International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 3
Article Received: *25 September 2026* ***Revised:*** *12 October 2026* ***Accepted:*** *15 November 2026*

Reversible Government with Digital Twins

Abdullah Mohamed Zaky Elsaied
Cloud and Disaster Recovery Expert Dammam, Saudi Arabia
Abdullah_zaky@live.com

Yousof Mohamed Salaheldin Mohamed
Cyber security technical lead Dammam, Saudi Arabia
Salah.yousof@gmail.com

Amr Mohamed Abdelfatah Elsayed
Cybersecurity GRC Expert Riyadh, Saudi Arabia
eng.amrelsayed@outlook.com

Nourah Basher Alenazi
General Manager Of Digital Infrastrucure Eastern Province Municipality Khobar, Saudi Arabia
nourah.alonazi@gmail.com

Hossam Mohamed Sherif
Strategy, Governance, Risk, Compliance & Business Continuity Consultant Khobar, Saudi Arabia
hosamdba@hotmail.com

Abstract—Governments are moving from descriptive digital services to systems that predict, recommend, and increasingly implement interventions in areas such as transport, energy, emergency response, and public administration. Current AI governance approaches focus on principles, risk classification, documentation, and human oversight. Existing digital twin programs focus on operational awareness, prediction, and optimization. However, neither approach adequately answers a fundamental practical question: What evidence is required before allowing an algorithmic public decision to alter reality, and how can a government recover if the decision proves harmful? This paper presents the Reversible Government Digital Twin (RGDT), an original conceptual framework that treats the civic digital twin as a pre-implementation authorization environment, not merely a conceptualization or planning tool. This framework requires simulating proposed machine-assisted actions under multiple population and infrastructure scenarios, validating their legal, security, equity, and resilience implications, defining clear mandates, and pairing them with a pre-implementation reversal plan. This research identifies seven architectural layers, five reversibility categories, a Decision Reversibility Index (DRI), authorization gateways, a Decision Evidence Package (DEP), and a phased maturity model. A practical municipal traffic management scenario demonstrates how AI-generated interventions can be tested, scaled, monitored, and reversed. The research underscores the need for reversibility to become a core quality attribute of digital government, alongside security, privacy, interoperability, and interpretability. The RGDT model does not eliminate uncertainty or replace democratic governance; rather, it transforms implicit assumptions into testable evidence and reveals the cost, time, and limitations of corrective action before implementation.

Keywords—Digital government; digital twins; artificial intelligence governance; smart cities; algorithmic accountability; reversibility; simulation; public-sector innovation; human oversight.

Introduction

Digital government is undergoing a fundamental transformation. The first-generation digitized forms, payments, and records. The next generation connected agencies through shared platforms and data. The emerging generation adds artificial intelligence, spatial computing, the Internet of Things (IoT), and digital twins to systems capable of sensing conditions, predicting outcomes, and recommending or implementing interventions. Recent international guidelines point to a trend toward AI-enabled urban environments, where connected infrastructure increasingly supports proactive public management [1, 2]. Similarly, the Saudi Digital Government Authority is measuring the government's readiness to adopt emerging technologies, reflecting a shift from isolated experiments to institutional capabilities [3].

This transformation is changing the nature of risks associated with public technology. For example, a chatbot that drafts an answer before posting it can be corrected. An optimization service that alters traffic light timing, prioritizes inspections, reallocates emergency resources, or modifies energy demand can impact safety, access, and rights very quickly. The concern is not simply the average accuracy of an AI model. The government must also determine whether the proposed measure

is legal in its context, whether its benefits and burdens are unevenly distributed, whether the services relied upon can sustain it, and whether the measure can be paused or reversed if reality differs from expectations.

Existing governance mechanisms are necessary but incomplete. The National Institute of Standards and Technology (NIST) AI Risk Management Framework regulates activities related to governance, mapping, measurement, and management, and explicitly calls for the definition of human roles, testing, incident response processes, and contingency plans [4, 5]. EU AI law requires risk management, registration, transparency, and human oversight of high-risk systems [6]. Similarly, Saudi Arabia's directives for government use of generative AI emphasize ethics, accountability, privacy, security, and human oversight [7]. These tools improve governance but generally regulate the AI lifecycle at the system level. It does not provide a reusable operating mechanism through which each high-impact action can be tested as a specific intervention against an existing civil situation.

Digital twins provide a missing technological infrastructure. A digital twin is more than just a static model when it is linked to physical asset data, a process, or a system [8]. Connected digital twins can represent infrastructure dependencies and enable scenario analysis before any changes are made [9]. However, most programs use digital twins for monitoring, prediction, and optimization. A digital twin provides advice in decision-making and rarely acts as an authorization gateway that determines the evidence upon which a decision can be implemented.

This paper poses the question: How can a government digital twin be designed as a pre-implementation governance mechanism for AI-powered public decisions? The paper proposes a Reversible Government Digital Twin (RGDT), a conceptual framework in which high-impact interventions move through a closed loop of monitoring, suggestion, simulation, challenge, authorization, limited implementation, observation, and reversal. The term "reversible government" does not imply the complete reversal of every public decision, but rather describes a methodology: the government should define the limits of reversibility before implementation, prioritize interventions with limited impact, maintain a state of security, identify compensatory measures, and present irreversible consequences to a higher authority and a stronger public justification.

This paper offers four contributions. First, it distinguishes between administrative, operational, and planning twins. Second, it presents a seven-tiered reference framework linking the civil state to public security. Third, it identifies five categories of reversibility and a Decision Reversibility Index (DRI) for comparing interventions. Fourth, it provides an implementation and evaluation agenda suitable for digital government programs at both the municipal and national levels. This contribution is conceptual and design-based; it is intended to be refutable and testable, not adopted simply because it is plausible.

Application Context

Research in digital twinning began primarily in engineering and manufacturing, where a digital representation remains linked to an observable element throughout its lifecycle. ISO 23247 provides the principles, reference architecture, and information-sharing requirements for twins in manufacturing [10, 11]. Recent studies address digital threads and the formation of integrated, unified, and interconnected twins [12, 13]. Although the standards are specific to particular fields, three ideas are directly applicable to the public sector: a twin requires a reliable relationship to the observable reality; the source must be continuously traced across the digital thread; and interconnected systems need clear rules for interoperability.

Public sector programs have expanded the concept from individual assets to systems of systems. The Gemini Principles in the UK frame digital twinning around the common good, value creation, and trust, with security, openness, and quality being key considerations [9, 14]. The National Vision for Digital Twins emphasizes a system of interconnected twins rather than a single, all-knowing model. European programs support local digital twins, interoperable data spaces, and testing facilities for smart communities [15, 16]. Similarly, the ITU's 2026 work on an AI-enabled "civic environment" integrates AI, spatial intelligence, digital twins, and connected infrastructure, with a focus on interoperability, human oversight, and data sovereignty [1, 2]. These initiatives establish a strong position for civic simulation, but their broad scope leaves room for a more nuanced, transactional governance model.

Government AI governance has developed in parallel. An OECD analysis indicates the expanding use of AI in government, particularly in internal processes, but with varying adoption rates across different government functions [17]. Its study of government AI use cases reveals significant activity in service automation and decision support, while a relatively small number of systems allow third parties to directly use government AI [18]. The National Institute of Standards and Technology (NIST) AI Risk Management Framework defines reliable and organizational risk management characteristics [4]. The EU AI Act adds binding obligations for specific high-risk uses, including technical documentation, records, accuracy, robustness, and human oversight [6]. These frameworks define what responsible governance should address, but operators still need to translate general requirements into evidence that supports concrete intervention under changing circumstances.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 3
Article Received: *25 September 2026* ***Revised:*** *12 October 2026* ***Accepted:*** *15 November 2026*

Human oversight alone is an unreliable safeguard when the user interface encourages automated bias, when the reviewer lacks time, or when responsibility is distributed. Therefore, NIST advises organizations to clarify human roles, document the appropriate level of involvement, and prepare responsible individuals to handle changes in model performance [5]. This supports the RGDT principle that oversight should not be a last-minute "apply" button. Reviewers should receive structured default evidence, information on uncertainties, affected groups, legal constraints, the feasibility of reversing, and a clear scope of authority.

Studies also reveal a gap between simulation and accountability. Digital twins can test alternatives, but high-fidelity simulations may include missing data, outdated sensors, or faulty causal assumptions. AI governance may require impact assessments, but the documentation prepared upon project approval may not reflect the conditions on implementation day. The RGDT system bridges these two aspects: each significant proposal generates an archived evidence package using the existing digital version, while governance controls define the required evidence and approvals.

Therefore, the novelty of the RGDT system lies in its operating model, not in any claim to invent digital twins, test environments, or AI safeguards. It integrates these elements into a pre-implementation control level and makes reversibility measurable. A literature search can reduce the potential for conceptual overlap, but it doesn't eliminate it entirely; justifiable originality depends on clearly defining the new structure, its outputs, and its testable propositions.

Research Methodology

This study employs a design-oriented research methodology. The outcome is a framework designed to address a specific governance problem: the ability of high-impact algorithmic actions to move from recommendation to implementation without specific intervention evidence related to safety, legality, fairness, or recovery. This framework was developed in three steps.

First, requirements were compiled from established standards and general guidelines covering digital twin architecture, AI risk management, high-risk AI commitments, digital government readiness, and responsible urban technology [1]-[18]. The compilation focused on recurring needs: reliable data, data source, interoperability, human accountability, testing, monitoring, contingency planning, transparency, and public value.

Second, these needs were transformed from lifecycle controls into transaction-level design requirements. For example, "human oversight" became a specific approval authority with clear mandate, evidence presentation, and a defined validity period. "Robustness" became scenario stress testing and post-deployment variance thresholds; "Contingency Planning" became a pre-authorized reversal or compensation procedure.

Third, the framework was analytically demonstrated through a designed municipal traffic scenario. This is an analytical study, not a pilot study. The scenario examines the framework's conceptual completeness and applicability by assessing whether it can represent affected groups, dependencies, uncertainties, authority, control, and reversal potential. Future research should assess accuracy, administrative burden, and outcomes through controlled pilot studies, test team exercises, and comparisons with existing change management processes.

Framework

The RGDT framework is defined as a dynamically updated, policy-sensitive digital representation of a public system, used to assess and approve critical interventions before their implementation, and to monitor, reverse, or compensate for their effects afterward. This framework differs from the traditional operational twin in four key ways:

First, it is intervention-oriented. The governance module is a proposed action with a defined objective, timeframe, authority, purpose, and expected impact. Second, it generates evidence. Each simulation creates an archived Decision Evidence Pack (DEP), not just a visual representation. Third, it is authority-oriented. Technical feasibility does not necessarily imply legal authorization; licensing rules and responsible roles are coded separately from the models. Fourth, it is recovery-oriented. Approval requires the identification of a declared reversal category, incentives, and a safe state plan or compensation plan.

The normative principle of this framework is simple: no proposed, automated, high-impact public action may be implemented in the physical or administrative environment until its anticipated value, uncertainties, impact on rights, risks of dependence, and recovery pathway have been assessed in proportion to its consequences. Exceptions remain possible, such as emergencies that threaten the safety of life, but these must be subject to an emergency authority mechanism, minimum safeguards, and retrospective review.

A. A Seven-Layer Structure

The first layer, the civil state, represents individuals, infrastructure, services, environmental conditions, budgets, policies, and relevant legal constraints. This layer must include not only average conditions but also the most vulnerable distributions and groups. For example, improving average travel time at the city level could mask significant delays around a hospital or reduced accessibility in a low-income area.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 3
Article Received: *25 September 2026* ***Revised:*** *12 October 2026* ***Accepted:*** *15 November 2026*

The second layer, the digital thread, maintains event time, data strings, sensor quality, semantic definitions, model versions, identity, and approvals. The work of the International Organization for Standardization (ISO) on digital threads and twin formation underscores the importance of lifecycle connectivity and unified interoperability [12, 13]. For governments, this system also requires data retention schedules, data purpose definitions, and access controls.

The third layer, "Models and Agents," includes predictive, optimization, causal analysis, generative, and rule components. Each component is logged with the intended use, owner, validation status, known constraints, and scope of permitted actions. Generative models may generate alternatives, but deterministic policy rules must impose non-negotiable constraints.

The fourth layer, "Simulation," implements baseline, proposed, and alternative scenarios. This layer should embody uncertainty rather than producing a single, convincing number. Required scenarios may include peak loads, sensor failures, cybersecurity degradation, unusual weather conditions, a major public event, and distributive effects across clusters or sites.

The fifth layer, "Reversibility," assesses whether an intervention can be stopped, recovered, or compensated for. This layer stores the last known safe state, dependencies, rollback orders, recovery time, and residual damage. The concept of reversibility is broader than just a technical reversal: if someone is unjustly deprived of a benefit, restoring the database record is insufficient unless the issues of delay, appeal, and financial consequences are addressed.

The sixth layer, the authorization layer, links risk and reversibility to the responsible authority. This layer establishes the principle of segregation of duties, approval criteria, time-limited authorizations, and limits of execution. The person authorized to approve must understand both public policy and evidentiary limitations; a technical operator cannot inherit a political mandate without their knowledge.

The seventh layer, the civil safeguards layer, evaluates legality, safety, fairness, accessibility, privacy, cybersecurity, public value, and appealability. At this layer, societal priorities and democratic constraints govern the improvement process. This layer provides appropriate general explanations while maintaining the security of sensitive and personal information.

B. The Cancelable Public Intervention Loop

The loop begins by monitoring the current situation and assessing the suitability of the data. An AI or human system then proposes an intervention with a clear objective and constraints. The system simulates the intervention based on baseline and stress scenarios. Independent challenge services test the model's assumptions, policy compliance, varying impacts, and security failure patterns. The system calculates risk and reversibility scores and develops an implementation plan. An authorized entity approves, modifies, rejects, or escalates the proposal. Implementation is authorized with a defined geographic scope, duration, maximum change limit, and stop conditions. Live telemetry data compares observed results with expected ranges. If any limit is exceeded, the system pauses, reverts, or initiates a compensation process. The results update the digital twin and the model log.

This loop redefines the concept of "human intervention." Humans are not expected to detect hidden model errors without assistance. The system organizes the evidence, highlights the variation between models, identifies the affected groups, and demonstrates reversibility limitations.

C. Decision Evidence Package

Each regulated intervention is assigned a unique identifier and a decision evidence package. At a minimum, the package must include: the overall objective; the initiating entity and responsible owner; the legal basis and policy constraints; the target audience and assets; a data snapshot and quality report; custom templates and versions where appropriate; a baseline and alternatives; outcomes of uncertainties in scenarios; distributive impact; cybersecurity and privacy checks; a reversibility category; a reversal or compensation plan; a consent log; authorization to implement; a comparison of direct outcomes; an incident log; and a final disposition.

The decision evidence package is machine-readable and human-reviewable. The brief consent presentation should be linked to more detailed technical evidence. Segmentation and annotation can only enhance integrity, but immutability must be reconciled with privacy law and retention rules. The goal is not random persistent storage, but a reliable source with structured deletion and structured access.

The goal is not random persistent storage, but a reliable source with structured deletion and structured access.

D. Five Categories of Reversibility

Category R0: Immediately reversible: The system can automatically revert to a safe state within minutes, with minimal residual damage. Temporary information display may fall into this category. Category R1: Operationally reversible: Restoration is reliable but requires controlled action or may take several hours. Category R2: Compensable reversible: The original state cannot be fully restored, but affected parties can be identified and substantially compensated through rectification, payment, priority, or appeal. Category R3: Partially reversible: Material effects remain after reversal, or

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 3
Article Received: *25 September 2026* ***Revised:*** *12 October 2026* ***Accepted:*** *15 November 2026*

some affected parties cannot be fully identified. Category R4: Functionally irreversible: Intervention may cause permanent material, legal, financial, or rights-related effects. Category R4 actions require the highest authority, public justification where legally permissible, narrower trials, and often a prohibition on automatic implementation.

This classification prevents a common category error: A configuration change may be technically reversible while its social consequences are irreversible. A road closure can be reversed by reopening it; however, an ambulance delayed during a closure cannot recover lost time. The classification should follow the least reversible physical consequences, not the easiest technical solutions.

E. Decision Reversibility Index

The proposed Decision Reversibility Index is a decision aid, not a universal measurement standard. The index assesses six dimensions from 0 to 5: feasibility of restoration (F), time to restoration (T), containment of the extent of damage (C), observability of damage (O), compensability (P), and isolation of dependency (D). Higher values indicate stronger reversibility. Because catastrophic residual damage should not be masked by the average, the penalty for the severity of residual damage, H, which also ranges from 0 to 5, is given greater weight:

$$DRI = 4(F + T + C + O + P + D) - 8H.$$

The initial range is -40 to 120. Authorities should calibrate the thresholds by domain and verify the consistency of the assessment. An explanatory policy could consider values from 80 to 120 as highly reversible, from 50 to 79 as conditionally reversible, from 20 to 49 as weakly reversible, and below 20 as indicating low or critically limited reversibility. These thresholds are illustrative and require domain-specific calibration and validation before operational use. A separate assessment of the risk of impact on the public is still necessary; high reversibility does not make an unlawful or unfair action acceptable.

F. Licensing Gateways

Gateway (a) validates purpose and mandate: Is the objective legitimate, and is automation appropriate? Gateway (b) validates twin suitability: Is the data up-to-date, representative, and sufficiently complete? Gateway (c) validates scenarios: Have realistic alternatives and pressure conditions been tested? Gateway (d) validates rights and public value: Who benefits, who bears the risks, and is there a less intrusive option? Gateway (e) validates reversibility: Can the intervention be paused, resumed, or compensated for within the required timeframe? Gateway (f) issues a limited implementation permit.

Gateway (g) closes the procedure only after reviewing the actual results, complaints, and residual effects.

Gates should only be automated if the rules are stable and can be independently audited. The framework intentionally separates "evidence generation" from "authorization." A vendor model may predict congestion, but it should not set the legal threshold for restricting access or approving its own proposal.

G. Security and Emergency Modes

The digital twin may fail due to outdated data, model deviation, a cyberattack, the unavailability of cloud computing services, or faulty assumptions. Therefore, the RGDT system needs a reduced mode with predefined controls: freezing new independent procedures; maintaining existing secure configurations; switching to conservative rules; alerting responsible operators; logging evidence gaps; and requiring explicit emergency authorization in exceptional cases. The system design must prevent the digital twin from becoming a single point of government failure.

Emergency procedures do not exempt one from accountability. When a delay poses a greater risk than the uncertainty of the action, the authorized entity may accept a reduction in the simulation depth. The system must record the reason, scope, and end date, maintain stop controls, and require retrospective evaluation. This balances speed and traceability rather than pretending emergencies allow perfect analysis.

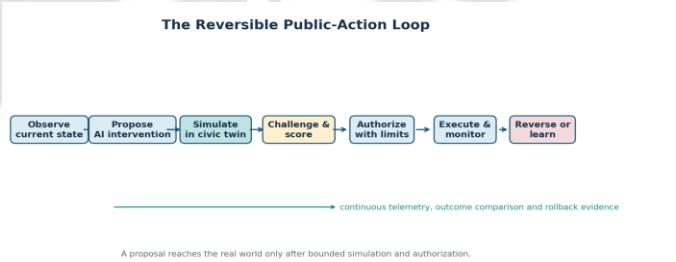


Figure 2. Reversible Public-Action Loop, from current-state observation through bounded execution and learning.

Table I. Reversibility Categories and Typical Authorization Levels

Class	Meaning	Primary test	Typical authority
R0	Immediate	Automatic safe-state restoration; negligible residual harm	System permit within policy
R1	Operational	Reliable controlled restoration within defined service time	Duty manager
R2	Compensable	State cannot fully return, but affected parties can be corrected	Service owner plus review

R3	Partial	Material residual effects or incomplete identification	Executive / statutory authority
R4	Irreversible	Enduring physical, legal, financial or rights effect	Highest authority; normally no autonomy

Practical Scenario: Adaptive Municipal Traffic Control System

All numerical values in the following scenario are illustrative assumptions used for analytical demonstration and do not represent empirical or pilot findings. Suppose a municipal operations center uses an AI-powered optimizer and a digital transport model to manage traffic signals across a main corridor. At 7:20 a.m., sensors detect an accident on a main road. The optimizer suggests extending the green light duration on a diversion lane by 18% and reducing the signal phases on side streets by two minutes for 45 minutes. The expected outcome is a 12% reduction in lane delays.

In a traditional system, an operator might view a dashboard and approve the recommendation. In an adaptive traffic control system, however, the suggestion becomes an intervention element. This intervention aims to reduce overall pedestrian delays without increasing emergency response times, increasing pedestrian risks, or impacting access to residential areas beyond defined limits. The implementation scope specifies 14 intersections, a maximum phase change of 18%, a 45-minute expiration period, and a mandatory five-minute review period.

The digital model first verifies the system's readiness. Two of the traffic sensors were found to be outdated, and pedestrian statistics at one school crossing were only available from the previous week. The fitness service reduced confidence and required a scenario that doubled the historical pedestrian demand. The digital system logged the missing data, so this limitation could not be hidden from the approval screen.

The simulation compared four alternatives: no action; an AI suggestion; a 10% phase change; and a route-only option. Each alternative was tested under the current incident estimate, a 30% shortfall in side street traffic statistics, high pedestrian demand, an ambulance dispatched from a nearby station, and the loss of one signal controller. The AI suggestion performed best in terms of average delay but increased pedestrian waiting time (90%) near the school beyond the policy. It also created a small probability of delaying the ambulance route.

The challenge layer rejected the original suggestion at Gate D. The optimizer produced a constrained alternative: a 12% phase change at 11 intersections, excluding the school crossing, and signal priority for emergency vehicles. The expected lane improvement drops from 12% to 8%, but all policy thresholds are met. This illustrates a core principle of the Reversible Government Digital Twin (RGDT): overall improvement should aim for acceptable action, not just numerical maximization.

The reversal layer defines the previous signal plan as the safe state. Recovery is automatic, expected within two cycles, and verified by controller notifications. The affected area is geographically limited; damage can be observed through traffic congestion, pedestrian movement, and emergency data. However, the emergency delay has a high residual severity. The proposal achieves a high score in most dimensions of the Decision Reversibility Index (DRI) but receives a material penalty of H. It is rated R1 instead of R0 and requires the assignment of a duty manager.

The Decision Evidence Package (DEP) provides the manager with the baseline, restricted options, uncertainties, affected locations, failed tests, DRI, rollback orders, and stop triggers. Approval issues a signed permit valid for 45 minutes. Expansion to additional intersections is prohibited without a new Decision Evidence Package and authorization. During implementation, traffic congestion on one side street exceeds the expected range by 95% for two consecutive periods. The system automatically reverts two intersections to the previous safe-state configuration while keeping the remaining nine active. The system records a partial deceleration rather than deeming the entire intervention a success or failure.

Upon closure, the actual lane delay improved by 6%. Pedestrian waiting times remained consistent, ambulance access was unaffected, and one side street experienced an additional temporary delay of four minutes. This result updates the calibration data. A public transparency record can illustrate the objective, geographic scope, duration, key safeguards, and observed outcome without disclosing sensitive operational details.

This scenario demonstrates six advantages: it prevents the city-wide average from obscuring any violation of local policy; it transforms missing sensor data into an explicit consent requirement; it compares a limited alternative rather than framing the decision as AI versus inaction; it limits authority in terms of location, scale, and time; it supports partial rollback; and finally, it preserves evidence for auditing and learning purposes.

The same pattern can be adapted to responding to energy demands, distributing water, locating emergency resources, and scheduling maintenance. This approach should be applied with caution to decisions concerning individuals, such as eligibility,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 3
Article Received: *25 September 2026* ***Revised:*** *12 October 2026* ***Accepted:*** *15 November 2026*

enforcement, or child protection, where population-level twinning may be ethically inappropriate, and due process cannot be reduced to mere simulation. In these areas, RGDT technology can test workflows and resource policies while maintaining individual segregation within specialized legal safeguards.

Implementation Roadmap and Maturity Model

Implementation should begin with a limited-scale, observable, and correctable service, rather than creating a replica for an entire city. The pilot phase, lasting 12 to 18 months, can be followed by five stages.

Stage 1: Defining the scope of the decision. The stakeholder selects the type of intervention, identifies the responsible authority, defines legal constraints, identifies the affected groups, and establishes a baseline performance level and incident log. This stage also defines what the digital version will not do.

Stage 2: Building the minimum viable digital version and digital thread. This does not aim for hyperrealism, but rather for a sufficient level of accuracy to compare alternatives, preserve the data source, and detect any discrepancies. User interfaces should be modular so that models and vendors can be changed.

Stage 3: Developing quality assurance scenarios and corrective mechanisms. Operators, domain experts, cybersecurity personnel, legal advisors, and community representatives contribute failure scenarios. The rollback capability is tested in a non-production environment and then through controlled drills.

Stage 4: Pilot Mode. The twin system generates data implementation proposals and plans, while the human element continues with existing operations. Expected and observed results are compared to measure calibration accuracy, false alarms, audit load, and missed scenarios. No autonomy should be granted simply because a pilot period has ended.

Stage 5 enables limited implementation for a limited class with the possibility of reversal. Authorizations begin on a small scale and for a short duration. Scaling up requires evidence of stable calibration, successful reversals, acceptable distribution results, and a manageable supervisory workload.

Six levels support the maturity of the evaluation process. Level 0, observable, includes dashboards and historical analysis, but does not include intervention simulation. Level 1, simulated, compares alternatives, but does not produce standardized governance evidence. Level 2, managed, creates data implementation plans and requires specific approval. Level 3, reversible, implements limited interventions with tested reversals, variance monitoring, and compensation. Level 4, Unified, coordinates between agencies or twin systems using shared semantics, shared authority, and cross-domain recovery rules. Level 5, Adaptive Accountability, learns from results and proposes new scenarios while maintaining democratic and human oversight of thresholds, mandates, and irreversible procedures.

Key evaluation indicators should include more than just operational efficiency: simulation calibration error; the proportion of high-impact actions that include complete implementation plans; the proportion of interventions modified or rejected by reviewers; average approval time; reversal success rate; average time to reach a safe state; rate of unnoticeable damage; variance between expected and actual outcomes by affected group; number of complaints and resolution time; violations of permit expiration dates; completeness of the model or data source; and vendor reliance concentration.

A pilot project is not considered successful if the system never disrupts operations. An effective reviewer should sometimes reject attractive proposals. Agencies should monitor the "safety theater," where evidence packages become large but ineffective, and "threshold drift," where performance pressures gradually weaken constraints. Therefore, independent review of threshold changes is essential.

Conclusion

As governments adopt AI-powered digital twin technologies, the central question should shift from "Can the system improve this service?" to "What needs to be proven before allowing this intervention to impact the public?" The "Reversible Government with Digital Twins" model answers this question by transforming the civic twin into a pre-implementation governance environment. It connects the current state, data source, models, virtualization, rights analysis, power, and recovery in a single loop to generate evidence.

The most important aspect of this framework is its institutional, not computational, nature: the correction must be designed before any action is taken. A government that knows how to stop, reverse, compensate for, and explain an intervention is better prepared than one that relies on post-implementation audits. Reversal technology cannot eliminate political discretion, uncertainty, or irreparable harm. But it can reveal these limitations early enough to change the decision.

The next step is controlled experimentation in specific, observable areas, with shadow-mode operation, independent challenge, public interest scenarios, and published evaluation. If this proves to be true, rollback capability could become a standard quality attribute for digital government, alongside security, privacy, interoperability, accessibility, and resilience.

References

1- International Telecommunication Union, AI-Enabled Citi verse: A Strategic Blueprint for Cities in the Age of AI, 2026.
2- International Telecommunication Union, Autonomous Cities and AI: The Next Frontier of Urban Transformation, 2026.
3- Saudi Digital Government Authority, Emerging Technologies Adoption Readiness in Government Entities 2026, 2026.
4- National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, 2023.
5- NIST AI Resource Center, Appendix C: AI Risk Management and Human-AI Interaction, 2023.
6- European Parliament and Council, Regulation (EU) 2024/1689 laying down harmonized rules on artificial intelligence, 2024.
7- Saudi Data and AI Authority, Generative Artificial Intelligence Guidelines for Government, 2024.
8- UK Infrastructure and Projects Authority, Transforming Infrastructure Performance: Roadmap to 2030, 2021.
9- Centre for Digital Built Britain, The Gemini Principles, 2018.
10- International Organization for Standardization, ISO 23247-1:2021, Automation systems and integration - Digital twin framework for manufacturing - Part 1: Overview and general principles, 2021.
11- International Organization for Standardization, ISO 23247-4:2021, Digital twin framework for manufacturing - Part 4: Information exchange, 2021.
12- International Organization for Standardization, ISO 23247-5:2026, Digital twin framework for manufacturing - Part 5: Digital thread, 2026.
13- International Organization for Standardization, ISO 23247-6:2026, Digital twin framework for manufacturing - Part 6: Digital twin composition, 2026.
14- UK Government, National Data Strategy, 2020.
15- European Commission, Smart and Sustainable Cities and Communities (RP 2026), Interoperable Europe Portal, 2026.
16- European Commission, Local Digital Twin Toolbox, Interoperable Europe Portal, 2026.
17- OECD, Digital Government Outlook 2026: Adopting and Governing AI in Government, OECD Publishing, 2026.
18- OECD, Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions, OECD Publishing, 2025.