INSIGHTS | FRAMEWORKS

The Reform AI Evaluation Framework

Version 1

By Dr. Ben E. Magee, Founder | October 4, 2026

This is version one of the framework Reform Intelligence® uses to evaluate whether an enterprise's use of artificial intelligence can be trusted. We are publishing it because a standard that cannot be examined is not a standard. It will change as the field, the regulations, and our own engagements teach us more. We invite scrutiny.

The unit of evaluation is the deployment

A model is not trustworthy or untrustworthy in the abstract. What can be trusted, or not, is a specific system doing a specific job: this model, wired into this business process, trained or prompted on this data, acting with this much autonomy, under these controls, touching these decisions.

The same model can be sound in one deployment and dangerous in another. A language model that drafts internal summaries is a different risk than the same model approving credit or triaging patients. For that reason, the unit Reform Intelligence evaluates is the deployment, not the model. Everything below measures the deployment as a whole.

What we measure

The framework assesses a deployment across six dimensions. Each asks a question the deploying institution must be able to answer with evidence, not assertion.

1. Fitness for purpose. Does the system do what it claims to do, for this specific use, at the level of performance the decision requires? We measure against the real task, not a generic benchmark the builder chose.

2. Failure characterization. Where and how does the system fail? Every capable system has a boundary of competence. We map the edge cases, the failure modes, and the conditions under which the system should not be trusted, so the institution knows what it is relying on and what it is not.

3. Reliability and drift. Is the system stable over time and across the range of inputs it will actually see? Models change as data and conditions change. We assess sensitivity, robustness, and whether monitoring is in place to catch degradation before it causes harm.

4. Data and provenance. What data does the system depend on, where did it come from, and can it be trusted? We examine quality, representativeness, lineage, privacy exposure, and the risk of leakage or contamination.

5. Autonomy and control. How much can the system do on its own, and what stops it when it is wrong? We assess the degree of autonomy, the quality of human oversight, reversibility, containment, and the controls that bound an agentic system's actions.

6. Governance and accountability. Who owns this system, how is it documented, and can its behavior be audited? We assess ownership, documentation, auditability, incident response, and alignment to the frameworks institutions are measured against, including the NIST AI Risk Management Framework, ISO/IEC 42001, and the EU AI Act.

How an evaluation runs

An engagement moves through five stages, each producing evidence that can be examined later.

Scoping. We define the boundary of the deployment: the model, the process, the data, the decisions it touches, the stakeholders, and its risk tier. The scope is written down and agreed before any testing begins.

Evidence. We gather documentation, audit the data, interview the people who built and operate the system, and run structured testing and adversarial probing against the six dimensions.

Measurement. We apply documented, repeatable methods so that a result can be reproduced and checked. Where we cannot measure something, we say so, rather than implying a confidence we do not have.

Determination. We issue an evidence-based determination for each dimension and for the deployment overall: ready, ready with conditions, or not ready, with the specific controls required and the explicit limits of what the system should be trusted to do.

Monitoring. A determination is a point in time. Where a system is put into production, we establish ongoing drift surveillance so the determination stays honest as the system and its world change.

What you receive

Not a grade. A determination you can stand behind, backed by evidence you can inspect, challenge, and show to a board or a regulator. It states plainly what the system can be trusted to do, what it cannot, under what conditions, and what must be true for it to stay trustworthy. Because the evaluation is independent and the methods are documented, the determination carries weight that a self-assessment cannot.

Why this is version one

Evaluation of AI is a young discipline. The honest position is that no framework is finished, including ours. We will revise this as the regulations mature, as the systems change, and as our engagements teach us where the framework is right and where it is not.

We publish it openly for a simple reason. An institution that asks others to be transparent and accountable has to hold itself to the same standard first. If you work in AI governance, risk, or regulation and you see something we have missed, tell us. The standard improves when it is examined, which is the entire point.

Reform Intelligence AI, Inc. is an independent evaluation and governance company, organized as a Delaware public benefit corporation.
reformintelligence.com

© 2026 Reform Intelligence AI, Inc. All rights reserved. Reform Intelligence® is a trademark of Reform Intelligence AI, Inc. This material is provided for general informational purposes only and does not constitute professional, legal, or investment advice, or an offer of services. No part may be reproduced or distributed for commercial purposes without permission.

The Reform AI Evaluation Framework and the methodologies described are proprietary to Reform Intelligence AI, Inc. Publication does not grant a license to use them.