

Who gets to verify AI? Provenance, authority and accountability in digital government

Shuai Wang

Third Affiliated Hospital of Inner Mongolia Medical College

Suqin Wei

Baise University

Peilin Li

Kuala Lumpur University of Science & Technology

Sai-Keong Chan

Kuala Lumpur University of Science & Technology

Yibing Liu

879729144@qq.com

Xi'an Siyuan University

Article

Keywords: artificial intelligence, content provenance, digital government, sociotechnical systems, transparency, accountability, misinformation, politics of truth

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10445068/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Artificial intelligence (AI) is increasingly used to produce official notices, images and other communication in government settings, and regulators, standards bodies and technology firms have responded mainly through labelling and provenance requirements. Existing research on AI-content labelling, however, is dispersed across human–computer interaction, technical and legal literatures, and rarely treats labelling and provenance as a single governance system that distributes verification work and authority across citizens, platforms, providers and regulators. This article asks how AI-content transparency requirements are structured across the content lifecycle, and under what conditions they produce accountability rather than symbolic compliance. A structured comparative document analysis was conducted on 11 Chinese, European and transnational regulatory, technical and management instruments, including China’s labelling rules and mandatory standard, the EU AI Act, the C2PA specification, the NIST AI Risk Management Framework and relevant ISO standards, drawing on science and technology studies, infrastructure studies and algorithmic-accountability research. The analysis distinguishes three transparency functions that cannot substitute for one another: public legibility, infrastructural verifiability and institutional answerability. Visible labels can shift the work of verification onto citizens; provenance systems can concentrate epistemic authority in platforms, vendors and standards bodies; and audit records support accountability only when institutions have access, interpretive capacity and the authority to act. These functions are connected through a five-stage lifecycle model covering generation, publication, distribution, review and regulation, and six recurring failure modes explain how formal compliance can coexist with system-level failure. The article contributes a framework that treats content provenance as a governance infrastructure rather than a neutral technical fix, and specifies the institutional conditions under which AI-content labelling can support, rather than only signal, public accountability.

Introduction

Generative artificial intelligence has entered the communicative infrastructure of government. Public agencies use generative systems for drafting, translation, service chatbots, policy communication, image production and internal knowledge work. These applications can reduce production costs and expand communicative capacity, but they also make it easier to produce plausible official notices, synthetic representations of public officials, fabricated emergency information and administrative guidance whose origin becomes uncertain after editing and cross-platform circulation. In public administration, the problem is not limited to whether a media object is technically synthetic. Government information carries institutional authority, and citizens may act on it before verification is possible. As a result, synthetic media can affect public safety, access to services, institutional legitimacy and the practical conditions under which citizens decide what counts as an authoritative statement (Mergel et al. 2019; Wirtz et al. 2019; Zuiderwijk et al. 2021).

Labelling has become a prominent response. Regulatory initiatives increasingly require visible disclosure, machine-readable marking or both. Technical coalitions promote cryptographically signed

content credentials, while risk-management frameworks emphasise traceability, record keeping and organisational oversight. These measures are often narrated as a straightforward progression from opacity to transparency: identify synthetic content, attach a label, preserve provenance and restore trust. Empirical research, however, complicates this account. AI labels can improve recognition without consistently changing perceived accuracy or sharing intention (Li and Yang 2024); warning labels often leave most users unable to identify a deepfake (Lewis et al. 2023); and provenance information can help calibrate trust while also causing users to confuse the credibility of a media object with the credibility of its provenance record (Feng et al. 2023). Design choices such as wording, iconography and persistence further shape how warning labels are noticed and trusted, with no single design performing best across platforms and content types (Gamage et al. 2025).

Deepfakes may generate generalised uncertainty rather than simple deception, producing cynicism towards authentic as well as manipulated media, an effect documented specifically in electoral contexts (Vaccari and Chadwick 2020; Diakopoulos and Johnson 2021; Fallis 2021; Liu et al. 2025). These findings suggest that labelling is not merely an information-design problem. Labels classify content, define which forms of production require disclosure and assign interpretive work to particular actors. Provenance standards establish who may attest to origin, which events are recorded, whose identity is recognised and which transformations count as legitimate. Audit systems determine who can access evidence and what consequences follow when rules fail. Each layer thus embodies choices about authority, responsibility and the politics of truth. Transparency can reveal selected features while concealing the institutional relations through which those features are produced and interpreted (Ananny and Crawford 2018).

Critical AI social science has challenged two persistent binaries in public and academic debate: the separation of technological and social determinants, and the opposition between optimism and pessimism about AI. AI-content labelling provides a productive case for moving beyond both. Neither labels nor social trust exist independently of the institutions, standards, interfaces and practices that connect them. A visible notice is generated through classification rules, platform design and enforcement choices; machine-readable provenance depends on installed technical infrastructures and organisational adoption; and accountability emerges only through procedures that enable contestation, evidence access and responsibility allocation. The relevant object of analysis is therefore not an isolated label but the sociotechnical system through which public information becomes legible, verifiable and answerable.

Research on AI labelling remains fragmented. Human–computer interaction studies examine label wording, salience and user interpretation; technical studies focus on watermarking, metadata and content credentials; legal research analyses disclosure duties and liability; and public-administration scholarship examines trustworthy AI, administrative discretion and institutional capacity (Bullock 2019). A broader AI-governance literature maps ethical principles, institutional functions and cross-jurisdictional arrangements, but rarely engages content provenance as a distinct governance object in its own right (Floridi et al. 2018; Taeihagh 2021; Batool et al. 2025). What is missing is a system-level social analysis

of how these functions redistribute epistemic authority and verification burdens across citizens, platforms, providers, standards bodies and regulators. This gap is particularly important in digital government, where source authenticity, administrative authority and public trust are tightly connected.

This gap motivates three questions that guide the analysis. How do labelling and provenance regimes construct public visibility and institutional authority around AI-generated government content? How are human-facing disclosure, machine-verifiable provenance and regulatory evidence connected across the content lifecycle? And under what conditions does transparency produce operational accountability rather than symbolic compliance?

The article makes three contributions. Conceptually, it distinguishes public legibility, infrastructural verifiability and institutional answerability as related but distinct transparency functions that cannot substitute for one another. Analytically, it connects these functions through a lifecycle model and identifies failure pathways through which local compliance can produce system-level breakdown. Practically, it specifies the institutional conditions under which AI-content labelling can support, rather than only signal, accountable public communication. The framework does not claim that labelling alone produces trust; it identifies what else is required for labelling to function as intended.

Theoretical approach: from disclosure to sociotechnical accountability

AI labelling as classification and infrastructure

Science and technology studies treats technological order and social order as co-produced rather than as separate domains (Jasanoff 2004). Sociotechnical imaginaries connect visions of desirable futures to institutional programmes, material systems and forms of political authority (Jasanoff and Kim 2009). Applied to AI-generated content, a labelling regime embodies an imaginary of governable authenticity: if origin and transformation can be recorded, public communication can be made inspectable and trust can be stabilised. Yet the categories that make content inspectable—AI-generated, AI-assisted, manipulated, verified, suspicious or authentic—are not discovered in nature. They are negotiated classifications that simplify heterogeneous production practices and attach consequences to those categories.

Classification systems are simultaneously informational and political. They become embedded in standards, interfaces and routines, often disappearing from view when they work smoothly (Bowker and Star 1999). Infrastructure is relational: it functions across communities only when it connects to installed systems, conventions, skills and organisational practices, and it becomes most visible during breakdown (Star and Ruhleder 1996). Provenance is consequently not a self-contained metadata object. It is an infrastructural accomplishment involving capture devices, model providers, publishing systems, platforms, trust lists, identity services, verification interfaces and public institutions.

A sociotechnical approach also cautions against abstraction. Technical interventions that isolate one component from the organisational context can create portability and solutionism traps (Selbst et al. 2019). A watermark may be robust in a laboratory but irrelevant if platforms discard it; a content credential may be cryptographically valid but socially misleading if users interpret provenance as proof of truth; and an audit log may be complete yet politically ineffective if regulators lack access or authority. The unit of governance is therefore the connected arrangement, including its exclusions and maintenance work.

Transparency, visibility and answerability

Transparency is frequently presented as an intrinsic good, but visibility does not automatically generate understanding or control. Ananny and Crawford (2018) show that seeing components of an algorithmic system may fail to reveal the relationships, histories and institutional capacities needed for accountability. Algorithmic accountability instead requires clarity about who must give an account, to whom, for what object, through which forum and with what consequences (Kroll et al. 2017; Wieringa 2020). Transparency is thus relational and audience-specific: even accurate disclosure fails to inform action when no audience is positioned, resourced or authorised to use it (Kemper and Kolkman 2019).

These distinctions can be organised into a simple model rather than a loose set of categories. Three transparency functions are distinguished, each defined by a different audience, a different type of evidence and a different governance question. Public legibility asks whether citizens can recognise a content object's production status and know how to act on it; its evidence consists of visible labels, risk cues and source explanations. Infrastructural verifiability asks whether platforms and authorised third parties can inspect origin, transformation and credential continuity; its evidence consists of metadata, watermarks and cryptographic credentials. Institutional answerability asks whether evidence can be used to reconstruct decisions, contest outcomes and assign responsibility; its evidence consists of actor identifiers, event logs and review records. This differentiation extends recent proposals for contextual transparency, which similarly argue that what should be disclosed, to whom and for what purpose varies systematically across decision settings rather than following a single universal standard (Sloane et al. 2023).

The model's central claim is that these three functions are coupled rather than additive: each is necessary, but none is sufficient on its own, and expanding one function cannot compensate for weakness in another. A label without machine-readable evidence gives platforms nothing to check. Verifiable evidence without an accessible forum gives citizens and regulators nothing to act on. And an accessible forum without evidence cannot assign responsibility to anyone. The three functions are therefore best understood not as separate types of transparency but as interdependent parts of a single accountability system, connected through the practical work of the actors who must recognise, inspect and act on the same underlying claim about a content object's origin.

The distinction also directs attention to burden distribution. Disclosure policies can relocate responsibility from powerful organisations to individual users by telling citizens to be vigilant without providing accessible verification or remedies. Similarly, audit requirements can become rituals that demonstrate procedural compliance while leaving organisational incentives and power relations unchanged (Power 1997; Metcalf et al. 2019). The critical question is not how much information is disclosed, but whether the arrangement enables meaningful recognition, contestation and institutional response.

Provenance and the politics of truth

Synthetic media threatens public knowledge not only by introducing false objects but by weakening confidence in evidentiary practices. Deepfakes can reduce the information that recordings carry and make authentic evidence easier to deny (Fallis 2021). This “liar’s dividend” allows actors to dismiss inconvenient authentic media as fabricated (Chesney and Citron 2019). In this environment, provenance promises to document how an object was created and transformed. Yet provenance is not truth. A valid credential can establish a history or source relationship while the substantive content remains inaccurate, selective or misleading.

Labelling research reinforces this distinction. Process-based labels communicate that AI was used; impact-based warnings communicate that content may mislead. The two objectives may require different language, enforcement and distribution responses (Wittenberg et al. 2024). Labels can reduce belief in some misleading content but can also stigmatise accurate AI-assisted material, normalise generalised suspicion or generate an implied-authenticity effect for unlabelled content. The design of labelling regimes thus participates in the politics of truth by shaping which signals count as credible and which actors are authorised to produce them.

Authority is also unevenly distributed. Global provenance systems are developed largely by powerful technology firms and standards organisations, echoing broader concerns that consequential decision infrastructures are governed through proprietary logics that resist public scrutiny (Pasquale 2015). Governments with limited technical capacity may become dependent on foreign vendors, certification infrastructures and trust lists. Critical AI scholarship warns that universalised standards can reproduce geopolitical and epistemic asymmetries when local institutions and plural knowledge practices are treated as implementation contexts rather than as sources of design authority (Couldry and Mejias 2019; Mohamed et al. 2020). A critical account of provenance must therefore examine not only technical continuity but also who controls the infrastructure of authentication.

Materials and methods

Research design and unit of analysis

The study uses a structured qualitative comparative document analysis combined with theory-to-design synthesis. The purpose is not to estimate a causal effect of labels on trust or to benchmark a detection algorithm. It is to identify recurrent governance functions across heterogeneous regulatory and technical instruments, analyse how those functions distribute visibility and authority, and construct a framework that can guide future empirical research in digital-government settings.

This design is appropriate to the research questions for three reasons. The questions concern how governance functions are distributed and connected across an institutional field, not how individual users respond to a specific label; authoritative regulatory and technical texts document this directly, whereas user studies or detection benchmarks would answer a different question. A second consideration is comparative scope: a single-jurisdiction case study could describe one regime in depth but could not show which functions recur across regulatory paradigms and which are specific to one institutional tradition, so comparison across rule-based, risk-based, technical-standard and management-system instruments is needed to separate the two. A third consideration follows from the state of the literature itself, which currently offers separate functional descriptions of labelling, provenance and audit rather than an integrated account of how they relate to one another; a qualitative, theory-building design is better suited to this stage of the research than a hypothesis-testing design, which would require a framework that does not yet exist.

The unit of analysis is the governance function performed by a rule, standard, technical mechanism or organisational control. Examples include user-facing disclosure, machine-readable marking, provenance preservation, identity binding, audit logging, risk classification, human review, appeal, regulatory access and performance evaluation. This functional unit allows legal obligations and technical specifications to be compared without treating them as equivalent institutions.

Document corpus and inclusion logic

The core corpus comprises 11 authoritative instruments that directly specify obligations or management requirements relevant to AI-generated content disclosure, provenance, risk management or organisational oversight. The Chinese corpus includes the Measures for Labelling AI-Generated and Synthetic Content (Cyberspace Administration of China et al. 2025), the mandatory national standard GB 45438 – 2025 (State Administration for Market Regulation and Standardization Administration of China 2025), the Interim Measures for the Management of Generative AI Services (Cyberspace Administration of China et al. 2023), the Provisions on the Administration of Deep Synthesis in Internet-Based Information Services (Cyberspace Administration of China et al. 2022b) and the Provisions on the Administration of Algorithmic Recommendations (Cyberspace Administration of China et al. 2022a). The comparative corpus includes the EU AI Act (European Parliament and Council of the European Union 2024), the C2PA Technical Specification 2.1 (Coalition for Content Provenance and Authenticity 2024), the NIST AI Risk Management Framework and Generative AI Profile (National Institute of Standards and Technology 2023; Autio et al. 2024), and ISO/IEC 42001 and ISO/IEC 23894 (International Organization for Standardization 2023a, 2023b).

Documents were included when they met three criteria: they were issued by a public regulator, recognised standards body or widely adopted technical coalition; they contained operational provisions concerning disclosure, provenance, risk control, records or accountability; and they contributed a distinct institutional logic. The corpus was designed for analytical variation rather than exhaustive global coverage: China and the European Union provide prominent and institutionally distinct regulatory approaches to AI-content provenance, while C2PA, NIST and ISO/IEC instruments supply transnational technical and management-system logics that interact with national regulation. Together the instruments represent rule-based labelling, risk-based law, cryptographic provenance and organisational management-system approaches, summarised by material group in Table 1.

Table 1
Core corpus and analytical purpose

Material group	Representative instruments	Primary analytical purpose
China: labelling rules	AI-generated and synthetic content identification measures; GB 45438 – 2025	Visible and implicit marking duties, metadata fields, provider and platform responsibilities
China: adjacent platform governance	Generative-AI service measures; deep-synthesis and algorithmic-recommendation provisions	Provider duties, distribution governance and wider platform responsibilities
European Union	Regulation (EU) 2024/1689 (AI Act)	Risk-based transparency duties, machine-readable marking, oversight and enforcement
Content provenance	C2PA Technical Specification 2.1	Credentials, manifests, signatures, editing history, validation and trust mechanisms
Risk and management systems	NIST AI RMF 1.0; NIST AI 600-1; ISO/IEC 42001; ISO/IEC 23894	Risk identification, role allocation, monitoring, records and continual improvement

Coding and analytical procedure

Analysis proceeded in four stages. The first stage involved reading each instrument for explicit governance functions and the actors, information objects, interfaces and purposes associated with them. The second stage grouped recurrent functions according to their primary audience: citizens and content recipients; platforms and technical verifiers; or auditors, regulators and compliance units. The third stage mapped the functions to five lifecycle stages—generation, publication, distribution, review and regulation—to identify handoffs and dependencies. The fourth stage coded failure modes where information, authority or evidence could be lost, inaccessible or detached from consequences.

The comparison used six dimensions: risk differentiation; human-facing disclosure; machine-readable marking and provenance; interoperability across transformations and platforms; responsibility and auditability; and public verification and performance feedback. Four critical questions were applied

across the dimensions: What becomes visible? To whom? Who controls the relevant evidence or classification? What remedy or consequence follows when the system fails? A code was retained only when it could be connected to a defined actor, evidence object or interface and a recognisable governance purpose. The resulting codebook, with guiding questions and illustrative evidence for each dimension, is set out in Table 2.

The analysis was interpretive and iterative rather than a statistical reliability exercise. No intercoder-reliability coefficient was calculated. Conceptual discipline was supported through a documented codebook, negative-case checks and a traceability matrix linking analytical claims to instrument functions. The framework was stress-tested against boundary conditions, including low-risk creative content, high-stakes government communication, benign loss of provenance and deliberate removal. The coding protocol, claim-to-evidence traceability matrix and negative-case checks are provided in Supplementary Tables S2–S4.

Table 2
Analytical codebook

Dimension	Guiding question	Illustrative evidence objects	Critical concern
Risk differentiation	How is control intensity matched to expected public consequences?	Risk tiers, high-risk scenario lists, escalation rules	Overcontrol of low-risk content and undercontrol of authoritative or emergency information
Public disclosure	What can a recipient notice and understand?	Visible labels, wording, source explanations, verification links	Burden shifting, ambiguity, stigma and warning fatigue
Provenance	What can machines and verifiers inspect?	Metadata, watermarks, content IDs, credentials and signatures	Metadata loss, false confidence and dependence on proprietary infrastructure
Interoperability	What survives editing and cross-platform transfer?	Core schemas, transformation records, APIs and trust lists	Fragmentation, installed-base incompatibility and exclusion
Responsibility	Who must give an account and on what evidence?	Actor IDs, event logs, review and appeal records	Responsibility diffusion and inaccessible audit evidence
Feedback	How are errors, disputes and incidents used to revise governance?	Appeals, audit findings, incident records and performance measures	Static compliance and measurement of outputs rather than outcomes

Use of generative AI in manuscript preparation

During manuscript preparation, the authors used OpenAI ChatGPT to support English-language editing, structural refinement and editorial polishing. The tool was not used to select the document corpus, assign analytical codes, generate empirical data or determine the study findings. All AI-assisted text and references were checked against the underlying sources and substantively revised by the authors, who take full responsibility for the manuscript. The two schematic figures were redrawn by the authors using Python and Matplotlib; no generative AI image tool was used.

Results

Public legibility: labels make production visible but can relocate responsibility

Across the corpus, human-facing disclosure is treated as the first line of transparency. Visible labels, notices and risk cues are intended to lower recognition costs by enabling recipients to identify content as AI-generated or materially manipulated. In high-authority contexts such as official notices, emergency alerts or realistic depictions of public officials, persistence, placement and proximity to the content matter. A technically correct notice that is transient, ambiguous or easily cropped may satisfy a narrow duty while failing to support public understanding.

The social function of a label, however, is not exhausted by recognition. A label defines what kind of object the recipient is encountering and implies what the recipient should do next. Process labels ask users to interpret the significance of AI involvement; risk labels communicate that the content may mislead; source explanations invite verification. These functions can conflict. A neutral AI-generated label may improve process awareness while saying little about truth, whereas a warning label may reduce credibility even when the content is accurate. The corpus rarely specifies how these competing interpretations should be evaluated in different cultural and administrative contexts.

Visible disclosure also redistributes verification work. When platforms display a label without an accessible source-checking route, responsibility is implicitly transferred to citizens who may lack time, technical literacy or access to authoritative records. This is especially problematic in emergency communication, where rapid action is expected. For this reason, public legibility is meaningful only when labels connect to practical verification and correction pathways. Otherwise, disclosure can function as liability management: the organisation signals that it warned the user while leaving the underlying information asymmetry intact.

Infrastructural verifiability: provenance enables inspection but concentrates authority

Machine-readable provenance forms the second transparency function. Metadata, digital watermarks, content identifiers, cryptographic signatures and content credentials allow platforms and authorised verifiers to inspect origin, editing history and credential continuity. This layer addresses a limitation of

visible labels: front-end notices can be removed or separated from the media object, whereas embedded or linked evidence can support automated verification across workflows.

Cross-platform circulation is the principal stress point. Screenshots, screen recordings, compression, file conversion, editing and re-uploading can strip metadata or break credential bindings. Platform-specific schemas and verification interfaces fragment the chain. The instruments therefore converge on a need for coordination between visible and implicit signals, but they differ in the institutional mechanisms available to achieve it. Regulation can mandate duties; technical specifications can define representations; management standards can require controls; none alone guarantees adoption across the installed base.

Provenance also creates new centres of epistemic authority. Trust lists, signing certificates, identity providers and validation software determine which claims are recognised. Large platforms and technology vendors may control the interfaces through which credentials are created or displayed, a pattern consistent with wider evidence that platforms act as de facto custodians of what counts as visible and legitimate content (Gillespie 2018). Public agencies that depend on those infrastructures may be able to verify content only on terms set by private actors. The technical achievement of verifiability can therefore coexist with political dependence. The question is not only whether a credential validates, but who can issue, revoke, interpret and contest it.

The analysis further confirms that provenance and truth must remain analytically separate. A credential can demonstrate that an official agency published an image while leaving open whether the accompanying statement is substantively correct. Conversely, absence of provenance is not proof of deception because evidence can be lost through benign transformations, legacy systems or accessibility practices. Treating missing provenance as conclusive would enable automated exclusion and create disadvantages for smaller organisations and users without compliant tools, reproducing a pattern in which automated eligibility and verification systems disproportionately burden already disadvantaged groups (Eubanks 2018).

Institutional answerability: evidence becomes accountability only through forums and consequences

The third function is institutional answerability. Its objects include actor identifiers, generation records, publication declarations, verification events, transformation records, moderation decisions, human-review notes, appeals and regulatory access logs. These records can reconstruct what happened, who exercised control at each stage and whether an intervention was timely and proportionate.

Record creation alone is insufficient. Accountability requires an institutional forum with authority to request evidence, interpret it, hear contestation and impose remedies or sanctions. A platform may retain detailed logs that citizens cannot access; an auditor may obtain records but lack power to compel correction; a regulator may have legal authority but insufficient technical capacity to assess credential

chains. In each case, the presence of data can produce an appearance of accountability without answerability.

The multi-actor structure of digital government intensifies this problem. Model providers, application developers, public communication units, social platforms, verification services, auditors and regulators control different parts of the lifecycle. When harmful content circulates, each actor can point to compliance within its local boundary. The provider generated a marker, the publisher displayed a notice, the platform applied its policy and the regulator issued guidance. Yet if evidence does not transfer across boundaries, no actor can reconstruct the failure pathway. Accountability depends on institutional handoffs as much as on technical logs. Table 3 summarises the audience, core objects, social achievement and characteristic failure of all three transparency functions.

Table 3
Triple transparency as a sociotechnical arrangement

Function	Primary audience	Core objects	Social achievement	Characteristic failure
Public legibility	Citizens and content recipients	Visible labels, risk cues, source explanations and verification links	Makes production status recognisable and potentially actionable	Ambiguity, warning fatigue, burden shifting and implied authenticity of unlabelled content
Infrastructural verifiability	Platforms, automated systems and third-party verifiers	Metadata, watermarks, content IDs, credentials, signatures and APIs	Makes provenance inspectable across workflows and transformations	Metadata loss, incompatible schemas, vendor dependence and confusion of provenance with truth
Institutional answerability	Regulators, auditors, courts and compliance units	Actor IDs, event logs, review records, appeals, access logs and audit reports	Makes decisions reconstructable, contestable and attributable	Inaccessible evidence, weak interpretive capacity, responsibility diffusion and absent consequences

The lifecycle responsibility loop

The three transparency functions operate across five connected stages. At generation, model or application providers attach required markers and preserve creation time, content identifiers and provider information. At publication, public agencies and platform operators collect declarations, validate available provenance and determine how source status should be displayed. During distribution, platforms attempt to preserve or reconnect provenance through metadata protection, content-ID mapping, credential verification or transformation records. At review, automated signals and human

judgement are combined, especially for high-impact public information. At regulation, logs and evidence chains support inspection, responsibility allocation and revision of future controls.

The central system property is coupling. A label generated at source but never checked at publication is inert. A platform verification that is not preserved after transcoding cannot support cross-platform accountability. An audit record that cannot be accessed through defined procedures remains nominal. The system therefore depends on both technical interfaces and organisational handoffs.

Figure 1 represents the architecture as a set of non-substitutable transparency functions connected to the content lifecycle. It also makes visible the critical risks associated with each layer: burden shifting at the public interface, infrastructural dependence in provenance systems and responsibility diffusion in institutional processes. Feedback from disputes, appeals, incidents and audit findings is necessary to revise categories, interfaces and control intensity. Table 4 sets out the core actors, primary responsibility and evidence or interface associated with each stage.

[Figure 1 about here]

Table 4
Lifecycle responsibilities and evidence requirements

Stage	Core actors	Primary responsibility	Evidence or interface
Generation	Model and application providers	Apply required marks; assign identifiers; preserve creation records	Metadata, watermark, content ID, timestamp and generation record
Publication	Public-sector publishers and platform operators	Validate evidence; display disclosure; record declarations and exceptions	Validation result, display record, publisher identity and trusted-source status
Distribution	Platforms and distribution channels	Preserve or reconnect provenance after transfer and transformation	Transformation record, platform ID, verification API and recovery link
Review	Moderation teams, public agencies and third-party verifiers	Combine automated checks with human review; record decisions and appeals	Detection result, review note, handling decision and appeal file
Regulation	Competent regulators and authorised auditors	Inspect, compel evidence, allocate responsibility and feed findings back into rules	Audit report, access log, enforcement record and evidence-chain procedure

Failure pathways: how formal compliance becomes symbolic compliance

Six recurring failure modes emerge from the comparison. Coarse risk classification applies similar controls to low-impact creative content and high-authority government communication, wasting capacity while leaving high-consequence scenarios underprotected. Label–provenance disconnection separates human disclosure from machine evidence, allowing visible compliance without inspectable origin.

Interoperability loss breaks continuity during editing and cross-platform transfer. Inaccessible verification leaves citizens unable to resolve uncertainty. Fragmented responsibility enables actors to shift blame across organisational boundaries. Output-oriented evaluation counts whether a label or log exists rather than whether it is visible, retained, verifiable, contested and used for remedy.

These failures interact. Interoperability loss reduces verification; uncertain verification increases moderation error; opaque moderation weakens appeals; and weak feedback prevents recalibration. The result is not a single malfunction but a pathway through which a system produces trust theatre: visible signs of governance without the capacity to answer disputes or correct harm. This outcome is not necessarily accidental. Symbolic compliance can be locally rational: attaching a label or retaining a log reduces an organisation’s exposure to visible non-compliance and reputational risk, even when the underlying verification or accountability capacity remains absent, so that the costs of unresolved system-level failure are borne by citizens rather than by the organisations best placed to prevent it.

Figure 2 summarises three recurrent forms of symbolic compliance. A visible label without provenance shifts recognition burdens to citizens. Provenance without public access creates infrastructural opacity and dependence on platform interpretation. Audit records without institutional authority create an answerability gap. Operational accountability requires the coupling of all three functions with contestability and enforceable responsibility. Table 5 sets out how each failure mode manifests, who bears the resulting burden and the corresponding governance response.

[Figure 2 about here]

Table 5
System failure modes and distributional effects

Failure mode	System manifestation	Who bears the burden?	Governance response
Coarse risk classification	Similar controls across very different public consequences	Low-risk creators face unnecessary friction; high-risk audiences remain exposed	Scenario-based risk tiers and periodically revised control bundles
Label–provenance disconnection	Human disclosure is not linked to machine evidence or workflow decisions	Citizens must interpret labels without verification; reviewers lack usable signals	Dual-mode coordination and routing of provenance status into review
Interoperability loss	Transformations or platform changes destroy evidence	Smaller platforms and cross-border users face higher attribution costs	Minimum identifiers, shared semantics, APIs and recovery mechanisms
Inaccessible public verification	Provenance exists only in back-end systems	Citizens and journalists cannot resolve uncertainty or contribute usable reports	Accessible verification portal, explanation and feedback channels
Fragmented responsibility	Actor IDs and records are split across organisations	Affected persons face delay and organisational blame shifting	Lifecycle logging, access rules, escalation and responsibility mapping
Output-oriented evaluation	Compliance measured by attachment rates or record existence	Invisible or fragile safeguards pass formal inspection	Outcome measures covering visibility, retention, response, correction and attribution

Risk-tiered governance without technological determinism

Uniform control intensity is neither efficient nor socially neutral. Entertainment, routine administrative guidance, financial or medical information, emergency announcements and realistic representations of public officials create different expected harms. The corpus supports risk differentiation, but risk classification itself should not be treated as an objective technical calculation. It embeds judgements about whose harms count, which institutions are trusted and what level of uncertainty is tolerable.

A four-level design can nevertheless provide an initial governance scaffold. Low-risk content may require lightweight disclosure and optional verification. Medium-risk content adds machine-verifiable signals. High-risk public information requires coordinated visible and implicit marking, structured records and human review. Critical emergency or identity-related content requires persistent provenance, strong source authentication, routine audit, escalation procedures and full responsibility tracing. Thresholds should be locally validated and periodically contested rather than fixed as universal scores.

Public verification is a crucial counterweight to infrastructural concentration. Citizens should be able to submit a file, paste a link or scan a code and receive a concise status such as verified source, provenance available, provenance inconsistent or unable to verify. The interface must explicitly separate source history from substantive truth. It should also provide a reporting and appeal route so that public signals enter review and regulatory sampling rather than ending at the interface.

Discussion

These findings have both theoretical and practical implications. Conceptually, they show what content labelling and provenance systems are actually doing as a governance arrangement, beyond the specific rules examined here. Practically, they point to concrete design choices for digital-government provenance systems. The remainder of this section develops these two implications in turn, before discussing the limitations of the analysis and directions for future research.

Moving beyond the technical–social and optimism–pessimism binaries

The analysis contributes to AI social science by showing that content labelling is neither a technical solution imposed on society nor a social response external to technology. Categories, metadata, interfaces and audit procedures co-produce the object they claim to reveal. A media object becomes “AI-generated,” “verified” or “suspicious” through institutional definitions and technical representations. Those classifications then reorganise platform workflows, citizen expectations and regulatory authority. This analysis extends a growing body of Collection-adjacent scholarship on trust, ethics and governance in the artificial-intelligence era (Afroogh et al. 2024; Bhardwaj 2025; Hahn et al. 2026; Krueger et al. 2025) by specifying how transparency functions are institutionally distributed rather than treating trust or accountability as a general disposition towards AI.

This perspective also moves beyond optimism and pessimism. Labels and provenance are not inherently effective or futile. They can support public knowledge when integrated with accessible verification, review and institutional remedy. They can also create false reassurance, generalised suspicion and new dependencies. The relevant question is which configuration produces accountable relations for a particular context and which actors can challenge that configuration.

The triple-transparency framework provides a vocabulary for this analysis. Public legibility addresses interpretation, infrastructural verifiability addresses provenance continuity and institutional answerability addresses forums and consequences. The framework differs from simple transparency taxonomies because it treats the layers as power-laden relationships with distinctive failure modes. It also clarifies why increasing one form cannot compensate for the absence of another.

Provenance as public infrastructure and epistemic governance

Content provenance is often described as a chain of custody for digital media. The comparison shows that it should also be understood as epistemic governance: a system for authorising claims about origin and transformation. Credential issuers, trust-list managers, platforms and public agencies become institutional witnesses. Their authority is encoded in interfaces that ordinary users may not see or understand.

This raises a democratic design problem. A provenance system designed exclusively around cryptographic validity may centralise power in organisations that operate the signing and validation infrastructure. A system designed exclusively around visible labels may responsabilise users without changing platform incentives. A system designed exclusively around audit can privilege institutional observers while excluding affected citizens. Public infrastructure requires plural access, transparent governance of standards, contestability and mechanisms for smaller actors to participate without prohibitive costs.

The distinction between provenance and truth remains essential. Treating provenance as truth would convert institutional source authority into epistemic certainty. Yet official information can be wrong, incomplete or politically contested. Conversely, valuable evidence can lack credentials. The appropriate role of provenance is to make histories and claims inspectable, not to end substantive inquiry.

Global and institutional asymmetries

The corpus brings together Chinese administrative rules, European risk-based law, US-origin risk-management frameworks, international management standards and a private-sector provenance specification. This hybrid field illustrates geopolitical asymmetry. Powerful jurisdictions and firms can shape technical definitions that become de facto global requirements. Governments with fewer resources may face high costs for certification, key management, secure identity infrastructure and continuous auditing, extending a broader pattern in which the material and computational infrastructures of AI concentrate interpretive and economic power in a small number of firms and states (Crawford 2021).

Critical and decolonial AI scholarship emphasises that governance systems should not treat the Global South only as a site of adoption or risk (Birhane 2020; Mohamed et al. 2020). Local institutions must participate in defining risk categories, trusted-source arrangements, language and accessibility requirements, and acceptable evidence practices. Otherwise, a globally interoperable system may remain institutionally exclusionary.

The same evidentiary infrastructures also intersect with state power beyond routine administration. Identity binding, audit trails and content credentials developed to support public accountability can be repurposed for policing, border control and mass surveillance, converting an answerability mechanism into an instrument of population monitoring when institutional safeguards are weak or absent. Civil-society verification networks, independent fact-checkers and community-based provenance auditing illustrate a complementary, informal form of answerability that operates outside official state and

platform channels; such practices can contest official narratives and expose failures of formal transparency, but they also depend on access to the same infrastructures whose concentration this article critiques, and they cannot substitute for enforceable institutional responsibility.

The practical implication is a layered interoperability model, echoing broader proposals for layered AI-governance architectures that separate technical, organisational and legal responsibilities across levels (Gasser and Almeida 2017). A minimum common layer can define content identifiers, credential semantics and verification statuses, while national and local institutions retain authority to set high-risk scenarios, disclosure language, appeal rights and retention constraints. Interoperability should connect governance systems without erasing institutional plurality.

Design principles for accountable digital-government provenance

Taken together, these theoretical and geopolitical considerations carry direct practical implications for how digital-government provenance systems should be designed. The findings support seven design principles, extending recent proposals for the operational conditions under which generative-AI systems can be observed, inspected and modified by those affected by them (Ferrari et al. 2025). First, separate process disclosure from truth assessment so that an AI-generated label is not treated as a verdict on accuracy. Second, calibrate controls to public consequences and source authority rather than to AI use alone. Third, coordinate public-facing and machine-readable signals so that labels connect to inspectable evidence. Fourth, preserve contestability by treating missing or inconsistent provenance as a risk signal rather than automatic proof of deception. Fifth, establish institutional access and responsibility rules before expanding data retention. Sixth, evaluate outcomes such as recognition, verification, response, correction and attribution rather than counting labels. Seventh, distribute governance authority across regulators, standards bodies, platforms, public agencies, researchers and civil society. Table 6 links each principle to an operational implication and a candidate outcome that future research can evaluate.

Table 6
Design principles and evaluable outcomes

Principle	Operational implication	Candidate outcome
Separate provenance from truth	Display source and transformation status independently from factual assessment	Users distinguish “verified source” from “verified claim”
Use contextual risk tiers	Match disclosure, authentication, review and audit to public consequences	High-impact content receives faster and deeper review
Couple visible and machine-readable signals	Link labels to credentials, content IDs and workflow triggers	Higher provenance retention and successful verification
Preserve contestability	Use missing provenance as a risk signal; provide appeal and correction	Lower wrongful restriction and faster restoration
Govern logs proportionately	Define purpose, access, retention and oversight before collection	Complete but minimised evidence chains
Measure outcomes	Track visibility, comprehension, retention, response, correction and attribution	Reduced symbolic compliance and improved learning
Pluralise authority	Include public institutions, independent auditors and civil-society participation	More accessible standards and fewer single-point dependencies

Limitations

The study is conceptual and comparative. It does not measure whether a particular label changes citizen trust or whether a provenance scheme survives specific transformations. The corpus covers major Chinese, European and transnational instruments but is not an exhaustive inventory of national regimes. The functional coding is interpretive and was not assessed through a formal intercoder-reliability statistic. The proposed risk tiers and indicators are design propositions rather than empirically calibrated thresholds.

Future research directions

Future research should test each connection in the framework. Experiments can examine how label wording, persistence and access to verification affect recognition, trust calibration and sharing. Technical benchmarks can test credential survival and recovery after compression, screenshots, editing and platform transfer. Organisational studies can examine how provenance signals enter public-sector publishing, moderation and incident response. Comparative institutional research should analyse how regulatory capacity, administrative law and political culture affect access to evidence and remedies. Participatory studies are needed to examine whether citizens, journalists and marginalised groups understand and can contest provenance systems.

Longitudinal research is particularly important. The lifecycle is changing as generative models become embedded in editing suites, autonomous agents and distribution systems. Static labelling rules may quickly become misaligned with production practices. Governance is accordingly best studied as ongoing infrastructural maintenance rather than a one-time regulatory intervention.

Conclusion

AI-generated content labelling in digital government is a problem of sociotechnical ordering. Labels define what becomes visible to citizens; provenance systems define what becomes inspectable to platforms and verifiers; and audit arrangements define what becomes answerable to institutions. These functions can support trustworthy public communication, but they also redistribute authority, verification work and dependency.

The central conclusion is conditional. More disclosure does not necessarily produce more accountability. Public legibility must be connected to practical verification; provenance must remain distinguishable from truth and open to contestation; and evidence must be accessible to institutions capable of assigning responsibility and providing remedy. Without these connections, visible labels, technical credentials and audit logs can coexist with system-level failure.

This account treats provenance as a governance infrastructure rather than a neutral technical fix, and in that sense moves away from framing AI governance as either a simple technical solution or an unavoidable failure. The task is to design institutions and interfaces through which origin claims can be understood, challenged and acted upon. In digital government, accountable transparency is not a property attached to a content object. It is an ongoing relationship among citizens, technologies, platforms and public authority.

Declarations

Data availability

The analytical corpus, coding framework, claim-to-evidence traceability matrix and negative-case checks are provided in the Supplementary Information. The underlying regulatory, policy, technical and management instruments are publicly available from the issuing organisations and are cited in the reference list.

Competing interests

The authors declare no competing interests.

Ethical approval

Ethical approval was not required because the study analyses publicly available policy documents, technical standards and published literature and does not involve human participants, personal data or animals.

Informed consent

Informed consent was not required because the study does not involve human participants.

Author Contribution

Conceptualisation: S.W. and Y.L.; methodology: Y.L., P.L. and S.W.; formal analysis: Y.L., S.W. and S.-K.C.; investigation: Y.L., S.W. and S.Q.W.; writing—original draft: Y.L. and S.W.; writing—review and editing: P.L., S.-K.C. and S.Q.W.; supervision: S.-K.C.; project administration: S.W. All authors reviewed and approved the manuscript.

References

1. Afroogh S, Akbari A, Malone E et al (2024) Trust in AI: progress, challenges, and future directions. *Humanit Soc Sci Commun* 11:1568. <https://doi.org/10.1057/s41599-024-04044-8>
2. Ananny M, Crawford K (2018) Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc* 20(3):973–989. <https://doi.org/10.1177/1461444816676645>
3. Autio C, Schwartz R, Dunietz J et al (2024) Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. National Institute of Standards and Technology, Gaithersburg. <https://doi.org/10.6028/NIST.AI.600-1>
4. Batool A, Zowghi D, Bano M (2025) AI governance: a systematic literature review. *AI Ethics* 5:3265–3279. <https://doi.org/10.1007/s43681-024-00653-w>
5. Bhardwaj A (2025) Artificial intelligence and the limits of reason: a framework for responsible use in public and private sectors. *Humanit Soc Sci Commun* 12:1565. <https://doi.org/10.1057/s41599-025-05749-0>
6. Birhane A (2020) Algorithmic colonization of Africa. *SCRIPTed* 17(2):389–409. <https://doi.org/10.2966/scrip.170220.389>
7. Bowker GC, Star SL (1999) *Sorting things out: classification and its consequences*. MIT Press, Cambridge
8. Bullock JB (2019) Artificial intelligence, discretion, and bureaucracy. *Am Rev Public Adm* 49(7):751–761. <https://doi.org/10.1177/0275074019856123>
9. Chesney R, Citron DK (2019) Deep fakes: a looming challenge for privacy, democracy, and national security. *Calif Law Rev* 107:1753–1820. <https://doi.org/10.15779/Z38RV0D15J>

10. Coalition for Content Provenance and Authenticity (2024) C2PA Technical Specification 2.1. Coalition for Content Provenance and Authenticity
11. Couldry N, Mejias UA (2019) The costs of connection: how data is colonizing human life and appropriating it for capitalism. Stanford University Press, Stanford
12. Crawford K (2021) Atlas of AI: power, politics, and the planetary costs of artificial intelligence. Yale University Press, New Haven
13. Cyberspace Administration of China et al et al (2022a) Provisions on the Administration of Algorithmic Recommendations in Internet Information Services. Beijing
14. Cyberspace Administration of China et al et al (2022b) Provisions on the Administration of Deep Synthesis in Internet-Based Information Services. Beijing
15. Cyberspace Administration of China et al et al (2023) Interim Measures for the Management of Generative Artificial Intelligence Services. Beijing
16. Cyberspace Administration of China et al et al (2025) Measures for Labelling AI-Generated and Synthetic Content. Beijing
17. Diakopoulos N, Johnson D (2021) Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media Soc* 23(7):2072–2098.
<https://doi.org/10.1177/1461444820925811>
18. European Parliament and Council of the European Union (2024) Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off J Eur Union L* 2024/1689
19. Eubanks V (2018) Automating inequality: how high-tech tools profile, police, and punish the poor. St Martin's, New York
20. Fallis D (2021) The epistemic threat of deepfakes. *Philos Technol* 34(4):623–643.
<https://doi.org/10.1007/s13347-020-00419-2>
21. Feng KJ, Ritchie N, Blumenthal P, Parsons A, Zhang AX (2023) Examining the impact of provenance-enabled media on trust and accuracy perceptions. *Proc ACM Hum Comput Interact* 7(CSCW2):1–42.
<https://doi.org/10.1145/3610061>
22. Ferrari F, van Dijck J, van den Bosch A (2025) Observe, inspect, modify: three conditions for generative AI governance. *New Media Soc* 27:2788–2806.
<https://doi.org/10.1177/14614448231214811>
23. Floridi L, Cowls J, Beltrametti M et al (2018) AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach* 28:689–707.
<https://doi.org/10.1007/s11023-018-9482-5>
24. Gamage D, Sewwandi D, Zhang M, Bandara AK (2025) Labeling synthetic content: user perceptions of warning label designs for AI-generated content on social media. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, New York.
<https://doi.org/10.1145/3706598.3713171>

25. Gasser U, Almeida VA (2017) A layered model for AI governance. *IEEE Internet Comput* 21(6):58–62. <https://doi.org/10.1109/MIC.2017.4180835>
26. Gillespie T (2018) *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, New Haven
27. Hahn M, Samhammer D, Tretter M et al (2026) An actor-network-theory informed normative framework for the governance of AI-systems in medicine. *Humanit Soc Sci Commun* 13:884. <https://doi.org/10.1057/s41599-026-07957-8>
28. International Organization for Standardization (2023a) *ISO/IEC 42001:2023 Information technology –Artificial intelligence–Management system*. ISO, Geneva
29. International Organization for Standardization (2023b) *ISO/IEC 23894:2023 Information technology –Artificial intelligence–Guidance on risk management*. ISO, Geneva
30. Jasanoff S (ed) (2004) *States of knowledge: the co-production of science and social order*. Routledge, London
31. Jasanoff S, Kim SH (2009) Containing the atom: sociotechnical imaginaries and nuclear power in the United States and South Korea. *Minerva* 47(2):119–146. <https://doi.org/10.1007/s11024-009-9124-4>
32. Kemper J, Kolkman D (2019) Transparent to whom? No algorithmic accountability without a critical audience. *Inf Commun Soc* 22(14):2081–2096. <https://doi.org/10.1080/1369118X.2018.1477967>
33. Kroll JA, Huey J, Barocas S et al (2017) Accountable algorithms. *Univ Pa Law Rev* 165:633–705
34. Krueger F, Riedl R, Bartz JA et al (2025) A call for transdisciplinary trust research in the artificial intelligence era. *Humanit Soc Sci Commun* 12:1124. <https://doi.org/10.1057/s41599-025-05481-9>
35. Lewis A, Vu P, Duch RM, Chowdhury A (2023) Deepfake detection with and without content warnings. *R Soc Open Sci* 10:231214. <https://doi.org/10.1098/rsos.231214>
36. Li F, Yang Y (2024) Impact of artificial intelligence-generated content labels on perceived accuracy, message credibility, and sharing intentions for misinformation: web-based randomized controlled experiment. *JMIR Form Res* 8:e60024. <https://doi.org/10.2196/60024>
37. Liu Q, Wang L, Luo M (2025) When seeing is not believing: self-efficacy and cynicism in the era of intelligent media. *Humanit Soc Sci Commun* 12:274. <https://doi.org/10.1057/s41599-025-04594-5>
38. Mergel I, Edelmann N, Haug N (2019) Defining digital transformation: results from expert interviews. *Gov Inf Q* 36:101385. <https://doi.org/10.1016/j.giq.2019.06.002>
39. Metcalf J, Moss E, boyd d (2019) Owning ethics: corporate logics, Silicon Valley, and the institutionalization of ethics. *Social Research: Int Q* 86(2):449–476. <https://doi.org/10.1353/sor.2019.0022>
40. Mohamed S, Png MT, Isaac W (2020) Decolonial AI: decolonial theory as sociotechnical foresight in artificial intelligence. *Philos Technol* 33:659–684. <https://doi.org/10.1007/s13347-020-00405-8>
41. National Institute of Standards and Technology, Gaithersburg (2023) *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. NIST. <https://doi.org/10.6028/NIST.AI.100-1>

42. Pasquale F (2015) *The black box society: the secret algorithms that control money and information*. Harvard University Press, Cambridge
43. Power M (1997) *The audit society: rituals of verification*. Oxford University Press, Oxford
44. Selbst AD, boyd d, Friedler SA, Venkatasubramanian S, Vertesi J (2019) Fairness and abstraction in sociotechnical systems. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, New York, pp 59–68. <https://doi.org/10.1145/3287560.3287598>
45. Sloane M, Solano-Kamaiko IR, Yuan J et al (2023) Introducing contextual transparency for automated decision systems. *Nat Mach Intell* 5:187–195. <https://doi.org/10.1038/s42256-023-00623-7>
46. Star SL, Ruhleder K (1996) Steps toward an ecology of infrastructure: design and access for large information spaces. *Inf Syst Res* 7(1):111–134. <https://doi.org/10.1287/isre.7.1.111>
47. State Administration for Market Regulation, Standardization Administration of China (2025) GB 45438 – 2025 Cybersecurity technology—Labeling method for content generated by artificial intelligence. Beijing
48. Taeihagh A (2021) Governance of artificial intelligence. *Policy Soc* 40(2):137–157. <https://doi.org/10.1080/14494035.2021.1928377>
49. Vaccari C, Chadwick A (2020) Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Soc Media Soc* 6(1). <https://doi.org/10.1177/2056305120903408>
50. Wieringa M (2020) What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, pp 1–18. <https://doi.org/10.1145/3351095.3372833>
51. Wirtz BW, Weyerer JC, Geyer C (2019) Artificial intelligence and the public sector—applications and challenges. *Int J Public Adm* 42(7):596–615. <https://doi.org/10.1080/01900692.2018.1498103>
52. Wittenberg C, Epstein Z, Berinsky AJ, Rand DG (2024) Labeling AI-generated content: promises, perils, and future directions. *MIT Explor Generative AI*. <https://doi.org/10.21428/e4baedd9.0319e3a6>
53. Zuiderwijk A, Chen YC, Salem F (2021) Implications of the use of artificial intelligence in public governance: a systematic literature review and a research agenda. *Gov Inf Q* 38:101577. <https://doi.org/10.1016/j.giq.2021.101577>

Figures

Triple transparency as a sociotechnical arrangement

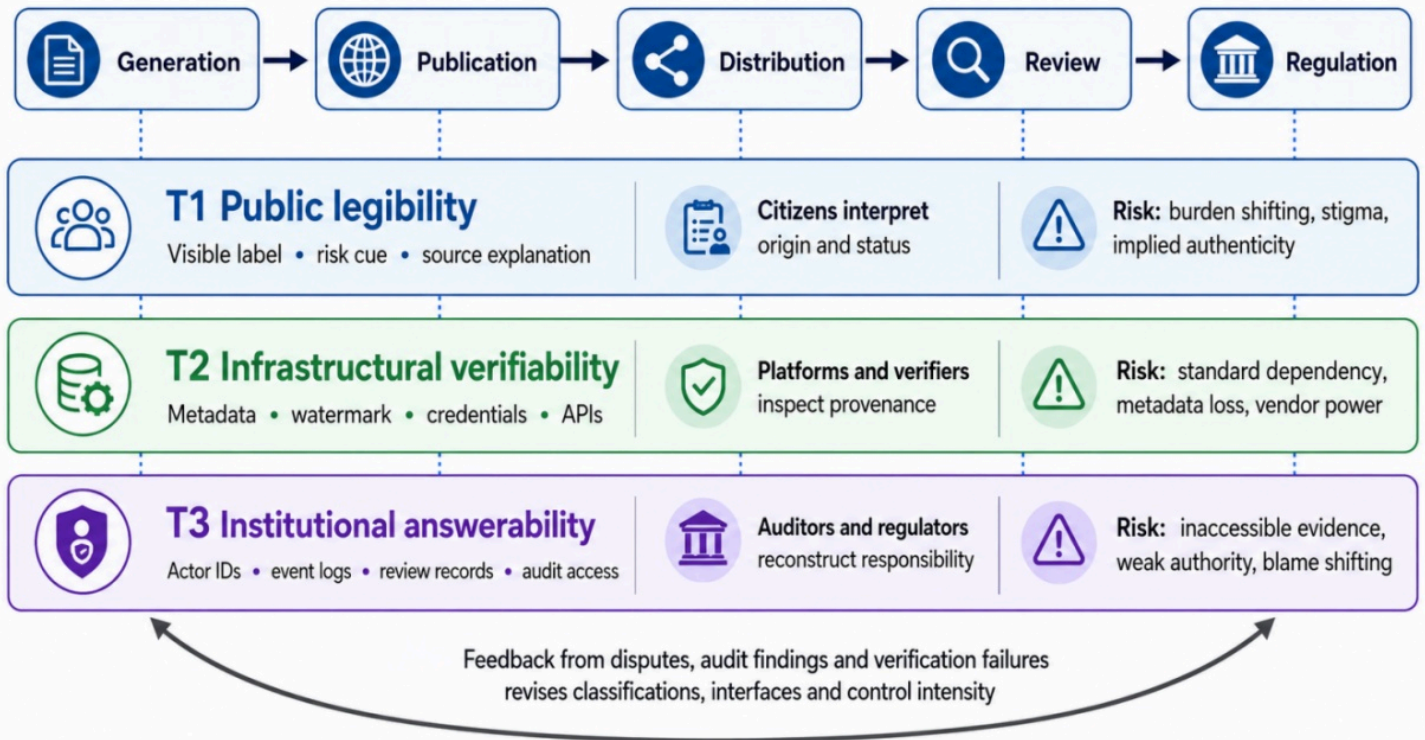


Figure 1

Triple transparency across the lifecycle of AI-generated government content. Public legibility, infrastructural verifiability and institutional answerability are non-substitutable functions. Each layer creates characteristic risks involving burden shifting, infrastructural dependence or responsibility diffusion. Feedback from disputes, incidents, appeals and audits revises classifications, interfaces and control intensity.

How transparency becomes symbolic compliance

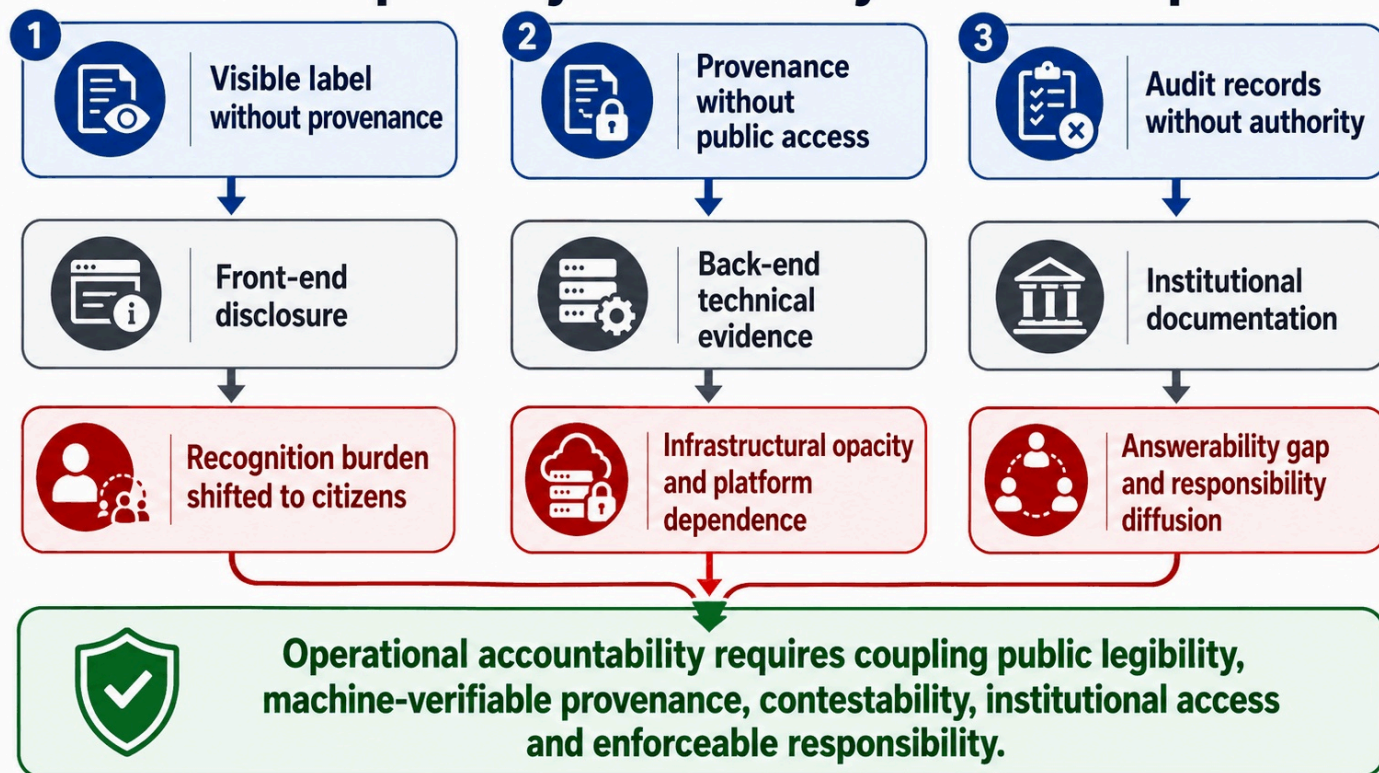


Figure 2

Pathways from transparency to symbolic compliance. Visible labels without provenance shift verification burdens to citizens; provenance without public access creates infrastructural opacity; and audit records without institutional authority create an answerability gap. Operational accountability requires coupling disclosure, provenance, contestability, institutional access and enforceable responsibility.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformationREVISED.docx](#)