connected to a credible chain of institutional action and measurable change.

Conceptual and Analytical Framework

AI governance can be understood as the system through which authority over the development, deployment, use, and consequences of artificial intelligence is distributed and exercised. It encompasses formal law, administrative regulation, organizational policies, industry standards, technical design requirements, professional norms, market incentives, public participation, and mechanisms of accountability. This definition is broader than AI ethics, which primarily specifies the values that should guide behaviour, and broader than AI regulation, which primarily establishes mandatory rules. Governance concerns the total arrangement through which decisions are made, implemented, contested, reviewed, and revised. Systematic reviews have shown that the field includes multiple levels of governance, ranging from model-level controls and organizational processes to national institutions and transnational coordination (13). This multilevel character is essential because an AI system may be technically reliable while being used irresponsibly by an organization, or an organization may adopt strong internal procedures while operating in a national environment without effective legal remedies. Responsible outcomes depend on alignment across these levels.

Responsible AI governance is the form of governance that directs AI development toward legitimate public purposes while preventing, detecting, and remedying unacceptable harm. Responsibility in this context has prospective and retrospective dimensions. Prospective responsibility concerns the duty to anticipate risks, design safeguards, allocate resources, and establish procedures before harm occurs. Retrospective responsibility concerns the capacity to identify accountable actors, investigate failures, impose corrective measures, and provide remedies after adverse effects occur. The prospective dimension is especially important for technologies with potentially systemic consequences because waiting for visible harm may make correction costly or impossible. Ethical and legal analyses of AI have therefore emphasized the need to coordinate technical opportunity with anticipatory governance (14). Yet anticipatory governance must remain proportionate and evidence-based. Regulation that assumes every AI application presents the same level of risk may unnecessarily restrict low-risk innovation, while permissive regulation may expose individuals and institutions to preventable harm. Responsibility thus requires differentiated governance based on context, purpose, affected rights, scale, reversibility, and the degree of human oversight.

A central distinction in responsible governance is that between principles and mechanisms. Principles express the direction of legitimate action, whereas mechanisms convert that direction into observable practice. Transparency becomes operational through documentation, traceability, disclosure requirements, accessible explanations, model and data records, and independent access to relevant information. Fairness becomes operational through representative data practices, impact assessment, performance testing across social groups,

complaint channels, and procedures for correcting discriminatory outcomes. Accountability becomes operational through assigned roles, reporting duties, audit rights, sanctions, and legal remedies. Privacy becomes operational through data minimization, access controls, anonymization, security standards, and restrictions on secondary use. Human control becomes operational through decision thresholds, rights of review, override procedures, and limits on autonomous action. Cross-industry studies demonstrate that organizations require governance mechanisms adapted to the risk profile and institutional setting of AI use rather than reliance on abstract principles alone (15). Responsible governance is therefore best evaluated by asking whether such mechanisms exist, whether they are assigned to capable institutions, and whether compliance can be verified.

The institutional dimension of AI governance concerns the configuration of authority. National systems must determine which body formulates strategy, which body regulates risks, which institutions finance infrastructure, which organizations produce or maintain data, and which actors implement applications. These functions may be combined, but combining them can create conflicts of interest. An institution responsible for promoting rapid adoption may be reluctant to impose restrictions on projects it sponsors, while an institution focused exclusively on risk may lack sufficient understanding of innovation processes. Effective governance therefore requires both coordination and functional differentiation. Governance-system research has stressed the importance of defined roles, vertical coordination across levels of government, and horizontal coordination across policy sectors (16). In the Iranian context, this requirement is intensified by the existence of multiple councils, ministries, research organizations, regulatory bodies, universities, and public enterprises with partially overlapping mandates. A national coordinating institution must possess sufficient authority to establish direction and resolve conflicts, but its authority should not eliminate specialized regulatory or sectoral expertise.

Institutional effectiveness also depends on organizational agility. AI technology develops through short innovation cycles, whereas governmental decision-making is often characterized by sequential approval, formal hierarchy, and lengthy legislation. A rigid institution may therefore govern yesterday's technology rather than emerging applications. Agile governance seeks to address this mismatch through iterative policymaking, experimental regulation, regulatory sandboxes, periodic review, rapid expert consultation, and provisional rules subject to revision. Comparative examinations of AI ethical guidelines have shown that static principles are insufficient when institutions cannot update them in response to new capabilities and harms (17). Agility, however, should be distinguished from institutional instability. Frequent changes in organizational structure, leadership, or legal status can interrupt implementation and weaken accountability. The responsible model is a stable institution with flexible procedures rather than an unstable institution repeatedly reconstituted in response to political change.

Data governance forms a second core dimension. Contemporary AI systems depend on the quality, relevance, accessibility, and lawful use of data. Public institutions possess valuable datasets in health, education, transportation, agriculture, taxation, industry, environment, and social services, but these datasets may be fragmented, poorly standardized, inaccessible, or legally sensitive. Treating all government data as open may endanger privacy, security, or legitimate confidentiality. Treating all data as sensitive may prevent research, innovation, accountability, and evidence-based policy. Responsible data governance therefore requires classification. Data should be differentiated according to sensitivity, public value, identifiability, security implications, permitted purposes, and conditions of access. Access can then be structured through open publication, licensed sharing, secure research environments, controlled APIs, or restricted use. The governing principle is neither unlimited openness nor unrestricted secrecy but proportionate and accountable accessibility. Data governance

must also clarify ownership, stewardship, responsibility for quality, conditions of reuse, and obligations to correct inaccurate information.

Computational infrastructure is inseparable from data governance because access to data has limited value when researchers and firms lack the capacity to train or deploy models. High-performance processors, cloud services, storage, secure networks, and reliable electricity represent strategic inputs. Unequal access to these resources can produce an AI ecosystem dominated by a small number of organizations, even when the formal policy encourages broad participation. Responsible infrastructure governance must therefore consider distributive fairness, strategic prioritization, and economic sustainability. Equal access does not necessarily require identical access for every actor; rather, allocation criteria should be transparent, non-discriminatory, and connected to public value. In a resource-constrained environment, universal provision of advanced computing may be unrealistic. Shared national facilities, sectoral consortia, subsidized access for qualified projects, and investment in efficient models may provide more feasible alternatives. Frameworks for AI governance increasingly acknowledge that ethical outcomes are shaped by the political economy of data and computing rather than only by model design (18).

The human-capital dimension concerns both technical expertise and interdisciplinary competence. Responsible AI cannot be developed by engineers alone because the social significance of a system depends on law, ethics, psychology, management, public policy, medicine, education, and domain-specific knowledge. Interdisciplinary teams are necessary to define appropriate problems, identify affected groups, interpret risks, and design meaningful oversight. National policy must therefore support education that combines technical knowledge with understanding of institutional and social contexts. It must also create incentives for skilled individuals to remain connected to domestic research and industry. Human-capital policy is not limited to increasing the number of graduates; it must improve the relationship between education, applied research, entrepreneurship, public-sector demand, and industrial scaling. Weak university–industry connections can lead to publication without application, while a purely commercial model can neglect long-term research and public-interest problems. Institutional mapping of Iran's AI ecosystem has highlighted the necessity of coordinating these actors through an explicit division of labour (9).

Human-centred governance provides the normative orientation for these institutional and infrastructural arrangements. A human-centred approach treats AI as an instrument for expanding human capabilities and improving collective welfare rather than as an end in itself. It requires preservation of dignity, autonomy, privacy, safety, and meaningful human judgment. Human-centred governance is especially important in high-impact settings where algorithmic recommendations can affect healthcare, employment, education, access to services, legal status, or personal reputation. The ethical challenges of AI in media further demonstrate that harms extend beyond individual decisions to the integrity of public communication, cultural representation, and collective trust (19). Synthetic media, automated persuasion, deepfakes, and large-scale content generation can alter the conditions under which citizens form beliefs. Responsible governance must therefore consider informational ecosystems as well as discrete technological applications. Research on media and AI governance indicates that platform incentives, strategic behaviour, institutional trust, and public responses interact in ways that cannot be controlled through technical accuracy alone (20).

The ecosystem dimension concerns the relationship among government, established industries, startups, universities, investors, and civil society. Government has a legitimate role in financing public goods such as foundational infrastructure, open datasets, basic research, standards, and education. Private actors are often better positioned to identify market demand, experiment rapidly, and commercialize applications. Responsible ecosystem

governance must combine these capacities while preventing public support from becoming permanent dependency or rent-seeking. Startup support should be connected to performance, knowledge creation, and pathways to market. Large firms can contribute capital and scaling capacity but may also consolidate control over data and infrastructure. Universities can supply research and expertise but require incentives to address practical national problems. Stakeholder-oriented governance recognizes that organizational policies are shaped by the distribution of interests, capabilities, and risks among these groups (21). Participation should therefore be structured rather than symbolic, with clear procedures for consultation, disclosure of conflicts, and incorporation of affected perspectives.

The evaluative dimension ensures that governance remains capable of learning. Indicators must measure more than the volume of publications, number of firms, level of investment, or position in international rankings. Such statistics may reflect technological activity without revealing whether AI systems are safe, equitable, useful, sustainable, or trusted. Responsible-governance indicators should combine capacity measures with institutional, ethical, distributive, and outcome-oriented measures. Examples include the availability of secure datasets, concentration of computational access, number and quality of impact assessments, response time to incidents, sectoral adoption, public-service improvements, documented discriminatory effects, accessibility of complaint mechanisms, and participation of affected groups. Evaluation frameworks should also identify causal contribution. An increase in AI publications cannot automatically be attributed to the national strategy, and an unsuccessful project may result from external restrictions rather than defective policy design. Evaluation must thus be connected to the causal assumptions of the strategy (22).

The present study integrates these dimensions through program theory. The theory of change begins with the problem situation, identifies its causes and consequences, specifies the intended long-term condition, and reconstructs the sequence of intermediate outcomes required to move from one to the other. The theory of action identifies the organizations, resources, activities, instruments, outputs, and implementation conditions required to activate that sequence. The analytical framework evaluates whether the National AI Strategy clearly defines the problem, establishes a plausible causal pathway, assigns responsibilities, identifies resources, links activities to outcomes, converts values into mechanisms, adopts measurable criteria, acknowledges external constraints, resolves major policy choices, and establishes procedures for monitoring and revision. It also treats policy dualities as indicators of strategic clarity. A document that simultaneously endorses open data and extensive security restrictions, private-sector leadership and state-directed development, or centralized authority and broad participation must explain how these orientations will be reconciled. The quality of a national strategy therefore depends not on eliminating all tensions but on governing them explicitly, proportionately, and adaptively.

Methodology

This study employed a qualitative, theory-based case-study design to evaluate responsible AI governance in Iran through an examination of the National Artificial Intelligence Strategy. A qualitative design was selected because the research problem concerned the internal logic of a policy document, the interpretation of institutional responsibilities, the plausibility of causal assumptions, and the contextual conditions influencing implementation. These questions could not be adequately addressed through a quantitative measurement of isolated indicators. The National AI Strategy was treated as a complex policy intervention embedded in a broader institutional and technological ecosystem. The unit of policy analysis was the strategy and its governance arrangements, while the unit of interview analysis was the individual expert's interpretation of the problems, mechanisms, resources, and

outcomes associated with responsible AI governance. This distinction ensured that participants were not treated as formal representatives of their organizations and that their statements were analysed as expert judgments rather than official institutional positions.

The case was bounded by the national governance of artificial intelligence in Iran and by the policy environment within which the strategy was formulated and initially interpreted. The empirical analysis focused on issues directly relevant to the strategy's program logic, including institutional authority, allocation of responsibilities, financing, data access, computational infrastructure, human capital, ecosystem development, regulation, ethical safeguards, policy indicators, and implementation constraints. Broad historical descriptions of artificial intelligence, technical discussions unrelated to governance, and administrative material associated with the original thesis were excluded. The study did not attempt to evaluate every possible application of AI or to measure the long-term effects of the strategy, because implementation remained at an early stage. Instead, it evaluated the design of the intervention and its readiness for implementation. Such a focus is appropriate where the primary analytical question is whether a strategy contains a sufficiently coherent and operational theory for producing its intended effects.

Two sources of evidence were combined. The first consisted of documentary material, principally the National Artificial Intelligence Strategy and directly relevant policy texts concerning AI development, national technological priorities, institutional responsibilities, and regulatory arrangements. Document analysis was used to identify the strategy's explicit vision, values, objectives, indicators, priority sectors, implementation institutions, and monitoring provisions. It also enabled the identification of omissions, ambiguities, internal tensions, and areas in which the document stated a desired result without specifying the mechanism required to achieve it. The strategy was read not simply as a descriptive text but as an intervention containing assumptions about how public authority, scientific capability, infrastructure, and market actors could produce national AI development. Each major provision was therefore examined in relation to the components of program theory: the problem addressed, proposed inputs, activities, outputs, intermediate outcomes, final outcomes, responsible actors, success criteria, and contextual assumptions.

The second source consisted of semi-structured interviews with 12 experts selected for their knowledge of artificial intelligence, science and technology policy, public governance, research, or industrial development. The final sample included five participants from academic settings, four participants working primarily in research or policy-oriented institutions, and three participants with experience in private-sector AI development or technology-based industry. This composition was intended to provide variation in theoretical, policy, and implementation perspectives. Expert selection was purposive and was supplemented by snowball referrals to individuals possessing specialized knowledge of the national AI ecosystem. The sampling process did not seek statistical representativeness. Its objective was to obtain analytically rich accounts from participants capable of identifying institutional relationships, implementation barriers, and causal mechanisms that might not be explicit in the formal policy document. Recruitment continued until the available accounts provided adequate coverage of the principal governance dimensions and additional interviews were no longer producing substantially different categories at the level required for program-theory reconstruction.

The interview protocol was organized around the logic of responsible governance rather than around a general request for opinions about artificial intelligence. Participants were invited to describe the principal governance problem, the causes of Iran's current limitations, the consequences of policy inaction, the institutional arrangements required for change, the resources necessary for implementation, the appropriate roles of government, industry,

and universities, the criteria by which success should be judged, and the external factors capable of influencing outcomes. Questions also explored responsibility for AI-related harms, access to public data, computational infrastructure, the balance between innovation and regulation, cultural and ethical priorities, and the relationship between domestic capability and international markets. The semi-structured format ensured comparability across interviews while allowing participants to elaborate on issues arising from their professional experience. Interviews were conducted through available in-person, telephone, and online formats. The research treated the participants' identities as confidential in the consolidated article and reported results by analytical theme rather than by named individual.

Data analysis was conducted through a multistage coding and synthesis process supported by MAXQDA. First, the interview material and documentary evidence were reviewed to identify statements related to governance problems, causal factors, expected outcomes, activities, resources, actors, risks, indicators, and external conditions. Initial coding remained close to the language and meaning of the material so that the analytical model would not be imposed prematurely. Second, related codes were organized into broader categories, including institutional governance, data governance, computational infrastructure, human capital, ecosystem development, regulation, ethical safeguards, strategic orientation, and monitoring. Third, the categories were interpreted through the distinction between theory of change and theory of action. Statements explaining why a change should occur were assigned to the theory of change, while statements specifying institutions, resources, instruments, and implementation procedures were assigned to the theory of action. Fourth, an individual program logic was reconstructed for each expert account. These individual logics were then compared to identify areas of convergence, disagreement, and complementary emphasis.

The consolidated program theory was produced through analytical integration rather than simple frequency counting. A view expressed by many participants was treated as evidence of broad agreement, but less frequent views were retained when they identified a plausible risk, institutional conflict, or causal mechanism. The synthesis sought to preserve disagreement where policy alternatives involved genuine trade-offs. For example, experts could agree that data access was necessary while disagreeing about whether access should be equal or strategically managed. They could agree on the need for a national institution while differing on whether it should be independent, located within an existing structure, or restricted to regulation rather than implementation. These disagreements were analytically valuable because they revealed policy dualities that the national strategy must resolve. The final program theory was therefore designed to represent both common prerequisites and alternative pathways.

The document was subsequently evaluated against the reconstructed theory. Each component of the consolidated theory was compared with the content of the National AI Strategy to determine whether it was clearly present, ambiguously present, or absent. The evaluation examined whether the strategy recognized the relevant issue, assigned an accountable actor, specified a feasible action, provided or identified necessary resources, and established a measurable criterion. Particular attention was given to the relationship among components. A strategic objective was not considered operational merely because it appeared in the text; it was assessed according to whether it was connected to an institution, instrument, intermediate result, and monitoring arrangement. This relational analysis was intended to avoid a checklist approach in which the mere mention of a concept is treated as evidence of adequate governance.

The study also extracted 12 policy dualities from the interviews and documentary analysis. These dualities concerned public data governance, responsibility for AI error, regulatory strictness, local and global ethics, equal

and managed data access, domestic and international market orientation, open and controlled data, domestic and international standards, public and private infrastructure investment, centralized and plural governance, human-centred and machine-centred development, problem orientation and technology orientation, and the relationship between self-regulation and independent oversight. Closely related formulations were consolidated so that the final analytical structure represented 12 distinct strategic choices. For each duality, the analysis identified the consequences of the alternatives, the position explicitly or implicitly adopted by the strategy, and the implementation problem created by ambiguity. The purpose was not to force every issue into a binary choice, but to evaluate whether the strategy offered a defensible hybrid arrangement where both orientations were necessary.

Trustworthiness was supported through triangulation between documents and interviews, comparison across participants from different professional sectors, preservation of divergent interpretations, and iterative review of the relationship between codes and the reconstructed program theory. Analytical decisions were recorded so that the transition from source material to categories and from categories to the final model could be followed. The integrated theory was assessed for internal coherence, empirical grounding, operational specificity, contextual plausibility, and evaluability. Negative and contradictory cases were examined rather than removed. The study also distinguished factors that could be influenced by the strategy, such as institutional mandates or funding mechanisms, from external factors such as sanctions, international technological concentration, or geopolitical competition. This distinction reduced the risk of attributing every anticipated success or failure to the policy document itself.

The study is interpretive and evaluative rather than predictive. It does not claim that every proposed mechanism will necessarily succeed or that all experts share an identical vision of responsible governance. Its contribution lies in making the assumptions of the national strategy more explicit and identifying the institutional conditions under which those assumptions could plausibly operate. AI-governance research has emphasized that effective systems must be adapted to organizational and national context rather than transferred mechanically from general frameworks (23). The present methodology therefore combined internationally recognized governance concerns with evidence about Iran's specific institutional, economic, infrastructural, and geopolitical conditions. This approach made it possible to evaluate the National AI Strategy as a context-dependent intervention rather than as an abstract statement of universal principles.

Findings and Discussion

The integrated findings indicate that responsible AI governance in Iran depends on a causal sequence beginning with political recognition of AI as a strategic national issue and progressing through institutional capacity, infrastructure, ecosystem development, responsible application, and continuous evaluation. The theory of change reconstructed from the expert accounts begins with a diagnosis of fragmented governance, insufficient computational resources, restricted and poorly organized data, weak connections among universities and industry, unstable support for skilled personnel, and uncertainty concerning the regulation of emerging risks. These problems are mutually reinforcing. Institutional fragmentation prevents coordinated investment; inadequate infrastructure discourages ambitious projects; limited applied demand weakens university–industry collaboration; and uncertainty regarding data access reduces incentives for both public research and private innovation. The intended change is therefore not simply an increase in the number of AI projects. It is the formation of a coherent national ecosystem capable of producing, adapting, deploying, and governing AI in ways that generate economic and public value without sacrificing cultural, legal, ethical, or security considerations.

The first causal requirement is durable political commitment. AI initiatives that depend on temporary enthusiasm or individual officeholders cannot sustain investments in infrastructure, education, datasets, and regulatory capacity. Political commitment must be translated into a stable allocation of authority and resources. The findings show that a powerful and agile national institution is widely regarded as a necessary condition for implementation. This institution requires an explicit legal mandate, a sustainable budget, access to high-level decision-making, and the authority to coordinate organizations whose responsibilities overlap. At the same time, the interviews indicated that such an institution should not become the direct executor of every AI project. Its primary functions should be strategic policymaking, coordination, standard-setting, regulation, resource facilitation, monitoring, and resolution of inter-organizational conflicts. Direct implementation should generally remain with ministries, universities, firms, sectoral institutions, and specialized organizations possessing the relevant operational capacity. This distinction reduces conflict between promotion and oversight and prevents the national body from becoming an additional bureaucratic layer competing with existing actors.

The National AI Strategy represents progress insofar as it recognizes the need for national coordination and provides a formal location for AI within the country's policy architecture. However, the evaluation identified ambiguity concerning the precise legal position, operational independence, financial authority, and regulatory instruments of the proposed governance institutions. Establishing a council or organization does not automatically create governance capacity. The institution must be capable of requiring information from implementing bodies, enforcing standards, coordinating budgets, evaluating compliance, and initiating policy revision. Iranian research on AI regulation similarly identifies institutional authority and the design of the regulatory system as unresolved issues (10). The study's findings add that the national institution must be evaluated not by its formal title but by whether it possesses the practical instruments necessary to influence the behaviour of ministries, firms, universities, and data-holding organizations. A formally independent institution without resources may be weaker than a unit located within an established structure, while a well-funded institution without regulatory accountability may intensify centralization without improving responsibility.

The second central requirement is the coordinated development of data and computational infrastructure. Expert accounts consistently treated high-quality data as a national strategic resource. Iran possesses substantial datasets in public administration, healthcare, education, transportation, energy, agriculture, culture, and industry, yet these resources are often fragmented across organizations, stored in incompatible formats, or inaccessible because of legal uncertainty and security concerns. The consolidated theory of action therefore includes the creation of a national platform or regulated marketplace for data exchange. Such a platform would not require the unrestricted publication of all public data. It would establish classification standards, procedures for anonymization, secure access, licensing conditions, technical interfaces, and responsibility for data quality. Government agencies would retain stewardship over sensitive datasets while enabling controlled research and innovation where legitimate conditions were satisfied. The objective would be to replace the current binary of complete closure or uncontrolled openness with graduated and accountable access.

The findings reveal that the National AI Strategy recognizes the importance of data but does not sufficiently operationalize the distinction among open, managed, restricted, and highly sensitive information. This ambiguity is particularly consequential in an environment where national-security concerns can lead organizations to classify broad categories of data as sensitive. Overclassification protects institutions from immediate administrative risk but transfers costs to researchers, startups, public-service innovators, and policymakers who lack access to evidence.

Conversely, an indiscriminate open-data policy could expose personal information, strategic infrastructure, or commercially valuable records. Responsible governance requires precise criteria, not general preferences. The ethical literature has long identified privacy and data governance as central components of legitimate AI (5). The present findings demonstrate that, in the Iranian case, the principal challenge is not merely endorsement of privacy but the design of a tiered access regime capable of reconciling public value and legitimate protection.

Computational capacity emerged as an equally important bottleneck. Advanced AI development requires processors, storage, cloud infrastructure, secure networks, and reliable electricity. International restrictions and concentration of semiconductor production make these inputs costly and strategically vulnerable. The theory of action therefore proposes gradual and organized expansion of shared computational infrastructure rather than an immediate attempt to replicate the largest global AI facilities. The government has a defensible role in financing capital-intensive infrastructure that individual firms cannot provide, while private actors can develop services and applications through transparent access arrangements. Access should be managed according to public value, technical merit, strategic priority, and inclusion of qualified smaller actors. Without such rules, national infrastructure could be captured by a limited number of politically connected organizations. The broader governance literature increasingly recognizes that responsible AI includes distributive questions about who can access the resources required to innovate (6). In this sense, infrastructure policy is also competition policy and social policy.

The expert evidence supported an emphasis on efficient and relatively lightweight models suited to conditions of limited data and computing. This approach does not imply rejection of foundational research or permanent technological inferiority. It reflects strategic adaptation. A country with constrained access to advanced hardware may generate greater value by developing specialized models for Persian language processing, healthcare, agriculture, industrial maintenance, public administration, cultural content, or environmental management than by attempting to compete immediately in every layer of large-scale model development. Efficient models can reduce costs, enable deployment on local hardware, improve privacy through on-premise use, and allow adaptation to sector-specific datasets. The finding is consistent with the broader proposition that national AI strategies must reflect domestic capabilities rather than simply imitate the priorities of leading technological powers (2). A responsible strategy should nevertheless preserve pathways for long-term capability-building so that short-term pragmatism does not become permanent dependence.

Human capital constitutes the third major component of the program theory. Participants did not define the problem solely as a shortage of technical graduates. They emphasized weak interdisciplinary integration, insufficient practical connection between research and national needs, migration of skilled personnel, and limited institutional demand for high-quality AI work. The theory of change assumes that sustainable capability emerges when education, research, industry, and public-sector problems are connected. Universities should support advanced scientific work and critical ethical inquiry; startups should experiment with applications and generate tacit knowledge; established industries should provide scaling opportunities and market access; public institutions should define high-value problems and create responsible procurement mechanisms. The national governing body should facilitate these connections rather than attempt to replace them. This finding reinforces stakeholder-based approaches in which responsible governance is produced by coordinated organizational roles rather than by a single regulator (21).

The analysis identified the university–industry relationship as one of the strategy's most important implementation weaknesses. General statements encouraging collaboration are insufficient when academic

promotion systems reward publication more than applied partnership, industries lack confidence in university delivery, public procurement is slow, and intellectual-property rules are uncertain. Systematic collaboration requires joint laboratories, shared doctoral projects, challenge-based funding, sectoral consortia, transparent procurement, mobility between academia and industry, and access to real datasets. These mechanisms should be connected to priority fields rather than dispersed across unrelated projects. The study therefore supports a vertical development strategy in which limited resources are concentrated in selected domains such as health, agriculture, education, industry, public services, and Persian-language content. Vertical development enables the formation of complete value chains from data and research to application, regulation, and evaluation. A purely horizontal approach that invests thinly in every field risks producing fragmented pilot projects without sufficient scale or institutional adoption.

The strategy's human-centred and culturally grounded orientation represents one of its principal normative strengths. It repeatedly positions AI as an instrument for improving human welfare, preserving dignity, supporting society, and advancing national development. This orientation is compatible with global responsible-AI principles while allowing attention to Iran's cultural and legal context. The findings nevertheless show that the operational meaning of cultural values remains underdeveloped. A commitment to national or religious values requires an institution authorized to interpret relevant principles, procedures for resolving disagreement, and safeguards against using cultural language to justify arbitrary restriction. Global studies of AI ethics have found considerable convergence around transparency, justice, responsibility, privacy, and human oversight, despite differences in emphasis (5). Iran can therefore articulate locally grounded priorities while maintaining compatibility with internationally recognized protections. The key requirement is to define local values in a manner that produces clear and reviewable governance standards.

The theory of action also requires a risk-based regulatory architecture. Participants generally rejected both unrestricted innovation and uniform strict regulation. Different applications create different levels and types of risk. A recommendation system for entertainment content should not be regulated in the same manner as an AI system influencing medical diagnosis, legal decisions, critical infrastructure, or access to public benefits. Responsible regulation should classify applications according to potential harm, affected rights, scale, reversibility, vulnerability of affected populations, and degree of autonomous decision-making. Low-risk applications may be governed through disclosure, voluntary codes, and basic documentation. Higher-risk applications require impact assessments, technical validation, human oversight, incident reporting, auditability, and rights of appeal. Prohibited or exceptionally dangerous applications may require strict limitation. The risk-based logic of collaborative regulation provides a means of protecting society without imposing identical requirements on every innovator (12).

The National AI Strategy recognizes ethical and social concerns but does not establish a sufficiently detailed regulatory ladder connecting risk categories to obligations. It also leaves important questions of liability unresolved. When an AI-assisted medical decision causes harm, responsibility may involve the model developer, data provider, hospital, professional user, software vendor, or regulator. Assigning all responsibility to the developer ignores the organizational context of use; assigning it entirely to the end user allows developers and deploying institutions to avoid foreseeable duties. The findings support a distributed model of responsibility in which obligations correspond to control, knowledge, benefit, and capacity to prevent harm. Developers should be responsible for documentation, testing, known limitations, and foreseeable technical risks. Deploying organizations should be responsible for suitability, staff training, monitoring, and integration into professional processes. Users should be responsible for complying with defined procedures, while public authorities should establish standards, supervision, and remedy.

Contemporary analysis of generative AI similarly shows that responsibility cannot be attributed exclusively to data, models, or regulation but must be allocated across the system (3).

Transparency was identified as another area in which ethical aspiration requires institutional mechanisms. Transparency does not necessarily mean publication of source code or proprietary model parameters. It means that relevant actors can obtain the information necessary to understand purpose, limitations, data provenance, performance, decision pathways, and available remedies. Different stakeholders require different forms of transparency. Regulators may need detailed technical documentation, professional users may need information about reliability and limitations, affected citizens may need understandable explanations and appeal channels, and investors or managers may require performance and risk indicators. Business-intelligence systems can support responsible governance by integrating compliance, performance, and risk information into organizational decision-making (24). The National AI Strategy would therefore benefit from defining documentation and reporting requirements appropriate to different levels of risk.

The reconstructed program theory identifies monitoring and policy learning as essential because both AI capabilities and governance risks change rapidly. Long-term visions can provide strategic continuity, but they must be supported by shorter implementation cycles. The strategy should be translated into time-bound action plans specifying responsible actors, budgets, outputs, and indicators. These plans should be reviewed periodically through a process involving regulators, technical experts, sectoral institutions, industry, and affected social groups. Monitoring should distinguish capacity indicators from responsible-governance indicators. Publication counts, patent numbers, firm creation, investment volume, or international rankings measure activity but not necessarily public value or responsibility. A country may improve its scientific position while simultaneously increasing unequal access, privacy violations, or dependence on imported infrastructure. Composite indicators are therefore needed to capture institutional quality, data accessibility, safety, sectoral impact, inclusion, public trust, and accountability. Impact-oriented evaluation is a central requirement of meaningful AI governance frameworks (22).

The analysis of policy dualities provides a more detailed account of the strategy's unresolved choices. The first cluster concerns data, access, and responsibility. In the duality between open public data and sensitive data, open access can promote innovation, transparency, research, and accountability, whereas restriction can protect privacy, national security, and legitimate institutional interests. The strategy generally supports data development and sharing but lacks precise boundaries. A responsible position would combine presumptive openness for low-risk non-personal data with controlled access for sensitive datasets and strict limitation for data presenting serious security or privacy risks. A related duality concerns equal access versus managed access. Equal access reduces monopoly and allows smaller actors to participate, but scarce resources may require prioritization. Managed access is defensible when criteria are transparent, reviewable, and connected to public value. Without such safeguards, strategic allocation can become administrative favouritism. The third issue concerns responsibility for error. The document acknowledges responsibility in general terms but does not provide a clear allocation across developers, deploying organizations, users, and the state. This is a significant program-theory gap because actors cannot implement undefined obligations.

The second cluster concerns regulatory intensity, values, and oversight. Strict regulation can prevent serious harms and establish public trust, but excessive ex ante requirements may impose costs that only large organizations can bear, thereby reducing competition. Flexible regulation can support experimentation but may externalize risk onto citizens and public institutions. The study's findings favour adaptive and risk-proportionate regulation

supported by sandboxes, provisional approvals, monitoring, and escalation of obligations as risk increases. This position is consistent with arguments for agile governance in fast-changing technological fields (11). The duality between local values and global ethics should similarly be managed through complementarity rather than isolation. Local principles can guide culturally appropriate applications and public legitimacy, while international norms support interoperability, rights protection, and participation in global markets. Comparative governance research warns that fragmented and incompatible regulation can create uncertainty and restrict cross-border technological exchange (8). The strategy's emphasis on national values is clear, but mechanisms for alignment with international standards remain insufficiently developed.

The choice between industry self-regulation and an independent regulator presents another critical tension. Self-regulation allows technically informed and rapidly updated standards, lowers administrative costs, and can create professional responsibility. It also involves conflicts of interest because firms may benefit from underestimating risk or limiting disclosure. An independent regulator can improve accountability and public trust but may become slow, under-resourced, or detached from technical change. The findings support a co-regulatory structure in which industry associations and professional bodies develop detailed standards under the authority and supervision of an independent public regulator. The regulator should define minimum obligations, audit rights, incident-reporting requirements, and sanctions, while industry contributes technical expertise and implementation guidance. Governance research across sectors indicates that combined mechanisms are often more effective than exclusive reliance on either voluntary organizational policy or centralized command (15).

The third cluster concerns institutional concentration, investment, markets, and standards. Centralized governance can create strategic coherence, prevent duplication, and resolve conflicts, whereas plural governance can incorporate knowledge from government, industry, universities, professional communities, and civil society. The National AI Strategy adopts a mixed orientation by establishing national-level institutions while also recognizing multiple stakeholders. This combination is potentially appropriate, but the relationship between central authority and distributed implementation must be clarified. The national institution should define direction and mandatory safeguards, while sectoral actors adapt implementation to their contexts. Participation should not create veto points that paralyse action, and centralization should not eliminate transparency or stakeholder knowledge. Responsible governance systems require both authoritative coordination and channels for distributed expertise (16).

The public-versus-private investment duality is particularly consequential for computational infrastructure. The strategy emphasizes private and knowledge-based firms but provides less clarity regarding the state's responsibility for expensive shared infrastructure. Private firms are generally more agile and responsive to markets, but the domestic private sector may lack the capital to finance large computational facilities under conditions of economic uncertainty. State investment can overcome this barrier but may produce inefficiency, corruption, or politically allocated access. The preferred program logic is therefore complementary. Government should invest in public-good components that markets are unlikely to provide, while private actors should compete in services, applications, optimization, and commercialization. Public–private arrangements require transparent procurement, performance criteria, independent oversight, and open access rules. Otherwise, public investment may socialize cost while privatizing strategic advantage.

The domestic-versus-international market duality concerns the scale and sustainability of Iran's AI ecosystem. Domestic applications are essential because they address national needs, use local data, and build trust among institutions. Yet the domestic market alone may be insufficient to sustain firms requiring continuous research and

infrastructure investment. International orientation can generate revenue, expose firms to higher standards, and support technological learning, but sanctions, regulatory incompatibility, and geopolitical risk create substantial obstacles. The strategy emphasizes national capability but does not fully articulate a pathway to international markets. The study suggests that export potential should be incorporated from the beginning in fields where Iran can develop specialized capability, including Persian-language technologies, regionally relevant industrial applications, efficient models, and culturally adapted services. This requires attention to international technical, ethical, and documentation standards rather than an exclusively domestic regulatory framework.

The relationship between domestic standards and international regulation is therefore not a secondary issue. Exclusive reliance on imported standards may neglect local institutional capacity, cultural context, and national priorities. Exclusive reliance on domestic standards may isolate firms, reduce interoperability, and limit international trust. A layered framework can preserve mandatory national requirements while mapping them against recognized international principles and sectoral standards. The literature on global AI governance consistently identifies international compatibility as necessary because data, software, platforms, and supply chains cross national boundaries (18). The National AI Strategy's strong emphasis on localization should be complemented by a systematic process for monitoring external regulatory developments and identifying areas of equivalence, adaptation, or justified divergence.

The fourth cluster concerns the direction of technological development. The strategy is clearly human-centred in its stated values, and this represents a significant strength. Nevertheless, human-centred governance requires more than declaring that technology should serve humanity. It requires measurable protection of autonomy, privacy, safety, equality, and the right to challenge consequential decisions. The alternative machine-centred orientation prioritizes efficiency, automation, and technical capability, potentially treating human adaptation as the primary solution to social disruption. The study's findings reject this orientation as a national governing principle, while recognizing that technical performance remains necessary. Human-centred AI is not technically weak AI; it is high-performing AI constrained and directed by legitimate human purposes. Ethical challenges in digital media demonstrate how technical optimization for engagement or scale can undermine autonomy and public trust when human consequences are treated as externalities (19).

The final duality concerns problem-oriented and technology-oriented development. Problem orientation directs resources toward identifiable national needs and can demonstrate public value in health, water, agriculture, education, industry, or government services. Technology orientation builds foundational knowledge, models, infrastructure, and research capacity whose applications may not be immediately visible. An exclusive problem orientation risks dependence on external foundational technologies, while an exclusive technology orientation risks expensive research disconnected from social demand. The findings therefore support a portfolio approach led by vertical priorities. Each priority domain should connect applied problems to selected foundational capabilities. A healthcare initiative, for example, may require domestic language models, secure data infrastructure, clinical validation, and professional training. This arrangement allows immediate applications to contribute to long-term capacity rather than competing with it.

External factors significantly influence whether the program theory can succeed. Participants identified restrictions on advanced processors and international services, currency instability, geopolitical competition, migration of skilled personnel, unreliable electricity and internet infrastructure, administrative rigidity, security-driven data closure, and rapid technological change. These factors should not be treated as reasons to abandon strategic

planning. They should be incorporated as assumptions and risks. The program should establish alternative supply strategies, prioritize computational efficiency, diversify technical partnerships where possible, improve retention incentives, and create contingency plans for infrastructure disruption. Geopolitical analyses of AI indicate that technological capability increasingly affects the distribution of international power (7). Responsible governance in Iran must therefore protect individual and social rights while also confronting the structural conditions of technological dependency.

The overall evaluation indicates that Iran's National AI Strategy possesses a recognizable normative vision and identifies many relevant components of an AI ecosystem. Its strengths include political recognition of AI, human-centred values, attention to national culture, identification of priority fields, support for private and knowledge-based firms, and acknowledgement of data, education, research, and governance institutions. Its principal deficiency lies in the incomplete connection among these components. Goals are often clearer than responsibilities; responsibilities are clearer than enforcement instruments; technological aspirations are more developed than resource commitments; and general indicators are more visible than measures of responsible outcomes. This pattern reflects a common transition in AI governance from ethical and strategic declaration to operational implementation (23). The strategy should therefore be understood as an important starting framework rather than a complete governance architecture.

The theoretical contribution of the study is the demonstration that responsible AI governance should be evaluated as a program of institutional change. The relevant question is not merely whether a strategy mentions fairness, innovation, privacy, or national capability. The question is whether it contains a credible causal chain through which political commitment generates institutional authority, institutional authority mobilizes resources, resources enable coordinated activities, activities produce sectoral and technological outputs, outputs generate socially valuable outcomes, and monitoring corrects unintended consequences. Responsible governance fails when any link remains undefined. Recent reviews similarly emphasize that responsible AI is an organizational and institutional capability rather than a static list of principles (6). Program theory makes this capability visible and evaluable.

Conclusion

Artificial intelligence has become a strategic domain in which technological ambition, public responsibility, economic development, cultural legitimacy, and national autonomy are inseparably connected. A national AI strategy must therefore do more than express support for innovation or reproduce generally accepted ethical principles. It must establish a coherent system through which institutions, resources, rules, expertise, and accountability mechanisms work together. The present study evaluated Iran's National Artificial Intelligence Strategy from this perspective by reconstructing the theory of change and theory of action required for responsible AI governance and examining the extent to which the strategy contains the necessary components.

The reconstructed theory of change begins with the recognition that Iran's AI challenges are systemic. Fragmented institutional responsibilities, limited computing capacity, inaccessible or poorly organized data, weak university–industry collaboration, migration of skilled personnel, uncertain regulatory responsibilities, and international technological restrictions cannot be resolved through isolated projects. Progress depends on a coordinated sequence of changes. Stable political commitment must support a capable national institution. That institution must coordinate sustainable investment, clarify responsibilities, facilitate data and infrastructure access, promote interdisciplinary human capital, support priority applications, establish proportionate regulation, and

continuously evaluate outcomes. The final objective is not merely increased technological activity but a nationally sustainable ecosystem capable of producing public value while preserving dignity, security, privacy, fairness, and social trust.

The findings indicate that a powerful and agile national governing institution is necessary, but power must be accompanied by clearly defined functions and accountability. The national body should concentrate on strategy, regulation, coordination, facilitation, standard-setting, and monitoring rather than attempting to execute every technological project. Implementation should remain distributed among competent ministries, universities, public institutions, firms, and sectoral actors. This model combines centralized strategic direction with plural implementation and reduces the risk that a new organization will become another layer of bureaucracy. Its effectiveness depends on legal authority, stable funding, access to information, the ability to coordinate budgets and standards, and protection from repeated institutional restructuring.

Data and computing are foundational public-policy concerns. Iran requires a differentiated system that separates open, managed, restricted, and highly sensitive data. Government datasets should not be closed merely because no access mechanism exists, nor should they be released without adequate privacy and security protection. A national data-exchange framework should define stewardship, anonymization, licensing, secure access, data quality, and responsibility for misuse. Computational infrastructure should be developed gradually through shared facilities, transparent allocation criteria, and complementary public and private investment. Priority should also be given to efficient and specialized models that can operate under resource constraints and address identifiable national needs.

Human capital must be treated as an ecosystem rather than as a numerical supply of graduates. Interdisciplinary training, meaningful research opportunities, access to data and computing, links with industry, and demand from public institutions are necessary to retain and effectively use expertise. Universities should contribute foundational and applied research, startups should conduct experimentation, large industries should provide scaling capacity, and government should create public-good infrastructure and responsible procurement. Vertical development in selected fields offers a practical means of connecting these actors. Concentrating resources in health, agriculture, education, industry, public services, and Persian-language content can produce complete chains of capability rather than scattered projects with limited impact.

The strategy's human-centred orientation provides an appropriate normative foundation, but normative principles must be converted into enforceable procedures. Transparency requires differentiated documentation and disclosure. Fairness requires testing, impact assessment, monitoring, and correction. Human oversight requires defined authority to review or override consequential outputs. Privacy requires data minimization, access controls, anonymization, and restrictions on secondary use. Accountability requires a distributed allocation of duties among developers, data providers, deploying organizations, professional users, and public regulators. Citizens affected by high-impact systems should have access to explanation, review, complaint, and remedy.

Regulation should be risk-based, adaptive, and proportionate. Uniformly strict regulation may suppress beneficial innovation, while excessive flexibility may transfer risks to the public. Low-risk systems can be governed through basic transparency and organizational controls, whereas high-impact applications require validation, documentation, human supervision, incident reporting, independent audit, and appeal mechanisms. Regulatory sandboxes and provisional authorization can support experimentation when they operate within clear boundaries and produce evidence for subsequent policy revision. Self-regulation can contribute technical expertise, but it should

function under the authority of an independent public regulator capable of establishing minimum obligations and imposing consequences.

The evaluation of policy dualities demonstrates that strategic clarity is one of the most important functions of a national AI document. The strategy must explain how it will balance open data with sensitive-data protection, equal access with strategic allocation, strict regulation with innovation, national values with international principles, centralized coordination with stakeholder participation, domestic capability with global-market orientation, public investment with private initiative, and self-regulation with independent oversight. Hybrid arrangements are often preferable, but a hybrid model must specify the boundary between its components. Ambiguity should not be mistaken for flexibility.

The National AI Strategy contains important strengths. It recognizes artificial intelligence as a national priority, adopts a human-centred vocabulary, emphasizes scientific and technological development, acknowledges the role of universities and private firms, identifies priority sectors, and seeks to create a coordinated governance structure. These elements establish a valuable foundation. The principal problem is that the strategy's aspirational architecture is more developed than its implementation architecture. The document does not always connect objectives to accountable institutions, resources, intermediate outputs, regulatory instruments, and measurable effects. This gap could create implementation friction even where political support exists.

Policy improvement should therefore focus on operationalization. The legal mandate and financial authority of the national governing institution should be clarified. Policymaking and regulation should be distinguished from direct execution. A tiered national data-governance system should be established. Shared computing infrastructure should be supported through transparent public–private arrangements. Liability should be distributed according to control, knowledge, and capacity to prevent harm. Priority sectors should receive integrated support connecting research, data, infrastructure, application, and regulation. University–industry collaboration should be institutionalized through joint laboratories, challenge-based funding, shared projects, and responsible procurement. The strategy should also adopt short implementation cycles within its long-term vision so that targets, risks, and regulatory instruments can be revised as technology changes.

Evaluation must move beyond isolated statistical indicators. Publication counts, patents, company formation, and international rankings are useful but cannot demonstrate responsible governance by themselves. A national evaluation framework should also measure the accessibility and quality of data, distribution of computational resources, effectiveness of regulatory institutions, quality of impact assessments, incidence and correction of harm, sectoral public value, participation of affected stakeholders, and public confidence. Indicators should be linked to the causal chain of the strategy so that policymakers can determine whether weak outcomes result from inadequate design, poor implementation, or external constraints.

The study is limited by its qualitative design and the size of the expert sample. The findings provide analytical depth rather than statistical generalization. The institutional environment is also subject to rapid change, and some arrangements examined during the study may subsequently be revised. Expert accounts inevitably reflect professional position and experience, even though variation across academic, policy, and industrial participants strengthened the analysis. The study primarily evaluates policy design and implementation readiness rather than long-term outcomes, which cannot be measured before sustained implementation has occurred. Program theory also provides one analytical perspective and does not fully replace innovation-system, political-economy, legal, or comparative institutional approaches.

Future research should develop measurable composite indicators of responsible national AI governance, conduct longitudinal evaluation of the strategy's implementation, and compare Iran with countries facing similar resource or geopolitical constraints. Sector-specific research is needed in healthcare, education, public administration, finance, agriculture, media, and critical infrastructure because responsible governance requirements differ across applications. Further studies should examine the legal allocation of AI liability, public attitudes toward automated decision-making, the institutional design of an independent regulator, secure models of public-data access, and mechanisms for retaining interdisciplinary expertise. Comparative analysis of national and international standards could also identify areas in which local adaptation is necessary and areas in which alignment would improve interoperability and trust.

The central conclusion is that responsible AI governance is not achieved through a single organization, ethical charter, technological investment, or regulatory law. It is produced through the alignment of institutions, resources, incentives, expertise, values, and evaluation. Iran's National Artificial Intelligence Strategy can provide the foundation for this alignment, but its effectiveness depends on transforming broad commitments into an explicit, funded, accountable, and adaptable program of action.

Acknowledgments

We would like to express our appreciation and gratitude to all those who helped us carrying out this study.

Authors' Contributions

All authors equally contributed to this study.

Declaration of Interest

The authors of this article declared no conflict of interest.

Ethical Considerations

All ethical principles were adhered in conducting and writing this article.

Transparency of Data

In accordance with the principles of transparency and open research, we declare that all data and materials used in this study are available upon request.

Funding

This research was carried out independently with personal funding and without the financial support of any governmental or private institution or organization.

References

1. Ng A. Why AI Could Be as Transformational as Electricity. 2017.
2. Fatima S, Desouza KC, Dawson GS, Denford JS. National Strategic Artificial Intelligence Plans: A Multi-Dimensional Analysis. Brookings Institution, 2020.

3. Raza S, Qureshi R, Zahid A, Fioresi J, Sadak F, Saeed M, et al. Who Is Responsible? Data, Models, or Regulations: A Comprehensive Survey on Responsible Generative AI for a Sustainable Future. 2025.
4. Janssen M. Responsible Governance of Generative AI: Conceptualizing GenAI as Complex Adaptive Systems. Policy and Society. 2025;44(1):38-51. doi: 10.1093/polsoc/puae040.
5. Jobin A, Ienca M, Vayena E. The Global Landscape of AI Ethics Guidelines. Nature Machine Intelligence. 2019;1(9):389-99. doi: 10.1038/s42256-019-0088-2.
6. Papagiannidis E, Mikalef P, Conboy K. Responsible Artificial Intelligence Governance: A Review and Research Framework. Journal of Strategic Information Systems. 2025;34(2):101885.
7. Mazdak E, Hamidi S. The Artificial Intelligence Juncture and the Transformation of the Geopolitical Order. Strategic Studies. 2022;10(3):120-45.
8. Zidan I, Ibrahim E. AI Governance in a Complex and Rapidly Changing Regulatory Landscape: A Global Perspective. Global Regulatory Review. 2024;12(1):23-45.
9. Mehraban MM, Akbari I. Institutional Mapping and National Division of Labor in the Field of Artificial Intelligence Development and Data-Driven Governance. Iranian Technology Studies. 2023;7(1):5-23.
10. Mahmoudi M, Mohammadzadeh M, Akbari A. Artificial Intelligence Governance (5): The Artificial Intelligence Regulatory System in the Country. Governance Studies Office, Islamic Parliament Research Center, 2024.
11. Wallach W, Marchant GE. Toward the Agile and Comprehensive International Governance of AI and Robotics. Proceedings of the IEEE. 2018;107(3):505-12.
12. Cancela-Outeda C. The EU's AI Act: A Framework for Collaborative Governance. Internet of Things. 2024;27:101291. doi: 10.1016/j.iot.2024.101291.
13. Batool A, Zowghi D, Bano M. AI Governance: A Systematic Literature Review. Research Square, 2024.
14. Cath C. Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2018;376(2133):20180080. doi: 10.1098/rsta.2018.0080.
15. Liao W, Zhao Z, Li H. Governance Mechanisms for Responsible AI: A Cross-Industry Review. Journal of Technology Ethics. 2023;2(3):198-210.
16. Athrad B, Lyons C. AI Governance Systems. Journal of Emerging Technology Governance. 2024;10(2):112-29.
17. Wakuri K, Yamamoto T, Saito M. Challenges in AI Governance: A Comparative Study of Ethical Guidelines. Journal of Ethical AI. 2021;3(1):45-62.
18. Lou Y, Tang X, Song L. Frameworks for AI Governance: Addressing Ethical Challenges and Global Policies. Journal of AI Policy and Governance. 2022;5(3):201-15.
19. Kalateh Aghamohammadi A. Ethical Challenges of Artificial Intelligence in New Media and Strategies for Confronting Them. 2023.
20. Balabanova N, Bashir A, Bova P, Buscemi A, Cimpeanu T, Correia da Fonseca H, et al. Media and Responsible AI Governance: A Game-Theoretic and LLM Analysis. 2025. doi: 10.48550/arXiv.2503.09858.
21. Zoughi A, Zimney R. AI Ethics and Governance: Exploring Stakeholder Roles and Organizational Policies. International Journal of AI Research. 2023;12(4):345-70.
22. Field A, Jones T, Smith L. AI Governance Frameworks: Evaluating Impacts and Policy Recommendations. AI & Society. 2020;35(4):857-73.
23. Batool A, Zowghi D, Rimini FD. Responsible AI Governance: A Systematic Literature Review. 2023.
24. Agbadamasi TO, Opoku LK, Adukpo TK, Mensah N. The Role of Business Intelligence in AI Ethics: Empowering U.S. Companies to Achieve Transparent and Responsible AI. EPRA International Journal of Economics, Business and Management Studies. 2025;12(3). doi: 10.36713/epra20314.