FAIR LABS

FAIR ARTIFICIAL INTELLIGENCE RESEARCH LABS

STANDARDS DOCUMENT ◆ FAIRNESS & BIAS ◆ 2025

AI Transparency Standards

Normative standards for the transparency and explainability of artificial intelligence systems, with technical specifications, documentation schemas, and a tiered conformance framework.

PROGRAM	DOCUMENT TYPE	AUDIENCE	PUBLISHED
Fairness & Bias	Standards Document	Industry & Compliance	2025

FAIR Labs · A 501(c)(3) nonprofit research institute · Washington, DC

CONTENTS

Table of Contents

EXECUTIVE SUMMARY

What this standard requires, in brief

IN ONE PARAGRAPH

A conforming organization shall document its AI systems across four layers—system, model, data, and decision—using standardized artifacts with specified required fields; shall offer explanations calibrated to their audience and honest about their faithfulness; shall give individuals subject to consequential automated decisions meaningful notice, an intelligible explanation, and a route to appeal; and shall manage the genuine tensions between transparency and privacy, security, gaming, and intellectual property through documented, defensible tradeoffs rather than blanket opacity. The depth of every obligation is set by one of three conformance levels, chosen according to the system's consequence and context.

4
transparency layers that together define what "understandable" means

3
conformance levels—Baseline, Enhanced, Comprehensive—scaling every requirement

5
distinct audiences whose differing needs the standard must serve at once

Core requirements

1 **Document across all four layers.**
Every conformant system *shall* maintain current documentation at the system, model, data, and decision layers, using the standardized schemas in Section III. Documentation is the substrate of transparency; without it, every other requirement is unverifiable.

2 **Match explanation to audience and be honest about faithfulness.**
Explanations *shall* be intelligible to their intended recipient and *shall* disclose their method and its known limits. A plausible explanation that does not reflect the model's actual reasoning is worse than none.

3 **Owe the affected individual notice, explanation, and appeal.**
For consequential automated decisions, the individual *shall* be told a system was involved, given a reason they can understand and act on, and offered a contestation route that reaches a human with authority to change the outcome.

4 **Resolve tradeoffs deliberately, not by default.**
Where transparency conflicts with privacy, security, IP, or gaming resistance, the organization *shall* record the conflict, the mitigation chosen, and the residual disclosure gap—rather than invoking any of these interests as a reflexive shield against scrutiny.

The sections that follow specify each requirement in normative detail, with field-level schemas, conformance tables, and assessment evidence that an organization can lift directly into policy and procurement.

SECTION I

Purpose, Scope & Audiences

Transparency is invoked so often in artificial-intelligence policy that it has begun to lose its edges. It is asked to mean, at once, that a system is documented, that its outputs are explainable, that its data are disclosed, that its decisions are contestable, and that its very existence is announced to those it touches. These are related but distinct obligations, owed to different people for different reasons, and a standard that runs them together will satisfy none of them. The purpose of this document is to give the word structure: to specify, at the level of artifacts and fields, what an organization *shall* do to substantiate a claim of transparency, and to scale those obligations to the stakes.

This standard rests on a distinction that recurs throughout. **Transparency** is the availability of accurate information about a system—what it is, how it was built, what data shaped it, and what it decided. **Explainability** is the narrower discipline of rendering a system's behavior, especially a particular decision, intelligible to a particular audience. Transparency is a property of the record; explainability is a property of the account given from that record. An organization can be transparent about a model it cannot fully explain, and it can produce fluent explanations while concealing the information that would let anyone check them. The standard requires both, and treats them as separate obligations with separate evidence.

Scope of application

This standard applies to any deployed system whose behavior is materially determined by a statistical model learned from data, together with the data pipelines, human procedures, and interfaces surrounding it. It applies whether the system is built in-house, procured, or accessed as a service, and whether the model is a narrow classifier or a general-purpose foundation model adapted to a task. The unit of conformance is the *deployed system in its context of use*, not a model file in isolation; identical model weights placed in two different decision processes are two different objects of this standard, because the transparency each context owes to those it affects differs.

Certain uses fall outside the core obligations while remaining subject to the standard's honesty requirements. Purely internal productivity tools whose outputs are reviewed and rewritten by a human before they reach anyone, and research systems not used to make decisions about people, are exempt from the individual-disclosure requirements of Section V but *shall* still be documented under Section III at the Baseline level. Nothing in this standard licenses an organization to mischaracterize a consequential system as internal or experimental to escape its obligations; the classification *shall* reflect actual use, and drift in use *shall* trigger reclassification.

The audiences transparency must serve

The central design problem of any transparency standard is that "transparency to whom?" has no single answer. Information that satisfies a regulator may bewilder an affected member of the public; detail that empowers a downstream developer may hand an adversary a manual for evasion. The standard therefore names its audiences explicitly and requires that each disclosure identify the audience it is written for. Five audiences recur.

THE FIVE AUDIENCES AND WHAT EACH ONE NEEDS FROM TRANSPARENCY

Audience	What they need to do	Information that serves them
Affected individuals	Understand and contest a decision that affects them	Plain-language notice, a specific and actionable reason, an appeal route
Operators & deployers	Use the system correctly and know when to doubt it	Intended use, limitations, performance by condition, failure modes
Auditors & regulators	Verify claims and assess compliance independently	Full documentation, evaluation evidence, logs, data provenance
Downstream developers	Integrate the system safely into a larger product	Interfaces, training context, out-of-scope uses, update policy
The general public	Form a legitimate view of an institution's use of AI	System existence, purpose, governance, aggregate performance

These audiences are not merely different in sophistication; their interests can point in opposite directions. The affected individual wants the maximum intelligible detail about the decision that touched them. The security team wants the minimum detail that would let an attacker reconstruct the model or engineer inputs to defeat it. A standard that pretends this tension away will be quietly ignored in practice. Section VI treats the tradeoffs directly; here we establish only the principle that **audience is a first-class attribute of every disclosure**, and that a conforming organization *shall* be able to state, for any piece of information it withholds, which audience it is withheld from and why.

A FRAMING RULE

Transparency is not the same as data dumping. Publishing a gigabyte of undifferentiated logs satisfies no audience and can obscure more than it reveals. Under this standard, a disclosure is transparent only if it is *accurate*, *relevant to a named audience*, and *intelligible to that audience*. Volume is not a virtue.

Relationship to law and other standards

This document is a voluntary standard. It is written to be compatible with, and to operationalize, the transparency and accountability expectations found in mainstream frameworks—among them the

risk functions of national AI risk-management guidance and the intergovernmental principles on trustworthy AI—without restating them. Where a legal obligation is more demanding than a requirement here, the law governs; where this standard is more demanding, adopting it is a way for an organization to exceed the floor deliberately. The conformance framework in Section VIII is designed so that a claim of conformance is meaningful: it names a level, identifies the evidence, and can be checked.

The remainder of the standard proceeds from structure to specifics. Section II defines the four layers at which transparency is produced. Sections III through V specify the artifacts, the explanations, and the disclosures owed to individuals. Section VI confronts the tradeoffs, Section VII adapts the requirements by sector, and Section VIII sets out how conformance is claimed and assessed.

SECTION II

The Four Layers of Transparency

Transparency is produced at four distinct layers—system, model, data, and decision. Each answers a different question, each can be strong while another is weak, and a claim of transparency is only as good as its weakest relevant layer.

Organizations often speak of "explaining the model" as though that alone discharged their duty. It does not. A perfectly explained model trained on undisclosed data, embedded in an undescribed system, producing decisions no individual can trace, is opaque in every way that matters to the people it affects. This section defines the four layers so that an organization can locate its obligations and an assessor can see where a program is hollow. The layers are cumulative in scope but independent in strength: doing one well does not excuse neglecting another.

The system layer

The system layer describes the whole deployed capability in its context: what problem it addresses, where it sits in a human process, who is accountable for it, and what its existence means for the people who encounter it. This is the layer at which the public and affected individuals first meet a system. A conforming organization *shall* maintain a *system card* that states the system's purpose, its operational context, the decisions it informs or makes, the humans in the loop, and the governance under which it operates. The system layer is where transparency about *existence* lives: the plain fact that an automated system is involved at all, which is logically prior to any explanation of how it works.

The model layer

The model layer describes the learned component: what kind of model it is, what it was trained to do, how it performs across conditions and groups, where it is known to fail, and what uses fall outside its intended scope. This is the layer served by the *model card*. The model layer answers the operator's and developer's questions—*can I rely on this here?*—and supplies the auditor with the performance evidence needed to test the organization's claims. Crucially, model-layer transparency includes disaggregated performance: an aggregate accuracy figure that hides a concentrated failure in one subgroup is a model-layer transparency failure even if every number reported is true.

The data layer

The data layer describes what the model learned from: where the data came from, whom it represents and whom it omits, how it was collected and labeled, and what historical decisions—possibly biased ones—are encoded in the training signal. This is the layer served by the *datasheet*. Data-layer transparency is frequently the weakest in practice, because data provenance is laborious to reconstruct and commercially sensitive to disclose. Yet it is often the most explanatory layer of all:

a great deal of what a model does, and much of how it fails, is inherited directly from its data. A standard that does not require data-layer transparency permits organizations to disclaim responsibility for pathologies they built in.

The decision layer

The decision layer describes individual outputs: for a given case, what the system produced, on what basis, and how an affected person can understand and contest it. This is the layer of *local explanation*, of logging, and of the appeal right specified in Section V. Decision-layer transparency is what turns the abstract fact of a documented system into something an individual can actually use. A system may be exemplary at the other three layers and still fail a person entirely if, when she asks why she was denied, no one can produce a specific, truthful, actionable reason.

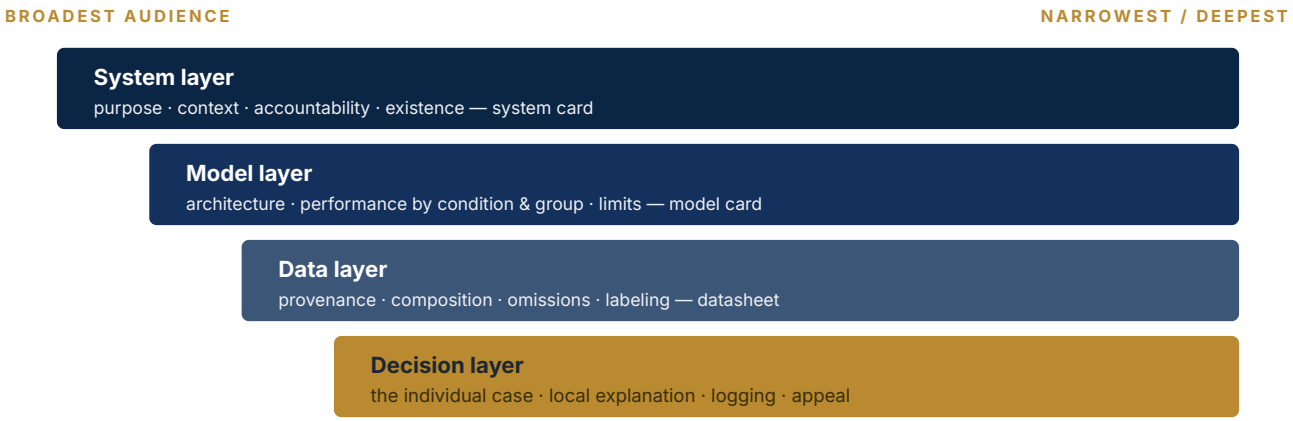

Figure 1. The four layers of transparency, from the system as a whole down to the single decision. Higher layers reach a broader audience; lower layers reach fewer people but at greater depth. A conforming program produces evidence at every layer relevant to its use.

LAYER RULE

A transparency claim *shall* specify which layers it covers. "This system is transparent" is not a conformant claim; "this system meets Enhanced conformance at the system, model, and decision layers, and Baseline at the data layer" is. Naming the layers prevents strength in one from being marketed as strength in all.

The four layers structure the rest of this standard. Section III specifies the documentation artifacts for the system, model, and data layers; Sections IV and V address the decision layer, where explanation meets the affected individual. Keeping the layers distinct is what allows an organization to be precise about what it has done—and honest about what it has not.

SECTION III

Documentation Standards & Required Fields

Documentation is the substrate of transparency. Every other requirement in this standard—explanation, disclosure, appeal, audit—depends on the existence of an accurate, current record of what a system is and how it was built. Where that record is absent, transparency claims cannot be verified, and an assessor is reduced to taking the organization's word. This section specifies the three primary documentation artifacts—the *model card*, the *datasheet*, and the *system card*—and the fields each *shall* contain at each conformance level.

The artifacts are deliberately modeled on established practice. Model cards, introduced to report model performance across conditions and groups, and datasheets, introduced to document data provenance and composition, are widely used and well understood; this standard adopts them, fixes a minimum field set, and binds each field to a conformance level. The point is not novelty but discipline: a model card that omits its intended-use boundaries or its subgroup performance is not a conformant model card, however polished it looks.

The model card

A conformant model card *shall* allow a competent reader to answer four questions: what is this model, how well does it work and for whom, where does it fail, and what must I not use it for. The required fields below are the minimum; an organization *may* add more.

MODEL CARD — REQUIRED FIELDS BY CONFORMANCE LEVEL

Field	Baseline	Enhanced	Comprehensive
Model identity & version	Required	Required	Required, with change history
Intended use & out-of-scope uses	Required	Required	Required, with rationale
Architecture & type (summary)	Required	Required	Required, with detail for audit
Aggregate performance metrics	Required	Required	Required
Disaggregated (subgroup) performance	—	Required	Required on deployer population
Known limitations & failure modes	Summary	Required	Required, with examples
Evaluation data & method	Named	Described	Reproducible by an auditor
Ethical & safety considerations	—	Required	Required

The datasheet

A conformant datasheet *shall* let a reader understand what population the model actually learned from and how that population was assembled. The datasheet is the layer at which many fairness problems become visible before they become harms. Its required fields track the lifecycle of the data itself: motivation, composition, collection, preprocessing, and intended and discouraged uses.

DATASHEET — REQUIRED FIELDS BY CONFORMANCE LEVEL

Field	Baseline	Enhanced	Comprehensive
Purpose & motivation for the dataset	Required	Required	Required
Composition (what instances, how many)	Summary	Required	Required, with distributions
Population represented & omitted	—	Required	Required, with gap analysis
Collection process & consent basis	Summary	Required	Required, with provenance chain
Labeling method & labeler context	—	Required	Required, with quality evidence
Known biases & sensitive attributes	Summary	Required	Required, with mitigation notes
Preprocessing & cleaning steps	Named	Described	Reproducible
Recommended & discouraged uses	Required	Required	Required

The system card

The system card sits above the model and data artifacts and describes the deployed whole. Where a model card describes a component, a system card describes the capability an affected person actually encounters. It is the artifact most useful to the public and to accountability bodies, and it *shall* be written to be intelligible without specialist knowledge. A conformant system card states the system's purpose and context, the decisions it touches, the human oversight around it, the models and data it relies upon (by reference to their cards), and the governance and contact points for questions and complaints.

CURRENCY REQUIREMENT

Documentation *shall* be current. A model card that describes a superseded model version, or a datasheet that predates a change in the data pipeline, is not merely incomplete—it is misleading, and a misleading record is worse than an acknowledged gap. Every artifact *shall* carry a version and a last-reviewed date, and material changes to the system *shall* trigger review of the affected artifacts before the change takes effect on consequential decisions.

Documentation as a living obligation

These artifacts are not deliverables produced once at launch and filed away. Under this standard they are living records, reviewed on a defined cadence and upon material change, and their existence and currency are themselves subject to assessment in Section VIII. An organization that produces immaculate documentation at procurement and never updates it has satisfied the letter of a checklist while defeating the purpose of the standard. The obligation is continuous because the systems are.

> *A model card that is not maintained is not documentation; it is archaeology.*
>
> — Working maxim of this standard

Finally, documentation *should* be layered for its audiences. The same underlying facts can be rendered as a one-paragraph public summary in a system card, a structured technical model card for developers, and a fully reproducible evaluation appendix for auditors. Producing these views from a single maintained source is both efficient and honest: it ensures that the public summary and the auditor's detail are two windows onto the same truth, not two incompatible stories.

SECTION IV

Explainability Methods & Their Limits

An explanation is a claim about why a system behaved as it did. Under this standard an explanation is conformant only if it is intelligible to its audience, faithful to the system it describes, and honest about the method that produced it and that method's limits.

Explainability is where transparency meets the hardest technical realities. Many high-performing models are not directly interpretable; their behavior emerges from millions of parameters no human reads. In response, a field of explanation methods has grown up—some producing global accounts of what a model has learned, others producing local accounts of a single decision. These methods are valuable, but they carry a danger the standard treats seriously: a fluent, plausible explanation can create the *impression* of understanding while bearing little relation to the model's actual computation. The central requirement of this section is therefore not "provide an explanation" but "provide an explanation whose faithfulness you can characterize."

Global versus local explanation

A **global** explanation describes the model's general behavior: which features it weighs, how its outputs vary across the input space, what patterns it has learned. Global explanations serve operators, developers, and auditors trying to decide whether a system is fit for a use. A **local** explanation describes a single decision: why *this* applicant, *this* case, received *this* output. Local explanations serve the affected individual and the appeal process. The two are not interchangeable. A faithful global account of a model can coexist with local explanations that are unstable or misleading, and vice versa. A conforming organization *shall* be explicit about which it is providing to whom.

GLOBAL AND LOCAL EXPLANATION — PROPERTIES AND APPROPRIATE AUDIENCES

Property	Global explanation	Local explanation
Question answered	How does the model behave in general?	Why this specific decision?
Primary audience	Operators, developers, auditors	Affected individuals, appeal reviewers
Typical form	Feature importance, behavior across ranges, learned rules	Contributing factors for one case, counterfactuals
Characteristic risk	Averages that hide local exceptions	Instability; plausible but unfaithful attributions
What it cannot do	Justify an individual outcome	Certify the system as a whole

Faithfulness: the property that matters most

An explanation is **faithful** to the degree that it reflects the actual basis of the system's behavior rather than a post-hoc rationalization that merely sounds right. Faithfulness is the hinge on which the value of any explanation turns, because an unfaithful explanation does not reduce opacity—it disguises it, and does so more dangerously than silence because it invites misplaced trust. This standard requires that any explanation offered for a consequential decision be produced by a method whose faithfulness has been characterized, and that known faithfulness limitations be disclosed to the audiences who might otherwise over-rely on the explanation.

THE CARDINAL RULE OF EXPLANATION

A plausible explanation that does not reflect the system's real reasoning is *worse* than no explanation. It manufactures false confidence, deflects legitimate challenge, and can make a biased decision look principled. Where an organization cannot produce a faithful account of a consequential decision, it *shall* disclose that limitation rather than substitute a comforting fiction.

Interpretable models versus post-hoc explanation

There is a standing choice between two routes to explainability. The first is to use an inherently **interpretable** model—one simple enough that its logic can be read directly—accepting whatever performance cost that entails. The second is to use a complex model and attach **post-hoc** explanation methods that approximate its reasoning. Each has a place. For the highest-consequence decisions, this standard *should* favor interpretable models where they can do the job, precisely because their explanations are faithful by construction. Where a complex model is genuinely necessary, the burden shifts to demonstrating that its post-hoc explanations are faithful enough for the use, and to disclosing where they are not.

The limits of explanation, stated plainly

This standard does not pretend that every system can be fully explained. Three limits recur and *shall* be acknowledged rather than papered over. First, some faithful explanations are simply complex, and rendering them intelligible to a lay audience necessarily loses information; the remedy is layered explanation, not false simplicity. Second, explanation methods can themselves be unstable, giving different accounts of nearly identical cases; an organization *shall* test for and disclose such instability where decisions are consequential. Third, explanation can be weaponized for evasion: publishing exactly which features drive a fraud model invites its defeat. Section VI addresses this last tension directly. The honest position is that explanation is a genuine good with genuine limits, and a standard earns trust by naming both.

A **Characterize faithfulness before you rely on an explanation.**

For any explanation offered for a consequential decision, the organization *shall* know, and be able to state, how well the method reflects the model's actual basis for that class of decision.

B **Prefer interpretable models for the highest stakes.**

Where a decision affects rights, liberty, or livelihood, an interpretable model *should* be used if it can meet the need, so that explanations are faithful by construction rather than by approximation.

C **Disclose the method and its limits.**

An explanation *shall* travel with a statement of how it was produced and where it is known to be incomplete or unstable, so that no audience mistakes a convenient story for the whole truth.

SECTION V

Disclosure to Affected Individuals

Everything to this point serves institutions, operators, and auditors. This section concerns the person on the other side of the decision—the applicant denied a loan, the traveler flagged for screening, the tenant refused a lease—whose stake in transparency is the most direct of all and the most easily neglected. For that person, the standard specifies three obligations that together turn transparency from an institutional virtue into an individual right: **notice** that a system was involved, a **meaningful explanation** of the decision, and a **route to appeal** it. For consequential automated decisions, these are requirements, not courtesies.

Notice

The most basic transparency owed to an individual is the fact that an automated system was materially involved in a decision about them. A person cannot question what they do not know exists. A conforming organization *shall* provide notice, at or before the point of decision, that an AI system contributed to the outcome; *shall* state in plain language what the system did; and *shall* identify how the person can obtain an explanation and pursue a challenge. Notice buried in a terms-of-service document is not notice for the purposes of this standard; it *shall* be provided in a form the affected person can reasonably be expected to see and understand.

Meaningful explanation

Notice without explanation informs a person that they were judged by a machine but leaves them unable to respond. A **meaningful explanation** is one that a non-specialist can understand and, critically, *act on*: it identifies the principal factors behind the specific decision in terms the person can recognize and, where applicable, address. "Your application did not meet our model's threshold" is not meaningful. "Your application was declined primarily because of the length of your credit history and a recent missed payment; here is how the decision is made and how to request review" approaches meaningfulness because it gives the person something to understand and something to do.

THE ACTIONABILITY TEST

An explanation offered to an affected individual is meaningful under this standard only if it passes a simple test: *could a reasonable person, reading it, understand why the decision went as it did and identify what, if anything, they could do about it?* An explanation that fails this test may satisfy a technical audience and still fail the person it was owed to.

Meaningfulness interacts directly with the faithfulness requirement of Section IV. An explanation to an individual *shall* be both intelligible and faithful; intelligibility purchased by misrepresenting the actual basis of the decision is not meaningful but manipulative. Where the true basis of a decision cannot be rendered both faithful and intelligible, that difficulty is itself material information: it bears on whether the decision should have been automated at all, a question Section VII and Section VIII return to.

The right to contest

Explanation exists so that a decision can be challenged. A conforming organization *shall* provide, for consequential automated decisions, a route by which the affected person can contest the outcome and have it reviewed by a human with the authority and the information to change it. The appeal *shall* reach someone empowered to act, not a queue that returns the same automated answer; and the reviewer *shall* have access to the case-level record needed to evaluate the challenge on its merits. A determination that cannot be explained cannot be justly appealed, and a determination that cannot be appealed *should* not be fully automated.

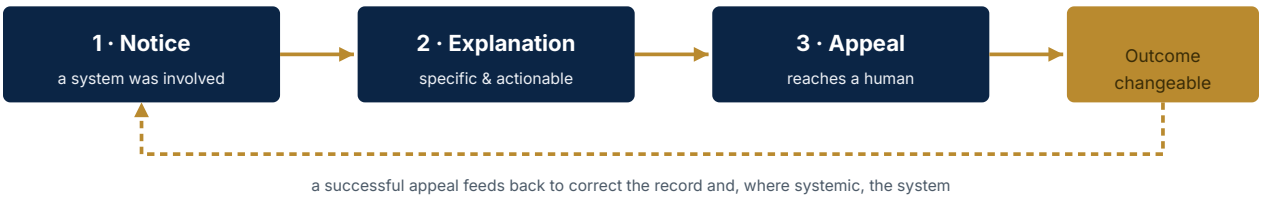

Figure 2. The individual-disclosure chain. Each link is necessary: notice without explanation cannot be acted on, explanation without appeal cannot be redressed, and appeal that reaches no empowered human is theater.

Proportionality of individual disclosure

The depth of these obligations scales with consequence, consistent with the conformance framework. A low-stakes recommendation may warrant only lightweight notice. A decision that materially affects a person's rights, benefits, safety, or livelihood *shall* carry the full set—notice, meaningful and faithful explanation, and a real appeal—at the Enhanced or Comprehensive level. The standard's insistence is not that every automated interaction be wrapped in due process, but that the interactions that can seriously harm a person never be stripped of it.

SECTION VI

Tradeoffs: Privacy, Security, Gaming & IP

Transparency is a good, but it is not the only good. It can conflict with privacy, security, gaming resistance, and legitimate intellectual property. A credible standard does not deny these tensions; it requires that they be resolved deliberately, documented, and never used as a reflexive excuse for opacity.

The weakest transparency programs treat disclosure as costless and are surprised when it collides with other obligations; the most cynical invoke privacy or trade secrecy to withhold whatever they prefer not to reveal. This section charts a course between them. Each tension below is real, each has a principled resolution, and in every case the standard requires the same discipline: name the conflict, choose a mitigation, and record what disclosure was foregone and why—so that the tradeoff can itself be reviewed.

Transparency versus privacy

Disclosing how a system works, and especially disclosing the data it learned from, can expose information about the individuals in that data. Detailed model or decision explanations can, in some settings, leak facts about training examples. This tension is genuine, and privacy is a protected interest, not an excuse. The resolution is that transparency about *the system* rarely requires exposing *individuals*: an organization can disclose data provenance, composition, and known biases at the population level, and can explain decisions in terms of factors, without publishing the personal records of the people in the training set. Where a specific disclosure would genuinely compromise privacy, the organization *shall* record the conflict and adopt the least-restrictive mitigation that preserves the interest, rather than withholding the whole class of information.

Transparency versus security and gaming

Some systems exist precisely because their subjects would defeat them if they understood them—fraud detection, security screening, abuse prevention. Full disclosure of the features that drive a fraud model is, in effect, a manual for committing undetected fraud. This is the hardest tension in the standard, because here the affected individual's interest in explanation and the public's interest in a working system point in opposite directions.

RESOLVING THE GAMING TENSION

The resolution is *differential disclosure by audience*, not blanket secrecy. An independent auditor or oversight body *shall* receive the full account needed to verify the system is fair and lawful, under confidentiality. The affected individual *shall* receive an explanation meaningful enough to understand and contest their case, even where the exact decision surface is withheld. What the standard forbids is using "gaming risk" to deny *all* transparency to *all* audiences, including those who could not game the system and to whom accountability is owed.

The distinction that makes this workable is between information that enables evasion and information that enables accountability. The specific numeric thresholds an adversary would exploit are often the former; the fact that a system is in use, its general purpose, its performance across groups, and its governance are the latter. A conforming organization *shall* distinguish the two and *shall not* let a legitimate need to protect the first justify concealing the second.

Transparency versus intellectual property

Vendors have real commercial interests in their models, and trade-secret protection is legitimate. But intellectual property is the interest most often over-claimed to resist scrutiny, and a public body or regulated firm cannot discharge its accountability by pointing at a supplier's confidentiality clause. The standard's position is that IP protects the *commercial* value of a system—its precise weights, its proprietary techniques—not the *public's* right to know that a consequential system is fair, lawful, and performing as claimed. These are reconciled, again, through differential disclosure: an auditor under confidentiality can inspect what a competitor may not see. Procurement *shall* secure the audit and disclosure rights the system's tier requires *before* purchase, because after deployment the leverage to obtain them is gone.

> *Trade secrecy protects a firm's advantage over its competitors. It was never meant to protect a system's advantage over the people it judges.*
>
> — On the limits of the IP shield

MANAGING THE TRANSPARENCY TRADEOFFS

Tension	Legitimate concern	Required resolution under this standard
Privacy	Disclosure may expose individuals in the data	Disclose at population level; explain by factors; record any withholding
Security / gaming	Explanation may enable evasion	Differential disclosure: full to auditors under confidentiality, meaningful to individuals
Intellectual property	Model details have commercial value	Protect weights & techniques; guarantee audit & accountability access by contract
Cost / burden	Documentation and explanation take effort	Scale to conformance level; reuse a single maintained source across audiences

The unifying discipline across all four tensions is documentation of the tradeoff itself. Whenever an organization withholds information it would otherwise disclose, it *shall* record what was withheld, from which audience, for which protected interest, and what alternative disclosure was provided instead. This record converts opacity from an unexamined default into a reviewable decision—which is the most transparency the hard cases will bear, and more than the reflexive invocation of privacy, security, or secrecy ever offers.

SECTION VII

Sector-Specific Adaptation

A single transparency requirement applied uniformly across every domain would be simultaneously too heavy for some uses and too light for others. The obligations owed by a system that recommends films are not those owed by a system that triages emergency care, and pretending otherwise discredits the standard. This section explains how the general requirements flex by sector and risk, while preserving a common spine so that "conformant" retains a stable meaning across domains.

Adaptation operates on two dials. The first is the **conformance level**—Baseline, Enhanced, or Comprehensive—chosen according to the consequence of the system's decisions. The second is **sectoral emphasis**: which of the four transparency layers and which audiences matter most in a given domain. A lending model and a clinical model may both operate at the Comprehensive level, yet emphasize different things—the lending model foregrounds decision-layer explanation and appeal for the applicant; the clinical model foregrounds model-layer performance evidence for the clinician who must decide whether to trust it.

How the layers re-weight by domain

ILLUSTRATIVE SECTORAL EMPHASIS ACROSS THE TRANSPARENCY LAYERS

Sector	Dominant concern	Layers emphasized
Consumer finance	Fair, contestable decisions about individuals	Decision & model; individual explanation and appeal
Healthcare	Clinician trust and patient safety	Model & data; performance by condition and group
Employment	Non-discrimination across protected groups	Data & model; disaggregated performance, provenance
Public benefits	Due process and the right to be heard	System & decision; notice, explanation, appeal
Content & recommendation	Aggregate influence and scope disclosure	System; existence, purpose, governance
Security & fraud	Accountability without enabling evasion	System & model to auditors; differential disclosure

The table is illustrative, not exhaustive, and the assignments are defaults to be reasoned from rather than rules to be applied mechanically. Its purpose is to show that sector adaptation is principled: it

follows from who is affected, how badly, and what those affected need in order to be protected. An organization *shall* be able to justify its sectoral emphasis by reference to the harms its domain makes possible, not by reference to convenience.

Regulated sectors and existing obligations

In several domains, transparency-adjacent duties already exist in law—adverse-action explanation in lending, non-discrimination duties in employment and housing, due-process expectations in public administration. Where such duties exist, this standard is designed to *operationalize* rather than displace them: its documentation artifacts and disclosure requirements give an organization concrete means to satisfy the transparency the law already contemplates. Where domain law is silent, the standard supplies a defensible default. In neither case does conformance to this standard substitute for legal advice, and an organization *shall* treat the more demanding of the two—law or standard—as controlling.

ADAPTATION RULE

Sector-specific adaptation adjusts the *emphasis and depth* of transparency requirements; it *shall not* be used to eliminate a layer entirely for a consequential system. A high-stakes system may weight the model layer over the data layer, but it may not go dark on either. The common spine—documentation across all relevant layers, explanation calibrated to audience, disclosure to the affected—holds across every domain.

Foundation models and the supply chain

Modern deployments increasingly build on general-purpose foundation models supplied by third parties and adapted downstream. This layering complicates transparency: the deployer often cannot see the base model's training data, and the supplier often cannot see the deployer's use. The standard addresses this by allocating obligations along the chain. The **supplier** of a foundation model *shall* provide model-layer and, to the extent feasible, data-layer documentation adequate for downstream risk assessment, and *shall* disclose known limitations and out-of-scope uses. The **deployer** *shall* document the system layer and the decision layer for its specific use, and *shall not* treat the supplier's documentation as discharging obligations that belong to the deployment. Neither party may point at the other to explain a gap the affected individual experiences.

This allocation matters because responsibility, like transparency, must have an owner at every layer. A foundation-model supply chain that leaves the data layer undocumented by the supplier and unexaminable by the deployer produces exactly the accountability vacuum this standard exists to prevent. Procurement is again the lever: the deployer *shall* obtain, by contract, the supplier documentation its conformance level requires, before the system is placed in consequential use.

SECTION VIII

The Conformance Framework

A standard is only as credible as its conformance framework. This section defines the three conformance levels, the evidence each requires, and how conformance is claimed and assessed—so that a claim of conformance is verifiable rather than rhetorical.

The framework rests on three levels of increasing rigor. An organization selects a level for each system according to the consequence and context of its decisions, using the same proportionality logic that runs through this standard: the more a system can affect a person's rights, safety, or livelihood, the higher the level it *shall* meet. A conformance claim always names a level and the layers it covers; a bare assertion of "transparency" is not a conformance claim.

The three conformance levels

1 **Baseline.**
For lower-consequence systems. The organization maintains current system, model, and data documentation at the Baseline field set, provides notice where individuals are affected, and can produce an account of the system on request. Baseline ensures a system is never wholly undocumented or unannounced.

2 **Enhanced.**
For systems that inform consequential decisions with a human materially in control. Adds disaggregated performance, described data provenance and labeling, faithfulness-characterized explanations, and meaningful individual explanation with an appeal route. Enhanced is the working level for most systems that touch people's interests.

3 **Comprehensive.**
For systems that determine or heavily steer decisions about rights, benefits, safety, liberty, or livelihood. Adds reproducible evaluation evidence, independent audit access, provenance chains, and full notice-explanation-appeal with faithful, actionable explanations. Comprehensive is the level at which the public's strongest claims to transparency are honored.

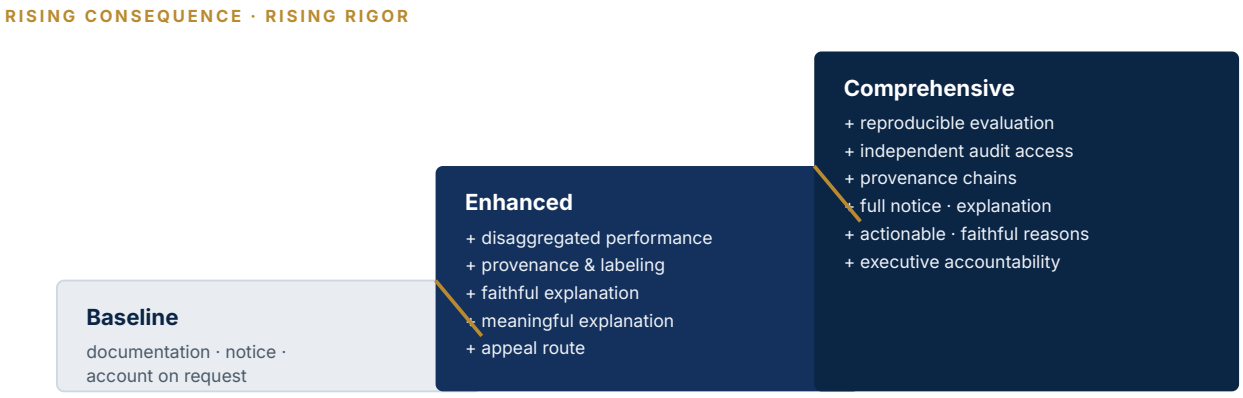

Figure 3. The conformance ladder. Each level subsumes the one below and adds obligations proportionate to the higher stakes it governs. An organization climbs the ladder as the consequence of a system's decisions rises.

Evidence and assessment

Conformance is demonstrated by evidence, not assertion. For each requirement at a chosen level, the organization *shall* be able to produce a corresponding artifact or record. Assessment may be internal (first-party), performed by a customer or partner (second-party), or performed by an independent body (third-party); the standard *recommends* that Comprehensive-level claims be supported by independent assessment, because self-graded conformance at the highest stakes carries the least credibility exactly where credibility matters most.

CONFORMANCE EVIDENCE BY LEVEL

Requirement area	Baseline	Enhanced	Comprehensive
Documentation artifacts	Present & current	Full field set, reviewed	Reproducible, audit-ready
Explanation faithfulness	Method stated	Characterized	Characterized & independently checked
Individual disclosure	Notice	Notice + explanation + appeal	Full chain, actionable & faithful
Tradeoff records	—	Withholding logged	Logged & reviewed by oversight
Assessment	First-party	First- or second-party	Independent recommended

Claiming conformance honestly

A conformance claim under this standard *shall* state three things: the level claimed, the layers it covers, and the assessment basis. A claim *shall not* imply broader coverage than the evidence supports; claiming Comprehensive conformance at the model layer while the data layer sits at Baseline is permitted only if the claim says exactly that. Overbroad or unqualified conformance

claims are themselves a transparency failure—an organization misrepresenting how transparent it is has, at the meta level, been opaque about its own opacity.

CLOSING PRINCIPLE

The purpose of this standard is not disclosure for its own sake. It is to ensure that the people affected by artificial intelligence, and the institutions accountable for it, can understand, question, and contest what these systems do. Transparency is the precondition for every other value we ask of AI— fairness, safety, accountability—because none of them can be verified in the dark. A conforming organization does not merely publish; it makes itself answerable.

Adoption is expected to be incremental. An organization *may* begin by bringing its highest-consequence systems to Enhanced conformance and its remaining systems to Baseline, then climb toward Comprehensive where the stakes warrant. What the standard asks throughout is consistent: that transparency be treated as a produced, maintained, and assessable property of a system—and that where it is limited, the limits be disclosed as plainly as everything else.

APPENDIX

Notes & Further Reading

This standard synthesizes established practice in AI documentation, explainability, and accountability, together with widely adopted public-sector guidance. The sources below are entry points into the literature that informs it; they are indicative rather than exhaustive.

1. Mitchell, M., et al. *Model Cards for Model Reporting.* On structured documentation of model performance across conditions and demographic groups, the basis for the model card in Section III.
2. Gebru, T., et al. *Datasheets for Datasets.* On documenting dataset motivation, composition, collection, and recommended uses—the basis for the datasheet.
3. Defense Advanced Research Projects Agency. *Explainable AI (XAI) Program.* On methods for producing explanations of machine-learning systems and the tension between performance and interpretability.
4. National Institute of Standards and Technology. *AI Risk Management Framework (AI RMF 1.0).* A voluntary framework whose govern, map, measure, and manage functions this standard operationalizes for transparency.
5. National Institute of Standards and Technology. *Four Principles of Explainable Artificial Intelligence (NISTIR 8312).* On explanation, meaningfulness, accuracy, and the knowledge-limits principle.
6. Organisation for Economic Co-operation and Development. *OECD AI Principles.* Intergovernmental principles including transparency and explainability for trustworthy AI.
7. Rudin, C. *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.* On preferring inherently interpretable models where the stakes are highest.
8. Ribeiro, M. T., Singh, S., & Guestrin, C. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier.* On local, post-hoc explanation and its interpretive risks.
9. Wachter, S., Mittelstadt, B., & Russell, C. *Counterfactual Explanations Without Opening the Black Box.* On actionable, contestable explanations for affected individuals.
10. Selbst, A. D., & Barocas, S. *The Intuitive Appeal of Explainable Machines.* On the limits of explanation and the risk of plausible but unfaithful accounts.
11. Raji, I. D., et al. *Closing the AI Accountability Gap: Internal Algorithmic Auditing.* On documentation and audit processes across the development lifecycle.
12. Diakopoulos, N. *Accountability in Algorithmic Decision Making.* On disclosure, transparency, and the governance of consequential automated systems.