Data-to-Rating (D2R):
A Public-Data Methodology for
Continuous Assessment and Structured Judgment

Author: Khaled Koubaa

Correspondence: research@atworthy.com

Abstract

Many assessment systems were formed in environments where meaningful evidence was difficult to observe without voluntary disclosure, questionnaires, site visits, internal documents, interviews, or formal attestations. That assumption remains valid in many physical, organizational, and high-stakes assurance contexts, but it is no longer sufficient as a general theory of assessment. Digital-native environments, especially software, demonstrate a different evidentiary logic: code, dependencies, configuration files, repositories, logs, tests, build artifacts, deployment records, and software bills of materials can often be assessed continuously or near-continuously. More broadly, organizations, public authorities, platforms, and jurisdictions now generate dense public digital traces through websites, registries, disclosures, technical interfaces, procurement records, service workflows, public repositories, and machine-readable artifacts. Public data has moved from being a supplementary source to becoming a measurable evidence environment.

This paper introduces Data-to-Rating (D2R), a methodological framework for converting publicly obtainable digital data into qualified signals and then into structured, comparable, and contestable assessments. D2R complements statutory audit, on-site inspection, and direct assurance by adding a public-evidence measurement layer that qualifies observable data as relevant signals, extracts features, maps them to evaluative dimensions, produces scores and ratings, and attaches explicit confidence and evidence lineage. D2R is formalized as an evidence-to-judgment architecture rather than a scoring formula or visual rating symbol.

The paper makes four contributions: it defines a strict D2R vocabulary that separates data, signals, features, indicators, scores, ratings, representations, confidence, and judgment; it develops an evidentiary theory for public-data assessment, including the principle of observable construct manifestation and a signaling-theoretic hierarchy that orders public evidence by the cost of counterfeiting it; it specifies an end-to-end methodological pipeline with human control points, a measurement model that separates latent constructs from their public manifestations, and explicit uncertainty and validation requirements; and it positions D2R as a complement to audits and questionnaires that preserves human authority over construct design, thresholds, disputes, and high-stakes interpretation. The central claim is bounded: public data cannot reveal everything, but where relevant qualities leave observable traces, a disciplined D2R method can convert public evidence into defensible structured judgment.

Keywords:

Data-to-Rating; public data; digital evidence; structured judgment; composite indicators; rating methodology; continuous assessment; public-signal measurement.

JEL Classification: C43; C81; D83; H83; M42.

1. Introduction

Modern institutions are already assessed through audits, certifications, questionnaires, compliance reports, benchmarks, maturity models, rankings, credit analyses, ESG scores, cybersecurity ratings, service reviews, software quality tools, and many other instruments. Many of these assessment instruments, however, were designed for physical, organizational, or document-bound environments in which meaningful evidence was difficult to observe without voluntary disclosure, questionnaires, site visits, internal documents, interviews, or privileged access.

This condition remains important. In many domains, especially where internal controls, confidential records, operational effectiveness, or legal assurance are at stake, structured judgment still depends on the ability to ask, inspect, sample, and verify evidence that is not externally visible. Direct audit, inspection, and attestation therefore remain essential. D2R leaves the need for privileged evidence intact while arguing that, in domains where relevant qualities leave observable digital traces, assessment no longer has to rely on privileged evidence alone.

Digital-native environments show why this distinction matters. In software, the object of assessment is often versioned, machine-readable, executable, testable, and instrumented by design. Source code, dependencies, configuration files, repositories, test results, issue histories, build logs, deployment artifacts, telemetry, and software bills of materials can support continuous or near-continuous assessment. For D2R, the important point is that digital artifacts can change the assessment model itself, shifting it from episodic inspection toward evidence-linked, event-triggered, and continuously refreshable structured judgment.

Beyond software, a different evidence environment now exists. Public entities and private organizations leave traces of their conduct, governance, capacity, risk posture, and service quality in digital form. A public agency may expose its maturity through online service design, AI use case inventories, accessibility pages, procurement notices, privacy statements, incident communication, public consultation channels, and machine-readable records. A company may expose aspects of its operational and governance posture through websites, security headers, app listings, public policies, repositories, regulatory filings, hiring signals, issue trackers, product interfaces, and public customer-review environments. A jurisdiction may reveal institutional capacity through open-data portals, digital service coverage, legislative transparency, registry availability, and public performance dashboards. Although partial, these traces can still provide meaningful observable manifestations of institutional life.

Data-to-Rating (D2R) is proposed as a methodological response to this shift. D2R is a framework for transforming publicly obtainable data into structured evaluative outputs. D2R goes beyond web scraping, natural language processing, benchmarking, and composite indexing by organizing them within an evidence transformation architecture: data are collected, qualified as signals, converted into features, mapped to evaluative dimensions, aggregated into scores, translated into ratings, represented to users, and governed through confidence, traceability, and contestability. The distinctive contribution is the discipline of the transformation chain. In this framework, a rating emerges from a governed public-evidence process, with evidence and rules behind the final category.

D2R does not emerge from an intellectual vacuum, and its claim to novelty should be stated precisely. The use of nonreactive public traces as evidence has a long pedigree: Webb and colleagues argued six decades ago that unobtrusive measures avoid the reactivity that contaminates interviews and questionnaires, precisely because the assessed party does not shape the evidence in response to being assessed (Webb et al., 1966). Computational social science has since demonstrated that digital traces can support inference about behavior at institutional scale (Lazer et al., 2009), while the failure of Google Flu Trends showed how trace-based measurement degrades when construct discipline, validation, and methodological transparency are absent

(Lazer et al., 2014). In parallel, the sociology of quantification has documented how ratings commensurate disparate qualities into comparable numbers and, in doing so, reshape the conduct of the entities they measure (Espeland and Stevens, 1998; Espeland and Sauder, 2007). Most recently, empirical work on sustainability ratings has shown that when rating methodologies are opaque, different raters assessing the same entities diverge so substantially that the ratings lose decision value (Berg et al., 2022). D2R synthesizes these strands into something none of them supplies individually: a general, domain-portable methodology for converting public digital evidence into contestable ratings, with explicit treatment of evidence qualification, uncertainty, and human control. The contribution is architectural rather than algorithmic.

The paper is intentionally conceptual and develops a general methodology rather than releasing a benchmark, ranking real entities, or validating a proprietary platform. Possible domains include digital public service quality, AI governance transparency, cyber exposure, public institutional transparency, procurement intelligence, vendor monitoring, sustainability disclosure quality, and other fields in which public evidence can be systematically interpreted. The purpose is scientific and methodological: to articulate how public data can support defensible assessment and judgment without pretending that external data can replace all forms of direct assurance.

The paper proceeds as follows. Section 2 describes the shift from access-dependent review to digital visibility and introduces software assessment as a reference case for continuous, artifact-based D2R. Section 3 clarifies the limits of conventional assessment models. Section 4 defines D2R and its scope. Sections 5 and 6 develop the conceptual vocabulary and evidentiary theory. Sections 7 and 8 specify the methodological pipeline, scoring logic, confidence treatment, and output structure. Sections 9 and 10 address automation, human judgment, validation, and quality assurance. Sections 11 to 14 discuss an illustrative application, the relationship with audit and questionnaire-based assessment, risks, safeguards, and a research agenda. Section 15 concludes.

2. The Evidentiary Shift: From Access-Dependent Review to Digital Visibility

Traditional assessment methods developed under conditions in which information about an entity was difficult to obtain from the outside. The visible surface of an organization was thin. External observers could see legal identity, premises, broad reputation, occasional public reports, and perhaps some market behavior, but they could not usually observe process maturity, governance design, service responsiveness, operational accessibility, technical configuration, or policy implementation without direct cooperation. In physical-world and institution-bound assessment, privileged access was often necessary.

The digital environment preserves the need for privileged access in some contexts while changing the measurement problem. Public data is now produced by ordinary institutional and technical operation. Websites, portals, mobile applications, open-data catalogues, repositories, service workflows, transparency registers, public consultations, application programming interfaces, disclosure notices, version histories, public procurement systems, regulatory databases, and machine-readable artifacts create an external evidence layer. This layer does not expose every relevant fact, but it expands what can be observed without asking the entity to complete a form or admit an assessor to its premises.

The shift is uneven. Some domains remain deeply access-dependent because the most important evidence is internal, confidential, physical, or legally protected. Other domains are increasingly digital-native, meaning that the assessed object or practice produces assessable artifacts as part of its normal operation. Software is the clearest reference case. It shows that where evidence is versioned, executable, timestamped, machine-readable, and continuously updated, assessment can move from periodic review toward ongoing measurement. D2R extends this logic beyond software to broader public-data environments, while preserving the limits and safeguards required when the assessed object is only partially observable.

2.1 Software Assessment as a Reference Case for Continuous D2R

Software provides a useful reference case because it shows that some domains produce abundant assessable evidence. In physical-world assessment, evidence often must be requested, inspected, sampled, or attested. In software assessment, by contrast, many relevant artifacts are digital by design. Source code, dependency manifests, configuration files, test outputs, build logs, software bills of materials, repository histories, release artifacts, vulnerability records, deployment records, and runtime telemetry can often be examined at commit time, build time, release time, or runtime. Software therefore illustrates a more advanced evidence environment: one in which the assessed object produces structured, timestamped, and machine-readable evidence as part of its ordinary lifecycle. This logic is reflected in contemporary software assurance practice, including secure software development frameworks, software supply-chain security guidance, CI/CD pipeline controls, and SBOM-based transparency mechanisms (Cybersecurity and Infrastructure Security Agency, 2025; National Institute of Standards and Technology, 2022, 2024).

Code alone cannot fully explain software behavior. Static analysis cannot prove all runtime behavior. Dependency scanning cannot establish secure architecture. Test coverage does not guarantee absence of defects. A software bill of materials can improve component transparency, but it does not prove that a system is secure (Cybersecurity and Infrastructure Security Agency, 2025). Runtime telemetry may reveal behavior but not all design intent. The lesson is narrower and more useful: where the object of assessment is digital, versioned, executable, and instrumented, assessment can become continuous, event-triggered, and evidence-linked rather than purely episodic and access-dependent.

D2R generalizes this lesson beyond software. A D2R domain becomes more suitable for continuous rating as its public evidence becomes more software-like: digital, timestamped, structured, machine-readable, versioned, attributable, comparable, testable, and update-sensitive. In such domains, rating systems can borrow practical

design principles from software assurance: versioned evidence snapshots, automated signal extraction, change detection, pipeline-based review, confidence scoring, issue-level remediation, and human review of exceptions. Software assessment therefore gives D2R a methodological reference case for continuous public-evidence assessment.

The translation from software to D2R should remain bounded. Software artifacts are unusually formal and machine-readable. Institutions, jurisdictions, services, and governance practices are less formally observable and more exposed to visibility bias, proxy distortion, and strategic presentation. For that reason, software should be used as a bounded reference case: it shows the direction of travel from periodic inspection toward continuous, artifact-based, event-triggered measurement, with confidence treatment and human judgment preserved where evidence is incomplete or consequential.

Table 1. Software assessment principles translated into D2R design

Software assessment principle	How it works in software	D2R translation
Version control	Code changes are timestamped, attributable, and reconstructable	Evidence snapshots, methodology versions, rating histories, and source timestamps
CI/CD gates	Tests and checks run at commit, build, package, or deployment stages, consistent with CI/CD pipeline security guidance	D2R updates triggered by public evidence changes, scheduled refreshes, or material events
Static and dynamic analysis	Code and behavior are assessed through automated inspection and execution signals	Public documents, interfaces, repositories, disclosures, and service behaviors are parsed as signals
Dependency scanning	Components are inventoried and checked for known issues	Public source classes and evidence components are catalogued and monitored
Software bill of materials	Software components and component relationships are recorded to support software transparency and supply-chain visibility	D2R evidence ledgers record source classes, evidence items, signal lineage, and attribution
Build and release provenance	The origin and production path of software artifacts are recorded	Rating provenance records source acquisition, extraction logic, scoring rules, and approval path
Runtime monitoring	Logs, metrics, traces, and alerts show behavior over time	Public dashboards, service behavior, update cadence, incident communication, and correction patterns show observable institutional behavior
Security scorecards	Automated checks produce security scores and identify areas for remediation in open-source software assessment	D2R ratings include score, confidence, missing evidence, limitations, and correction pathway
Human exception review	False positives, risk acceptance, and architectural context require human judgment	Contested findings, low-confidence outputs, threshold changes, and high-stakes uses require governed human review

Table 1 translates selected software-assurance practices into D2R design principles. The comparison draws on secure software development guidance, software supply-chain security guidance for DevSecOps and CI/CD pipelines, SBOM transparency practices, and automated open-source security scorecard methods (Cybersecurity and Infrastructure Security Agency, 2025; National Institute of Standards and Technology, 2022, 2024; Open Source Security Foundation, 2026).

The growth of open government data and digital public administration illustrates the broader trend. The OECD's 2023 OURdata Index benchmarks government efforts to design and implement open government data policies and reports that open government data has become a significant instrument for addressing policy issues and improving access to timely, relevant, and high-quality data (OECD, 2023). The OECD's 2025 Digital Government Index and OURdata results similarly treat digital government and open reusable data as measurable dimensions of public-sector transformation (OECD, 2026a). The World Bank's GovTech Maturity Index also measures digital transformation in the public sector across 198 economies and uses both survey and remotely collected data, which demonstrates that external observation already contributes to institutional measurement at global scale (World Bank, 2025).

AI governance provides another example. Recent public policy instruments increasingly require inventories, monitoring, ongoing evaluation, and public reporting. OMB Memorandum M-25-21 requires agencies to maintain AI use case inventories and establishes processes to measure, monitor, and evaluate the ongoing performance and effectiveness of high-impact AI applications (OMB, 2025). NIST’s AI Risk Management Framework likewise treats the measurement and management of AI risks as continuing functions rather than one-time exercises (NIST, 2023), and NIST’s 2026 report on post-deployment monitoring of deployed AI systems identifies monitoring as crucial for confident AI adoption and organizes monitoring challenges across categories such as functionality, operational performance, human factors, compliance, and societal impact (NIST, 2026). The Financial Stability Board similarly discusses monitoring AI adoption and related vulnerabilities through direct and proxy indicators, including publicly available information and vendor data (FSB, 2025). These developments show that the measurement environment is moving toward continuous observation, proxy indicators, and external evidence streams.

The significance for D2R is conceptual. If public digital evidence is abundant, then the primary assessment question changes. The older question was often: can the assessor obtain enough evidence through access, questionnaire, or audit? The new question is: what valid inferences can be drawn from the public evidence that already exists, and how should those inferences be governed? D2R is built around this second question. D2R treats public evidence as a distinct class of evidence with its own strengths, limits, and safeguards.

3. The Limits of Prevailing Assessment Models

D2R is introduced because existing assessment methods were optimized for different evidence conditions. Self-reported assessment, audit-based assurance, and composite indices remain valuable, but each has structural limits that become more serious when assessment must be frequent, scalable, externally comparable, and linked to public digital evidence.

Questionnaires and self-assessments are efficient when the assessed entity has the information and is willing to report it accurately. They allow the assessor to ask directly about internal processes, policies, and intentions. Their weakness is that they rely on cooperation, interpretation, and selective disclosure. A questionnaire can measure the quality of a response as much as the quality of the underlying practice. It can also become burdensome when repeated across many clients, vendors, agencies, or jurisdictions. In fast-moving domains, a questionnaire may be outdated as soon as it is completed.

Audit and assurance methods provide depth. They can examine internal controls, sample records, verify documentation, and test evidence unavailable to outsiders. Their legitimacy is often grounded in professional standards and direct access; indeed, the expansion of audit into a general organizing principle of institutional life, what Power (1997) called the audit society, reflects precisely the assumption that verification requires ritualized, privileged examination. Their main limitations are cost structure and timing. Deep review is expensive. It is periodic. It usually covers a limited sample. It may not scale across thousands of entities or update when a public-facing system changes the next day. In many domains, audit answers a different question than continuous public observation. Audit asks whether a claim or control can be supported by internal evidence at a review date. D2R asks what the external evidence field shows over time.

Composite indices and rankings broaden coverage, but they frequently rely on lagging indicators, broad proxies, and weighting choices that are difficult for users to inspect. The OECD and Joint Research Centre handbook on composite indicators emphasizes the importance of theoretical framing, indicator selection, imputation, normalization, weighting, aggregation, robustness, and transparent presentation (OECD & Joint Research Centre, 2008). Methodological opacity has practical consequences. In sustainability ratings, Berg et al. (2022) document that prominent ESG raters assessing the same firms produce only modestly correlated ratings, and they attribute the divergence primarily to differences in measurement rather than in scope or weighting. When users cannot reconstruct how evidence became a rating, divergence cannot be diagnosed, errors cannot be located, and results cannot be contested. D2R builds on the composite-indicator measurement tradition but responds directly to this failure mode by adding a public-signal layer with mandatory evidence lineage: the indicators are features extracted from qualified public digital evidence, with a reconstructable path from source to rating.

D2R therefore occupies a different methodological position. It is less intrusive than audit, less dependent on cooperation than questionnaires, and more updateable than many conventional indices. Because D2R cannot directly inspect non-public processes, its value lies in structured external judgment under conditions of public digital visibility.

Table 2. Positioning D2R among assessment models

Model	Primary evidence	Strength	Structural limitation
Questionnaire or self-assessment	Entity-supplied answers and documents	Direct access to internal claims; low initial friction	Selective disclosure, response burden, self-report bias, weak refresh frequency
On-site audit or inspection	Privileged access, interviews, sampled records, control testing	Depth, professional legitimacy, ability to test non-public evidence	High cost, periodicity, limited scale, sampling constraints
Traditional composite index	Standardized datasets, surveys, official statistics, selected proxies	Coverage and comparability across populations	Lag, proxy drift, weighting opacity, limited connection to live digital behavior
Data-to-Rating (D2R)	Public digital evidence qualified as signals and transformed into features	Non-intrusive, scalable, updateable, externally comparable, evidence-traceable	Partial observability, public-visibility bias, gaming risk, need for confidence treatment

4. Definition and Scope of Data-to-Rating

Data-to-Rating (D2R) is a methodology for transforming publicly obtainable evidence into structured evaluative outputs through an explicit chain of signal qualification, feature extraction, indicator construction, scoring, rating translation, confidence assignment, and governed interpretation. The method is public-data based and valid only where the assessed construct has external manifestations and where the public evidence is rich enough to support bounded inference.

The term publicly obtainable is broader than open data and narrower than everything technically accessible. It includes evidence that can be lawfully accessed without privileged internal permission, deception, circumvention of access controls, intrusion, or confidentiality breach. It may include public websites, public registries, formal disclosures, open datasets, policy documents, service interfaces, public repositories, metadata, app store listings, public customer reviews, public procurement notices, regulatory filings, public dashboards, and other digital artifacts. It does not include hacked data, leaked private information, confidential records, or data obtained through unauthorized access. The method therefore has to define the public evidence boundary from the start.

The boundary should also specify whether the assessed domain is physical-world, institution-bound, digital-facing, or digital-native, because each type produces different forms of evidence and supports different levels of refresh frequency, automation, and confidence.

Physical-world domains should not be treated as wholly non-digital. Many physical settings now produce digitally mediated traces through public photographs and videos, street-view imagery, satellite imagery, inspection photos, user-generated review images, IoT or sensor dashboards, app-based records, and public multimedia disclosures. Computer vision and vision-language models can extract observable features from these traces, such as visible amenities, accessibility features, signage, safety equipment, maintenance conditions, queues, infrastructure progress, or service-environment characteristics. In hospitality, for example, public guest photos or videos may reveal visible room attributes such as a television, desk, refrigerator, balcony, or accessibility feature. In urban governance, street-view or satellite imagery may reveal aspects of infrastructure, land use, mobility, or the built environment. Research using street-view imagery and computer vision already demonstrates that visual public data can support large-scale characterization of built environments, and hotel-room image datasets show that visual traces from travel websites and crowd-sourced sources can support hotel recognition tasks (Fan et al., 2023; Stylianou et al., 2019). These traces extend D2R into partially physical domains, but only as bounded public signals. They should be qualified for legality, privacy, timestamp, provenance, field of view, manipulation risk, representativeness, and corroboration before they become features or indicators.

D2R is also representation-agnostic. A rating may be shown as a letter grade, a star, a tier, a percentile band, a color label, a numeric band, or a textual judgment. These formats are only representations; the method resides in the evidentiary chain that produces the output. This distinction matters because a rating symbol can create the illusion of precision; quantification carries an authority of mechanical objectivity that can outrun its evidentiary basis (Porter, 1995). A responsible D2R system must make the evidence basis, confidence level, and construct boundary visible enough for users to understand what the rating does and does not mean.

D2R is best understood as an assessment architecture with three layers. The evidence layer concerns what public material is admissible and how it is captured. The inference layer concerns how observable material becomes signals, features, indicators, scores, and confidence. The judgment layer concerns how scores are translated, represented, interpreted, challenged, updated, and used. Weakness in any layer undermines the entire output.

5. Conceptual Vocabulary: From Data to Judgment

A frequent weakness in rating systems is that the same word is used for different levels of abstraction. Data, signal, indicator, score, rating, and judgment are often collapsed into one label. D2R requires a stricter vocabulary because the legitimacy of the method depends on preserving the chain of transformation.

Raw data becomes evidence only when the methodology defines why it matters for a construct. This applies to web pages, files, registry entries, metadata fields, application interfaces, and disclosure statements. A signal is raw data judged relevant. A feature is an extracted property of one or more signals. An indicator is a feature or feature set mapped to an evaluative dimension. A score is a normalized numerical output. A rating is a categorical interpretation of that score. A representation is the form in which the rating is shown. Confidence is an assessment of evidentiary sufficiency and interpretive stability. Judgment is the human or governed decision about what the output means in context.

This separation allows D2R to avoid two common errors. The first error is data positivism: assuming that because something can be observed, it automatically measures what matters. The second error is rating opacity: presenting a final category without showing how public evidence became that category. D2R treats both errors as methodological failures.

Table 3. D2R vocabulary

Term	Meaning in D2R	Example
Data	Raw publicly obtainable material before evaluative interpretation	A public AI inventory page, privacy notice, service form, repository, registry record, or metadata field
Signal	Data judged relevant to the construct under a stated methodology	A published recourse pathway in an assessment of accountability
Feature	Extracted property or coded element derived from signals	Presence of contact channel; date of last update; disclosure completeness; response option type
Indicator	Feature or feature group mapped to an evaluative dimension	User-facing transparency; lifecycle monitoring evidence; accessibility maturity
Score	Normalized quantitative output generated from indicators	82 out of 100; 0.74 probability-like posture score
Rating	Categorical interpretation of the score	A, B, C; high, medium, low; advanced, developing, basic
Representation	User-facing symbol or notation used to display the rating	Stars, grade, badge, percentile band, numeric band, or narrative label
Confidence	Degree of evidentiary sufficiency, source reliability, recency, corroboration, and interpretive stability	High confidence because multiple current public artifacts corroborate the same finding
Judgment	Governed interpretation or decision use of the output	Treating a result as a screening input, audit trigger, or public transparency finding

6. The D2R Theory of Public Evidence

The foundational principle of D2R is the principle of observable construct manifestation. A public-data rating is valid only to the extent that the quality being assessed leaves observable traces and those traces can be lawfully obtained, qualified, normalized, corroborated, and challenged. The principle avoids two extremes: treating public data as superficial by default and treating it as the full truth of an institution. D2R occupies the bounded middle: public evidence can support structured judgment when the construct is externally manifested.

The strength of observable construct manifestation varies by domain. A software project may manifest relevant qualities through code, releases, dependencies, tests, issues, and build artifacts. A public digital service may manifest relevant qualities through user pathways, accessibility behavior, service availability, public forms, recourse channels, and update cadence. A governance practice may manifest relevant qualities through policies, inventories, decision records, disclosures, procurement notices, and public monitoring statements. An internal culture, by contrast, may leave weak or indirect public traces. D2R should therefore ask whether the construct has a credible public manifestation strong enough to support bounded inference, alongside the basic question of whether the evidence is public.

Different public signals have different evidentiary strength. A statement on a web page is weaker than a structured disclosure. A structured disclosure is weaker than an operational artifact. An operational artifact is weaker than repeated behavior observed over time. Multi-source corroboration is stronger than a single isolated trace. D2R therefore requires an evidentiary hierarchy. The hierarchy guides interpretation and prevents the methodology from treating all visible signals as equal.

The hierarchy has a theoretical rationale beyond intuition. In signaling theory, a signal is informative when it is differentially costly: entities that possess the underlying quality can produce the signal more cheaply than entities that do not (Spence, 1973). Public assertions are nearly costless for everyone and therefore carry little separating power. Operational artifacts are costlier to fabricate, because a working recourse channel, a maintained inventory, or a functioning accessibility interface requires real implementation behind the visible surface. Longitudinal behavior is costlier still, because consistency over time cannot be staged retroactively. Corroboration across independent sources raises the cost of misrepresentation multiplicatively, since each additional source must be separately produced and maintained in a mutually consistent state. The D2R hierarchy therefore orders public evidence by the cost of counterfeiting it, and this is why higher levels warrant greater interpretive weight whenever the construct demands operational substance rather than stated commitment.

This signaling logic also explains why the hierarchy matters in practice: public evidence can be polished, staged, or gamed. Many entities can publish a policy; fewer can maintain consistent public artifacts that align with the policy; fewer still can demonstrate longitudinal evidence of implementation, responsiveness, and correction. D2R should reward higher evidentiary levels when the construct requires operational substance rather than mere disclosure.

Table 4. D2R evidentiary hierarchy

Level	Evidence type	Interpretive value	Typical risk
0	Silence or absence	May indicate no public evidence, not necessarily poor performance	False negative if absence is treated as failure
1	Public assertion	Shows stated commitment or claim	Rhetorical compliance; weak proof of implementation
2	Structured disclosure	Shows organized publication of relevant information	Completeness and accuracy may be untested
3	Operational artifact	Shows a visible mechanism, interface, record, repository, or process	Artifact may exist but be unused, inaccessible, or outdated
4	Versioned or executable artifact	Shows that an artifact can be inspected, tested, compared across versions, or linked to a release or operational state	May reveal technical or procedural existence without proving broader effectiveness
5	Observed behavior over time	Shows consistency, update cadence, responsiveness, correction, or persistence	Observation window may be too narrow
6	Corroborated public evidence	Multiple independent signals converge on the same inference	Sources may share the same upstream error or bias

7. The D2R Methodological Pipeline

The D2R pipeline converts public data into structured assessment through a sequence of controlled transformations. The sequence matters because errors at early stages propagate into later outputs. A rating that appears simple at the representation layer may depend on hundreds or thousands of micro-decisions about source admissibility, feature extraction, indicator mapping, weighting, aggregation, and confidence.

The pipeline begins with construct definition. The assessor must define what is being assessed before collecting public data. Digital accessibility, service reliability, governance transparency, cybersecurity exposure, and public accountability require separate construct definitions. A weak construct definition produces measurement drift: the system begins scoring whatever is easy to observe rather than what it claims to measure.

After the construct is defined, the public evidence perimeter is specified. This perimeter lists admissible source classes, exclusion rules, time windows, jurisdictions, languages, and acquisition constraints. The source registry is then populated and evidence is collected. Candidate signals are qualified for authenticity, recency, provenance, relevance, comparability, and resistance to manipulation. Features are extracted manually, computationally, or through hybrid review. Indicators are normalized and mapped to dimensions. Scores are generated through documented weighting and aggregation rules. Ratings and representations are then assigned. Finally, confidence, evidence lineage, change logs, review procedures, and contestability mechanisms are attached to the output. Figure 1 summarizes this transformation chain from public data to governed judgment.

This process can be automated in parts, and in digital-native domains it can sometimes operate as a continuous or event-triggered pipeline. Software assessment provides the reference model: evidence can be captured when code is committed, a dependency changes, a build is produced, a test fails, a release is issued, or runtime telemetry changes. D2R translates this idea into public-data assessment. A new disclosure, policy revision, public inventory update, procurement notice, service change, incident communication, repository update, or interface modification may trigger remeasurement. Automation can collect, parse, classify, compare, and update, while human governance retains authority over construct definition, source admissibility, weighting, thresholds, exceptions, appeals, and consequential use.

Table 5. D2R pipeline and control points

Stage	Core question	Control point
1. Construct formulation	What quality, risk, maturity, or capacity is being assessed?	Definition of construct, scope, exclusions, and theory of observable manifestation
2. Population and evidence perimeter	Which entities and public sources are in scope?	Source classes, access boundaries, time windows, language and jurisdiction rules
3. Source registry and acquisition	Where will evidence be obtained and how will it be recorded?	Provenance, timestamps, version capture, legality, reproducibility
4. Evidence versioning and change detection	How are evidence changes recorded, compared, and used to trigger review?	Evidence snapshots, source timestamps, version history, change logs, refresh triggers, and reproducibility
5. Signal qualification	Which public data points are relevant and reliable enough to count?	Authenticity, recency, relevance, comparability, manipulation resistance
6. Feature extraction	What properties can be extracted from the signals?	Feature dictionary, coding rules, extraction quality, reviewer/model agreement

Stage	Core question	Control point
7. Indicator construction	How do features map to evaluative dimensions?	Normalization, missingness rules, dimension mapping, non-compensability rules
8. Scoring and aggregation	How are indicators converted into scores?	Weights, caps, thresholds, sensitivity testing, calibration
9. Rating and representation	How is the score translated for users?	Band definitions, labels, visual symbols, narrative explanation
10. Confidence, review, and update	How reliable, current, and contestable is the output?	Confidence scoring, evidence trace, change logs, appeal route, refresh cadence

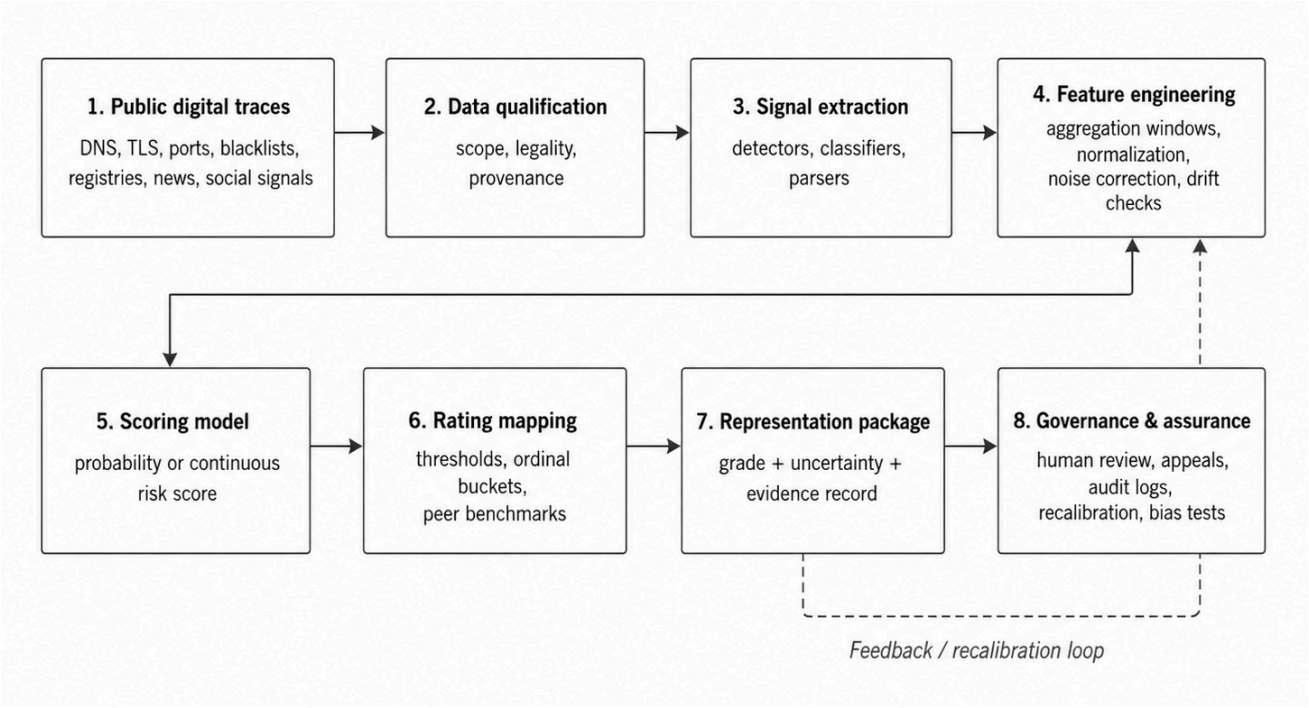

Figure 1. The D2R transformation chain, from public data to governed judgment

8. Scoring, Rating, and Confidence

D2R scoring should be simple enough to explain and flexible enough to fit different constructs. In formal terms, an entity e has a set of public data D_e. A methodology qualifies a subset of D_e as signals S_e. Features F_e are extracted from S_e. Features are mapped to indicators I_e across dimensions. A scoring function produces a score σ_e. A translation rule maps σ_e into a rating ρ_e. A representation function displays ρ_e to users. A confidence function assigns c_e based on evidence sufficiency and interpretive stability. The complete D2R output is therefore the package $(\rho_e, \sigma_e, c_e, E_e)$, where E_e is the evidence trace that supports reconstruction and challenge.

This notation can be extended into an explicit measurement model that makes the central inferential problem of D2R visible. Let θ_e denote the latent level of the construct for entity e. D2R does not observe θ_e directly. It observes public manifestations of that construct through indicators that can be represented as:

$$I_e = f(\theta_e) + v_e + g_e(M) + \varepsilon_e \quad (1)$$

In this notation, the subscript e refers to the assessed entity. I_e denotes the observed indicator or indicator vector for entity e. $f(\theta_e)$ denotes the public manifestation of the latent construct. v_e denotes visibility bias arising from differences in disclosure capacity, publication culture, and communications resources. $g_e(M)$ denotes strategic distortion produced when entities adapt their public-facing signals in response to the methodology M, including scoring rules, thresholds, and known indicators. ε_e denotes measurement noise introduced by collection, extraction, classification, attribution, or recency problems.

The equation is additive because it is a measurement-error formulation, not an aggregation formula. It does not say that D2R constructs are produced by adding independent dimensions. It says that the public indicator observed by a D2R system contains the public manifestation of the underlying construct plus identifiable sources of distortion. Multiplication, gates, or thresholds may still be appropriate later in the scoring model, especially where the methodology requires non-compensability.

A D2R score is therefore best understood as a bounded indicator of the publicly manifested component of θ_e, not as a direct estimate of the entity's complete internal condition. Signal qualification and population-relative normalization help address v_e. The evidentiary hierarchy and multi-source corroboration raise the cost of $g_e(M)$. Extraction controls, evidence snapshots, and review procedures reduce ε_e. Confidence, in this framing, is an explicit statement about the expected magnitude of the components that remain unsuppressed. The model clarifies the boundary of the claim: D2R estimates public manifestation under stated evidence conditions, and its validity argument must show that the gap between manifestation and construct is bounded, characterized, and disclosed.

The scoring model may be additive, hierarchical, threshold-based, probabilistic, rule-based, or hybrid. The choice depends on the construct. Some constructs allow compensability: weakness in one area can be offset by strength in another. Others do not. For example, in a public accountability assessment, a strong policy statement may not compensate for the complete absence of any public recourse channel. In an AI governance assessment, public transparency may not compensate for the absence of lifecycle monitoring where high-impact systems are involved. D2R should make these non-compensability rules explicit. Technically, non-compensability can be implemented through Boolean gates, minimum threshold rules, multiplicative aggregation, caps, or automatic downgrades that prevent strength in one dimension from masking the absence of a required condition in another.

Confidence should be treated as a core output. A low score and low confidence are different states. A low score may reflect strong negative evidence. Low confidence may reflect sparse evidence, uncertain attribution, inconsistent sources, or stale data. No evidence is also different from negative evidence. Public silence may be

a relevant signal in some transparency constructs, but it should not automatically be treated as operational failure in all constructs. D2R must therefore separate performance judgment from evidentiary sufficiency.

A practical confidence model should consider at least six variables: source authority, evidence recency, source coverage, cross-source corroboration, extraction reliability, and attribution certainty. A seventh variable, volatility, becomes important in continuous systems: if public signals change rapidly, the rating may require a narrower confidence interval or a shorter refresh cycle.

A further confidence factor is evidence instrumentability: the degree to which evidence can be captured, versioned, retested, and compared over time. Digital-native evidence such as repositories, structured datasets, machine-readable registries, public APIs, software bills of materials, and timestamped disclosures may support higher refresh frequency because changes can be detected and reconstructed. Narrative evidence, such as a static policy page or public statement, may still be relevant but usually supports lower confidence unless corroborated by operational artifacts or longitudinal behavior.

Table 6. D2R confidence states

State	Meaning	Responsible presentation
High score / high confidence	Positive evidence is current, relevant, and corroborated	Present rating with evidence summary and normal caveats
Low score / high confidence	Negative finding is supported by current and corroborated evidence	Present rating, evidence chain, and correction pathway
Any score / low confidence	Evidence exists but is sparse, stale, inconsistent, volatile, or weakly attributed	Attach low-confidence label; avoid strong categorical claims
No observable evidence	The method found no admissible public evidence for the feature or dimension	Separate from failure unless public visibility itself is the construct
Conflicting evidence	Public sources support different interpretations	Flag conflict; defer final rating or route to human review
Stale evidence	Evidence is relevant but outside the valid time window	Down-weight, mark expired, or trigger refresh

9. Automation, AI, and Human Judgment

D2R becomes operationally significant because public digital evidence is too large, heterogeneous, and dynamic for purely manual review at scale. Public evidence may be multilingual, semi-structured, distributed across thousands of websites, embedded in interface behavior, contained in public repositories, reflected in customer-review environments, or updated without notice. AI systems can assist in source discovery, document classification, entity resolution, signal extraction, feature coding, anomaly detection, change detection, translation, summarization, and cross-source comparison. Without computational assistance, D2R would remain possible in principle but weak in coverage, consistency, and refresh frequency.

The software domain shows how this assistance can become operational. Modern software assessment commonly uses automated checks across development and deployment pipelines: source inspection, dependency review, configuration analysis, vulnerability scanning, build verification, release monitoring, and runtime observation. Secure software development guidance, DevSecOps-oriented CI/CD pipeline guidance, and open-source scorecard methods all illustrate how software assessment can be organized as a repeatable evidence pipeline rather than as a purely manual review exercise (National Institute of Standards and Technology, 2022, 2024; Open Source Security Foundation, 2026). D2R adapts the architectural lesson from these tools: assessment can be organized as a pipeline in which evidence is continuously discovered, qualified, transformed, scored, reviewed, and refreshed. In D2R, the equivalent of a software build event may be a public disclosure update, a change in a registry, a new procurement record, a modified service interface, a new AI inventory entry, a public incident communication, a repository update, or a material change in user-facing evidence.

AI has three distinct roles in D2R, and these roles should not be collapsed. First, AI may be the object of assessment. This occurs when D2R is used to evaluate public AI governance transparency, AI tool worthiness, AI policy maturity, the accountability of AI deployments, or the observable governance of AI used in domains such as screening, recommendation, public administration, procurement, education, health, finance, or supply chains. In this role, the AI system or AI-enabled practice is what is being assessed. The relevant public signals may include AI use case inventories, model or system cards, procurement documents, user notices, responsible-use policies, risk classifications, monitoring statements, incident disclosures, complaint channels, appeal mechanisms, audit summaries, and public explanations of human oversight. D2R does not claim to reveal the full internal performance of the AI system. It assesses whether the public manifestations of governance, accountability, monitoring, and recourse are present, current, attributable, and credible.

D2R also intersects with AI benchmarking, but the two should not be conflated. AI benchmarks evaluate models, tools, or agents against defined task sets, such as reasoning, coding, safety, factuality, multilingual performance, or tool use. Benchmark results may provide useful public signals within a D2R evidence perimeter, especially when they are independently produced, current, reproducible, and aligned with the assessed construct. However, a benchmark score is not itself a D2R rating. D2R treats benchmark results as one evidence class among others and evaluates their relevance, provenance, contamination risk, update cadence, scoring transparency, and susceptibility to gaming. Conversely, D2R can also assess the worthiness of benchmarks themselves by examining whether they validly measure the claimed capability, resist overfitting, disclose methodology, support reproducibility, and remain meaningful as models improve. In this sense, AI benchmarks can inform D2R, and D2R can discipline the interpretation of AI benchmarks.

Second, AI may be an operational instrument inside the D2R pipeline. In this role, AI helps execute the method. It may discover candidate sources, classify documents, identify entities, resolve name variations, translate public materials, extract relevant passages, detect changes, compare sources, summarize long documents, identify anomalies, and convert raw material into candidate signals or coded features. This role is

especially important where evidence is fragmented, multilingual, or fast-changing. For example, a public-data assessment may need to compare an organization's website, regulatory filings, procurement records, public repositories, app store disclosures, customer-review patterns, and official policy statements. AI can reduce the cost of assembling this evidence field. Recent public-audit practice illustrates this operational role. An OECD review of AI in public audit reports emerging use cases in anomaly detection, risk assessment, document classification, predictive models, semantic search, retrieval-augmented knowledge management, intelligent document processing, drafting, summarisation, translation, and visual or spatial analysis (OECD, 2026b). These examples show that AI is already being used to expand the scale and granularity of oversight work, while also confirming a central D2R premise: computational tools may improve evidence processing, but they do not remove the need for source provenance, validation, explainability, and human review. However, operational AI does not by itself determine relevance, scoring, confidence, or final judgment. Its output must remain tied to source provenance, timestamps, extraction rules, and review status.

Third, AI may serve as a quality-control layer. Reviewer agents or judge agents can test whether extracted signals are relevant, attributable, current, corroborated, and consistent with the approved methodology. They can flag unsupported claims, stale sources, weak attribution, inconsistent coding, unexplained score changes, missing evidence, or narrative overstatement. They can also compare a draft assessment against the evidence trace and ask whether the rating language exceeds what the public evidence supports. This use is distinct from using AI to generate the initial assessment. One AI component may extract candidate signals, while another may review whether those signals have been used correctly. Even then, the reviewer agent is not a final authority. It is a control mechanism inside a governed process.

One specific technique is LLM-as-a-Judge, in which a large language model is used to evaluate another model output, extracted signal, coded feature, or draft assessment against a rubric. In D2R, such a technique may support quality control by checking whether public evidence has been classified consistently, whether a source supports a claimed signal, whether feature coding follows the approved methodology, or whether a draft narrative overstates the evidence. However, LLM-as-a-Judge should not be treated as a substitute for validation. It can reproduce position bias, verbosity bias, self-enhancement bias, rubric sensitivity, source-grounding failures, and reasoning errors (Zheng et al., 2023; Gu et al., 2024). Its outputs should therefore be logged, sampled, compared against human review, and treated as review assistance rather than final judgment.

These three roles imply different governance questions. When AI is the object of assessment, the question is whether public evidence supports a bounded judgment about the AI system or AI-enabled practice. When AI is the operational instrument, the question is whether the extraction and transformation process is reliable, reproducible, source-grounded, and auditable. When AI is the quality-control layer, the question is whether the reviewer system is itself calibrated, tested, and constrained. The failure to separate these roles creates methodological confusion. A system that uses AI to collect evidence should not be assumed to be able to judge that evidence. A system that judges extracted signals should not be assumed to validate the underlying construct. A system that assesses AI governance should not be confused with the AI tools used to perform that assessment.

This distinction is also important for autonomous or continuous D2R. AI can make continuous assessment operationally feasible by monitoring public sources, detecting material changes, proposing provisional updates, and triggering review when new evidence appears. But continuous operation should not mean autonomous legitimacy. A D2R system may automatically detect that an AI inventory has changed, that a recourse channel has disappeared, that a public repository has become inactive, or that customer-review patterns have shifted. It may also propose a confidence adjustment or trigger reassessment. Yet material downgrades, contested findings, threshold changes, high-stakes uses, and methodology revisions should

remain subject to human control. Automation can refresh evidence; it should not silently alter the meaning of the construct.

AI use in D2R also creates specific risks. Models may hallucinate sources, summarize selectively, confuse similarly named entities, miss jurisdictional context, mishandle multilingual evidence, over-weight fluent public statements, under-detect weak attribution, or convert ambiguous evidence into confident prose. Retrieval systems may exclude relevant sources because of indexing gaps. Entity-resolution systems may merge entities that should remain separate. Classification systems may treat a policy statement as implementation evidence. Reviewer agents may reward well-written narratives rather than evidentiary strength. These risks do not make AI unsuitable for D2R, but they require explicit controls: source grounding, evidence snapshots, model and prompt versioning, extraction logs, benchmark tests, human sampling, disagreement review, appeal procedures, and confidence penalties where automation is uncertain.

The broader direction of AI governance supports this cautious position. Work on post-deployment monitoring emphasizes the complexity of monitoring deployed AI systems and the need for continuing guidance (National Institute of Standards and Technology, 2026). Work on AI monitoring in financial systems similarly notes data gaps and the need for robust approaches to tracking adoption, vulnerabilities, and proxy indicators (Financial Stability Board, 2025). These are precisely the kinds of environments in which automation is valuable because evidence is large and changing, but governance remains necessary because interpretation is uncertain and consequential.

D2R therefore requires a human-control architecture. Humans should define constructs, approve evidence admissibility rules, validate feature dictionaries, set weights and thresholds, review sensitive downgrades, adjudicate contested results, approve material methodology changes, and decide how ratings may be used. AI can make D2R faster, broader, more consistent, and more continuously refreshable. It cannot supply the normative judgment that makes a rating legitimate. Governed judgment should determine the meaning, release, and consequences of any provisional result generated by automation.

10. Validity, Reliability, and Quality Assurance

A D2R output is only as strong as its validity argument. Validity requires more than predictive accuracy or face plausibility. A rating must be evaluated in relation to the construct it claims to measure, the evidence it uses, the transformation rules it applies, and the decisions it supports. This is consistent with classical construct-validity thinking, which treats validity as an evidentiary argument about interpretation rather than a property of a number alone (Cronbach and Meehl, 1955; Messick, 1995). Contemporary validity theory sharpens this further by framing validation as the construction and appraisal of an interpretive argument: a chain of inferences leading from observation to score to interpretation to use, each link of which must be separately supported (Kane, 2013). D2R adopts this argument-based structure directly. The pipeline stages in Section 7 each correspond to an inference that the methodology must defend, and the evidence trace keeps each link in the argument inspectable.

Content validity asks whether the selected public signals cover the relevant dimensions of the construct. If an assessment of public digital service quality measures only visual design, it may miss accessibility, recourse, language coverage, and transaction completion. Construct validity asks whether the rating measures the intended attribute rather than a convenient proxy. A public AI governance rating should distinguish polished policy language from actual monitoring arrangements. Criterion validity asks whether D2R outputs correlate with external evidence, such as audit findings, incident outcomes, user experience measures, or expert judgments, where such evidence is available. Discriminant validity asks whether the rating avoids measuring unrelated attributes, such as wealth, brand maturity, or communications capacity.

Reliability concerns reproducibility. If the same evidence is processed under the same methodology, the same output should result. If a methodology uses human coders, inter-rater agreement should be measured. If it uses AI extraction, model performance, drift, and error profiles should be monitored. If public sources change, the system should preserve timestamps and version histories so that prior ratings can be reconstructed.

Software assessment also suggests a stronger reproducibility principle for D2R: a rating should be reconstructable from an evidence snapshot, a methodology version, and a scoring configuration. A later reader should be able to determine which public sources were considered, when they were captured, how they were transformed into features, which weights and thresholds were applied, which confidence rules were used, and which human review decisions affected the final output. A D2R system can keep some operational details internal while maintaining an auditable record sufficient to distinguish methodological judgment from arbitrary scoring.

Quality assurance should include a minimum set of controls: source provenance, evidence snapshots, feature dictionaries, extraction logs, weighting documentation, sensitivity analysis, missing-data rules, threshold governance, confidence assignment, change logs, independent review, and appeal mechanisms. Composite-indicator practice already recognizes the importance of transparent indicator selection, weighting, aggregation, and robustness testing (OECD & Joint Research Centre, 2008). D2R extends those requirements into a public-digital-evidence environment. AI-assisted D2R adds a further quality-assurance problem: some evidence may be generated, extracted, classified, summarised, or interpreted by AI before it reaches the reviewer. This requires controls not only over the final score, but over the AI-mediated evidence chain itself. Public audit institutions report uncertainty about standards for handling evidence generated or analysed by AI, including how to validate, document, and explain such outputs (OECD, 2026b). A D2R system should therefore record when AI was used, which model or tool was involved, what source material was processed, what transformation occurred, what confidence or error checks were applied, and which human review decision accepted, rejected, or modified the AI-assisted output.

A mature D2R system should also be tested against strategic behavior. Once actors know that a signal is measured, they may optimize for the visible indicator rather than the underlying substance. D2R should therefore treat gaming as a validity threat and test whether measured signals remain connected to the construct after entities have incentives to improve their public-facing traces.

11. Illustrative Application: Public AI Governance Transparency

The illustration is synthetic and shows how D2R would operate in a public-data domain without evaluating real agencies, companies, or jurisdictions. Consider an assessment of public AI governance transparency. The construct is the visible governance of AI use, as demonstrated through public evidence, rather than overall AI safety, internal model quality, or legal compliance.

Concrete public artifacts already exist for this type of illustration. For example, the U.S. Department of Energy publishes a 2025 AI Use Case Inventory, and the U.S. Department of Health and Human Services maintains a public AI use case inventory (Department of Energy, 2026; Department of Health and Human Services, 2026). These examples show the type of public evidence artifact that a D2R method could use without implying any rating or evaluation of those agencies.

The evidence perimeter may include public AI use case inventories, agency or organizational AI policies, generative AI policies, model or system cards, procurement notices, impact assessment summaries, privacy notices, public feedback channels, documentation of waiver or risk determinations, incident communication, accessibility statements, and public repositories. These sources are public and can be reviewed without privileged access. They do not reveal all internal safeguards, but they can reveal whether governance commitments have visible public manifestations.

A D2R methodology might define five dimensions: public visibility of AI use, governance articulation, lifecycle monitoring evidence, user-facing transparency, and recourse or accountability. Features could include whether AI use cases are listed, whether high-impact status is identified, whether responsible offices are named, whether monitoring is described, whether affected users are told about AI involvement, whether feedback or appeal channels exist, and whether disclosures are updated over time. Each feature would be qualified by source recency and corroboration.

Table 7. Synthetic D2R output package for a public AI governance transparency assessment

Output element	Illustrative content	Methodological purpose
Construct	Public AI governance transparency	Prevents confusion with overall AI safety or legal compliance
Score	78 / 100	Summarizes normalized indicator performance
Rating	Advanced public transparency	Translates score into a categorical judgment
Confidence	Medium-high	Indicates evidence is current and corroborated, but some lifecycle evidence is only narrative
Evidence trace	Inventory page, policy page, updated recourse channel, public monitoring statement, timestamped copies	Allows reconstruction and challenge
Limitations	No internal testing data reviewed; public evidence does not prove operational effectiveness	Prevents overinterpretation
Update trigger	Inventory change, policy revision, new public waiver, incident disclosure, or quarterly refresh	Supports continuous assessment

12. Relationship to Audit, Inspection, and Questionnaire-Based Assessment

D2R should not be framed as a replacement for audit, inspection, or questionnaire-based assessment. Its proper role is complementary. Audit and inspection remain necessary where the relevant evidence is internal, confidential, physical, legally protected, or dependent on professional verification. D2R cannot, by default, test internal controls, inspect non-public records, interview responsible staff, verify confidential evidence, or issue statutory assurance. Its contribution lies elsewhere: it creates a continuous public-evidence layer that can reveal observable signals, detect changes between review cycles, identify low-confidence cases, and guide where deeper inquiry is needed.

This complementarity is visible in the emerging use of AI by public audit institutions. The OECD's review of AI in public audit describes AI as a tool that can improve productivity, responsiveness, anomaly detection, data extraction, and knowledge management, while remaining a complement to traditional audit methods rather than a substitute for professional judgment (OECD, 2026b). D2R occupies a similar position. It can strengthen the public-evidence layer around an assessment, but it should not be treated as statutory assurance, audit opinion, or proof of internal reality.

The relationship with audit is therefore best understood as a sequencing relationship. A D2R output may help decide where an audit should focus, what questions should be asked, which public claims require internal verification, and which changes have occurred since the last formal review. For example, a public-data assessment may show that an organization has recently changed its AI policy, removed a recourse channel, published a new inventory, modified a service interface, or disclosed an incident. None of these observations proves internal compliance or operational effectiveness. They do, however, create structured prompts for audit planning and expert review. D2R can therefore reduce the cost of preliminary evidence gathering while preserving the need for professional judgment where assurance is required.

The relationship with questionnaires is equally important. In many assessment systems, questionnaires are used as the primary evidence source because the assessor has no better way to obtain information. This makes the process dependent on self-reporting, respondent interpretation, and selective disclosure. D2R reverses the order. It begins with public observation and then uses questions to close evidence gaps, resolve conflicts, clarify attribution, or verify internal implementation. The questionnaire becomes a validation and clarification instrument rather than the first and only source of evidence.

This shift has practical consequences. A conventional questionnaire may ask whether an entity has a policy, a process, a contact point, an inventory, or an escalation pathway. A D2R process first checks whether those elements are publicly observable, current, attributable, and corroborated. The remaining questions become more precise: whether the published policy is implemented internally, whether the listed contact point is operational, whether the inventory is complete, whether the recourse channel is used in practice, and whether internal controls support the public claim. The burden moves from broad self-description toward targeted verification.

For audit and assurance professionals, D2R may shift value from manual collection of public evidence toward methodology assurance, evidence-chain verification, model oversight, exception review, and hybrid evaluation. A continuous public-data rating may flag anomalies, changes, or low-confidence cases. Human experts then focus on the points where judgment is most valuable: interpretation, contested evidence, threshold decisions, high-stakes use, and deeper investigation. In that sense, D2R does not weaken assurance. Properly governed, it can make assurance more selective, more timely, and more evidence-aware.

D2R also should not be presented as a replacement for online customer-review platforms or user-rating systems. Customer-review platforms capture subjective experience, satisfaction, complaint narratives, and

reputational signals from users who choose to participate. Those signals can be useful within a D2R evidence perimeter, but they serve a different function. D2R is not a substitute forum for consumer opinion, nor is it a mechanism for aggregating sentiment into popularity scores. Its role is to qualify public signals, test their evidentiary strength, compare them with other public artifacts, and incorporate them into a broader structured judgment only where the construct warrants it.

The boundary should remain clear. D2R can support screening, monitoring, benchmarking, prioritization, and structured public-evidence judgment. It should not be presented as proof of internal truth where only external traces have been assessed. Where decisions have legal, financial, reputational, or regulatory consequences, D2R outputs should be treated as inputs into judgment rather than substitutes for professional assurance.

13. Risks, Ethics, and Governance Safeguards

D2R carries risks that must be treated as design constraints. The first is visibility bias. Entities with stronger communication capacity may appear more mature because they publish more public material. Entities with fewer resources or different disclosure cultures may appear weaker despite sound internal practice. The second is proxy distortion. Public signals may correlate with the construct but not fully represent it. The third is gaming. Once measured signals are known, actors may optimize for visible indicators without improving underlying substance. Rankings, audits, and metrics create similar incentives, and public measures are known to reshape the behavior of the entities they evaluate (Espeland and Sauder, 2007). The pattern is general enough to have acquired the status of law: the more a quantitative indicator is used for consequential social decisions, the more it invites pressures that distort both the indicator and the process it is intended to monitor (Campbell, 1979), and a measure that becomes a target tends to cease being a good measure (Strathern, 1997). D2R should respond by using multiple signal classes, emphasizing operational artifacts and longitudinal behavior, rotating or auditing selected signals, and routing suspected gaming to human review. The fourth risk is attribution error, especially when public technical artifacts, domains, records, or third-party statements are incorrectly linked to an entity. Vendor-hosted traces should be attributed only when the methodology can establish a controlled relationship between the vendor artifact and the assessed entity, such as through an official domain link, procurement record, contractual disclosure, authenticated repository ownership, or corroborating public statement; otherwise, the evidence should be marked as weakly attributed or excluded. The fifth is automation bias: users may overtrust clean-looking computational outputs. The sixth is false precision, in which uncertain evidence is converted into a categorical label without adequate caution.

Customer reviews and user-generated ratings require separate caution. They may provide valuable public signals about service experience, accessibility barriers, responsiveness, trust, reliability, or recurring complaints, especially when patterns are consistent across platforms or over time. However, they are not neutral measures of performance. Review data may suffer from selection bias, demographic skew, platform moderation effects, fake or incentivized reviews, coordinated campaigns, complaint amplification, sentiment distortion, and unequal review cultures across countries, sectors, and customer groups. D2R should therefore avoid treating customer reviews as direct evidence of underlying quality. They should be used as weak-to-moderate signals, strengthened only through volume, recency, cross-platform corroboration, textual consistency, and alignment with other public artifacts such as complaint channels, regulatory records, service changes, or incident disclosures.

Visual and video-based public signals require additional safeguards. Images may contain bystanders, children, employees, private objects, location clues, or other personal data that are irrelevant to the assessed construct. They may also be stale, staged, selectively posted, edited, taken from atypical rooms or locations, or stripped of reliable metadata. D2R should therefore avoid relying on a single image or video to infer physical attributes unless the finding is low-stakes, clearly visible, time-bounded, and corroborated by other sources. Where personal data cannot be avoided, minimized, or separated from the relevant signal, the source should be excluded or handled under stricter governance.

Digital-native evidence creates an additional risk: the illusion of technical completeness. Because software-like artifacts can be captured, parsed, tested, and refreshed, users may assume that the rating is more complete than it is. A repository may be visible while deployment practices remain opaque. A software bill of materials may identify components and improve supply-chain transparency without proving secure design, secure deployment, or secure use (Cybersecurity and Infrastructure Security Agency, 2025). A public API may be testable without revealing governance quality. A service interface may appear functional without

demonstrating equitable access or user outcomes. D2R should therefore treat digital instrumentability as an evidence-collection advantage while still testing whether it supports construct validity.

There are also ethical and legal boundaries. Public accessibility does not automatically make every use appropriate. D2R should avoid personal data where the construct can be assessed at institutional level. It should respect access controls, robots policies where legally relevant, intellectual property limitations, privacy law, and contextual integrity, understood as the principle that information use should remain appropriate to the social context in which the information was produced or disclosed (Nissenbaum, 2010). D2R should not transform public traces into punitive judgments without a correction route. Where outputs may affect reputation, procurement, financing, access to services, or regulatory attention, procedural safeguards become essential.

A responsible D2R system should include at least seven governance safeguards. First, the methodology should be documented in a public or auditable methodology statement. Second, the evidence chain should be traceable. Third, confidence should be displayed with the rating. Fourth, results should be contestable through correction and appeal mechanisms. Fifth, methodology changes should be versioned and logged. Sixth, periodic recalibration and independent review should test drift, bias, and sensitivity. Seventh, high-stakes uses should require human review and should treat D2R outputs as inputs into judgment rather than final determinations.

These safeguards are integral to D2R and distinguish it from arbitrary public scoring. The scientific credibility of the framework depends on the ability to reconstruct, test, and challenge the path from public data to rating.

14. Research and Policy Agenda

The next stage of D2R research should move from conceptual architecture to domain-specific validation. Each application will require its own construct map, public evidence perimeter, feature dictionary, scoring logic, confidence model, and validity argument. A method for public AI governance transparency, for example, cannot simply reuse the same ontology as a method for cybersecurity exposure, digital service accessibility, procurement transparency, sustainability disclosure quality, sanctions screening, or AI tool worthiness. The architecture is portable, but the measurement design is domain-specific.

Future research should test six questions. First, construct validity: whether selected public signals actually measure the intended construct. Second, criterion validity: how D2R outputs compare with audit findings, expert assessments, incident data, user outcomes, benchmark results, or other independent reference points. Third, reliability: whether independent analysts or automated models produce consistent outputs under the same methodology. Fourth, uncertainty calibration: whether confidence labels accurately reflect evidence sufficiency and error risk. Fifth, cross-language and cross-jurisdictional robustness: whether the method treats entities fairly across different publication norms, legal environments, and disclosure cultures. Sixth, gaming resistance: how easily actors can improve visible signals without improving the underlying practice.

Future work should also examine how D2R relates to emerging AI-enabled oversight models. Public audit institutions are already exploring continuous intelligence, reasoning models, multimodal evidence analysis, automated quality assurance, and agentic AI systems that may support near-real-time awareness of public-sector risks (OECD, 2026b). These developments reinforce the need for D2R research to address not only scoring accuracy, but also evidence traceability, model governance, audit trails, human override, and the legal status of AI-assisted assessment outputs. As AI systems become more capable of collecting, interpreting, and reviewing public evidence, the methodological question becomes sharper: which parts of assessment can be automated, which parts require human judgment, and which parts require new standards before they can be used in high-stakes settings?

Empirical pilots should publish the construct definition, evidence perimeter, feature dictionary, weighting logic, missing-data rules, sensitivity analysis, confidence model, and limitations. Where independent ground truth is unavailable, pilots should state clearly that D2R measures public manifestations of a construct rather than internal reality. This distinction is essential to the credibility of the method.

D2R also has implications for transparency law and public-records policy. Many access-to-information regimes were designed around request-based disclosure: a person asks for a record, the public body searches for it, applies exemptions, and releases or withholds the material. That model remains essential for accountability, especially where records are not already public or where privacy, security, law-enforcement, or confidentiality review is required. However, continuous public-data assessment requires a complementary model: proactive publication of legally public records in machine-readable, timestamped, API-accessible, and versioned formats. Existing open-data and financial-transparency regimes already move in this direction, but coverage remains uneven across agencies, jurisdictions, and record types. Future policy could therefore distinguish between records that require case-by-case access review and records that should be continuously available as public digital infrastructure. For D2R, this would reduce dependence on episodic requests and allow public evidence to be refreshed, compared, audited, and challenged over time.

The distinction between public-sector transparency and market-disclosure regimes also matters. Public authorities and listed companies are subject to different transparency obligations. FOIA and public-records laws create request-based access rights against government bodies, subject to exemptions. Securities laws impose mandatory market-disclosure duties on public companies, including periodic and event-based filings, but they do not create a general right for the public to request corporate records. For D2R, both regimes matter

because both generate public signals. Their legal logic, coverage, exemptions, update frequency, and contestability mechanisms differ, however, and a D2R methodology should not treat them as equivalent source environments.

Policy pilots should therefore examine which public-record categories are suitable for proactive machine-readable release, which require privacy or security safeguards, which should remain subject to request-based review, and which can support continuous public assessment without creating unfair surveillance, false precision, or reputational harm. The policy challenge is not simply to make more data public. It is to make legally public data more usable, traceable, current, and contestable while preserving the boundaries that protect legitimate confidentiality, privacy, and security interests.

15. Conclusion

Data-to-Rating (D2R) responds to a structural change in the evidence environment. Assessment no longer has to begin only with a site visit, an internal file request, or a questionnaire. In digital-native domains, especially software, assessment already shows how structured artifacts can support continuous, pipeline-based, and event-triggered evaluation. Public digital evidence is now sufficiently dense in many broader domains to support systematic external observation, although with less completeness and formal precision than software. Public data cannot replace all internal evidence. The relevant question is whether, under a disciplined methodology, it can still be transformed into structured, comparable, updateable, and contestable judgment. D2R answers yes, under defined conditions.

The framework proposed in this paper treats rating as the end point of an evidence transformation chain. It distinguishes raw data from signals, features, indicators, scores, ratings, representations, confidence, and judgment. It introduces a theory of observable construct manifestation, an evidentiary hierarchy, a pipeline with control points, and a confidence model that separates low score, low confidence, no evidence, conflicting evidence, and stale evidence. It also insists that automation can support measurement execution but should not replace human authority over constructs, thresholds, disputes, and consequential use.

D2R is strongest where the assessed construct leaves public traces, evidence is sufficiently dense, signals can be normalized across a population, and outputs are presented with evidence lineage and uncertainty. It is weakest where evidence is sparse, private, easily staged, poorly attributable, or misaligned with the construct. D2R's promise lies in a new assessment layer that is non-intrusive, scalable, continuous, public-data based, and governed. If unobtrusive measurement was the social-scientific response to reactive research instruments (Webb et al., 1966), D2R is its institutional successor for a digitally visible world: nonreactive in collection, disciplined in inference, and accountable in judgment. In an era of digital visibility, that layer may become essential to the future of assessment and structured judgment.

Appendix A. Minimum D2R Methodology Statement

A deployable D2R assessment should publish or maintain an auditable methodology statement containing the following elements:

Appendix Table A1. Minimum methodology statement

Element	Required content
Purpose	The decision context and intended use of the rating
Construct	The quality, risk, maturity, or capacity being assessed and what is excluded
Target population	The entities, jurisdictions, systems, or services in scope
Public evidence perimeter	Admissible source classes, access limits, legal boundaries, and time window
Signal qualification rules	Criteria for relevance, authenticity, recency, provenance, comparability, and manipulation resistance
Feature dictionary	Definitions, extraction methods, coding rules, missingness treatment, and quality checks
Indicator mapping	How features connect to dimensions and why those dimensions represent the construct
Scoring and aggregation	Weights, caps, thresholds, non-compensability rules, and sensitivity analysis
Rating translation	Band definitions, category labels, representation rules, and user-facing explanations
Confidence model	Source authority, recency, coverage, corroboration, extraction reliability, attribution certainty, volatility, and evidence instrumentability
Governance	Human control points, change logs, review cadence, appeal process, and high-stakes-use restrictions
Limitations	Known blind spots, non-observable dimensions, bias risks, and conditions where the rating should not be used

Appendix B. Glossary

Appendix Table B1. Glossary

Term	Definition
D2R	Data-to-Rating: a methodology for transforming public data into structured evaluative outputs through governed evidence transformation.
Public evidence perimeter	The boundary defining which public sources are admissible, lawful, relevant, and within scope.
Observable construct manifestation	The principle that a public-data rating is valid only where the assessed construct leaves external traces that can be qualified and interpreted.
Evidence trace	The recorded path from source material to signal, feature, indicator, score, rating, and confidence.
Confidence	An assessment of evidentiary sufficiency and interpretive stability, distinct from the score itself.
Representation	The visual or symbolic format used to display a rating, such as grade, tier, star, or numeric band.
Contestability	The ability of affected parties or users to challenge source errors, attribution errors, stale evidence, or methodological misapplication.
Digital-native evidence	Evidence produced by systems or practices that are digital, versioned, machine-readable, executable, instrumented, or otherwise continuously observable by design.
Evidence instrumentability	The degree to which evidence can be captured, versioned, retested, compared over time, and used to trigger reassessment.
Event-triggered assessment	A rating update or review caused by a material change in public evidence, such as a new disclosure, repository update, service change, procurement record, incident communication, or registry modification.

Appendix C. Minimum Disclosure Template for a D2R Rating Report

A D2R rating report should include more than a standalone grade, score, star, tier, or category. Because D2R converts public evidence into structured judgment, the report must disclose enough context for a reader to understand what was assessed, what public evidence was considered, how current and reliable that evidence was, and what limits attach to the resulting judgment.

The template below defines a minimum disclosure structure for a D2R rating report. The template sets a minimum disclosure structure while allowing domain-specific applications to add peer comparison, indicator-level scoring, evidence ledgers, model-review notes, or validation results. However, a D2R report should not omit the elements needed for interpretation, reconstruction, and contestability.

Appendix Table C1. Minimum D2R Rating Report Structure

Report element	Required content	Methodological purpose
Rating outcome	The categorical rating assigned to the entity, such as advanced, developing, basic, high, medium, low, A/B/C, or another defined rating band.	Provides the user-facing judgment while making clear that the category is an interpretation, not the method itself.
Score	The numerical score and scale, such as 78/100 or 0.78, including the score range and whether higher or lower values are preferable.	Allows comparison and prevents the rating category from hiding the underlying measurement.
Confidence	A confidence label or score, with a short explanation of the evidence basis.	Separates the strength of the rating from the strength of the evidence supporting it.
Construct definition	The quality, risk, maturity, capacity, or condition being assessed, including what is excluded.	Prevents overinterpretation and clarifies that the rating applies only to the stated construct.
Unit of assessment	The entity, jurisdiction, service, system, platform, or organization assessed.	Clarifies the object of judgment and avoids attribution errors.
Assessment date	The date on which the rating was produced or approved.	Establishes the time of judgment.
Evidence window	The period during which public evidence was collected or considered.	Distinguishes current evidence from stale or historical evidence.
Evidence perimeter	The public source classes included and excluded, such as public websites, registries, disclosures, inventories, procurement notices, technical artifacts, or repositories.	Makes the boundary of public evidence visible and auditable.
Key positive signals	The strongest public signals supporting the rating.	Shows what evidence contributed positively to the assessment.
Key negative signals	The strongest public signals lowering the rating.	Shows what evidence contributed negatively to the assessment.
Missing or low-confidence evidence	Important areas where evidence was absent, stale, inconsistent, weakly attributable, or insufficient.	Prevents silence from being confused with verified failure and supports cautious interpretation.

Material limitations	Known limits of the assessment, including non-public information not reviewed, possible visibility bias, proxy limitations, and uncertainty.	Protects against false precision and misuse.
Evidence trace	A reconstructable record of the sources, timestamps, extracted signals, features, and scoring logic supporting the output.	Enables review, challenge, correction, and methodological accountability.
Last update	The most recent update to the rating or evidence base.	Supports continuous assessment and change tracking.
Next refresh or trigger	The scheduled refresh date or event-based trigger, such as new disclosure, policy update, system change, or corrected evidence.	Makes the rating dynamic rather than static.
Correction pathway	The procedure through which affected parties or users may challenge source errors, stale evidence, attribution mistakes, or methodological misapplication.	Establishes contestability as part of the rating process.

A minimal D2R rating report therefore contains four layers: the judgment layer consisting of rating, score, and confidence; the scope layer consisting of construct, unit of assessment, assessment date, and evidence window; the evidence layer consisting of positive signals, negative signals, missing evidence, limitations, and evidence trace; and the governance layer consisting of update rules and correction pathway.

This structure reflects a central principle of D2R: a rating becomes complete only when the category is accompanied by enough context to make the judgment interpretable, bounded, and contestable.

References

Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. Review of Finance, 26(6), 1315-1344. https://doi.org/10.1093/rof/rfac033

Campbell, D. T. (1979). Assessing the impact of planned social change. Evaluation and Program Planning, 2(1), 67-90. https://doi.org/10.1016/0149-7189(79)90048-X

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281-302. https://doi.org/10.1037/h0040957

Cybersecurity and Infrastructure Security Agency. (2025). Software Bill of Materials (SBOM). https://www.cisa.gov/topics/information-communications-technology-supply-chain-security/sbom

Department of Energy. (2026). DOE AI Use Case Inventory. https://www.energy.gov/cet/doe-ai-use-case-inventory

Department of Health and Human Services. (2026). HHS Artificial Intelligence Use Cases Inventory. https://www.hhs.gov/programs/topic-sites/ai/use-cases/index.html

Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. American Journal of Sociology, 113(1), 1-40. https://doi.org/10.1086/517897

Espeland, W. N., & Stevens, M. L. (1998). Commensuration as a social process. Annual Review of Sociology, 24, 313-343. https://doi.org/10.1146/annurev.soc.24.1.313

Fan, Z., Zhang, F., Loo, B. P. Y., & Ratti, C. (2023). Urban visual intelligence: Uncovering hidden city profiles with street view images. Proceedings of the National Academy of Sciences, 120(27), e2220417120. https://doi.org/10.1073/pnas.2220417120

Financial Stability Board. (2025). Monitoring adoption of artificial intelligence and related vulnerabilities in the financial sector. https://www.fsb.org/2025/10/monitoring-adoption-of-artificial-intelligence-and-related-vulnerabilities-in-the-financial-sector/

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Liu, W., Li, J., Wang, Z., & Guo, J. (2024). A survey on LLM-as-a-Judge. arXiv. https://arxiv.org/abs/2411.15594

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1-73. https://doi.org/10.1111/jedm.12000

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Computational social science. Science, 323(5915), 721-723. https://doi.org/10.1126/science.1167742

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. Science, 343(6176), 1203-1205. https://doi.org/10.1126/science.1248506

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

National Institute of Standards and Technology. (2022). Secure Software Development Framework (SSDF) Version 1.1: Recommendations for Mitigating the Risk of Software Vulnerabilities. NIST Special Publication 800-218. https://csrc.nist.gov/pubs/sp/800/218/final

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

National Institute of Standards and Technology. (2024). Strategies for the Integration of Software Supply Chain Security in DevSecOps CI/CD Pipelines. NIST Special Publication 800-204D. https://csrc.nist.gov/pubs/sp/800/204/d/final

National Institute of Standards and Technology. (2026). Challenges to the monitoring of deployed AI systems. NIST AI 800-4. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-4.pdf

Nissenbaum, H. (2010). Privacy in context: Technology, policy, and the integrity of social life. Stanford University Press.

OECD. (2023). 2023 OECD Open, Useful and Re-usable Data (OURdata) Index: Results and key findings. OECD Public Governance Policy Papers, No. 43. OECD Publishing. https://doi.org/10.1787/a37f51c3-en

OECD. (2026a). Digital Government Index and Open, Useful and Re-usable Data Index: 2025 results and key findings. OECD Publishing. https://www.oecd.org/en/publications/digital-government-index-and-open-useful-and-re-usable-data-index_6347ec74-en.html

OECD. (2026b). The state of artificial intelligence in public audit: Evidence from selected countries and the European Union. OECD Artificial Intelligence Papers, No. 58. OECD Publishing.

OECD & Joint Research Centre. (2008). Handbook on constructing composite indicators: Methodology and user guide. OECD Publishing. https://doi.org/10.1787/9789264043466-en

Office of Management and Budget. (2025). M-25-21: Accelerating Federal use of AI through innovation, governance, and public trust. The White House. https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf

Open Source Security Foundation. (2026). OpenSSF Scorecard. https://scorecard.dev/

Porter, T. M. (1995). Trust in numbers: The pursuit of objectivity in science and public life. Princeton University Press.

Power, M. (1997). The audit society: Rituals of verification. Oxford University Press.

Spence, M. (1973). Job market signaling. The Quarterly Journal of Economics, 87(3), 355-374. https://doi.org/10.2307/1882010

Stylianou, A., Xuan, H., Shende, M., Brandt, J., Souvenir, R., & Pless, R. (2019). Hotels-50K: A global hotel recognition dataset. arXiv. https://arxiv.org/abs/1901.11397

Strathern, M. (1997). 'Improving ratings': Audit in the British university system. European Review, 5(3), 305-321.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). Unobtrusive measures: Nonreactive research in the social sciences. Rand McNally.

World Bank. (2025). GovTech Maturity Index (GTMI). https://www.worldbank.org/en/programs/govtech/gtmi

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv. https://arxiv.org/abs/2306.05685