The Evidence Ladder: Make AI Prove Itself Before It Judges Us

Kostakis Bouzoukas *

Posted Date: 18 December 2025

doi: 10.20944/preprints202512.1615.v1

Keywords: artificial intelligence; algorithmic accountability; evaluation; governance; public policy

 doi:10.20944/preprints202512.1615.v1

Article

The Evidence Ladder: Make AI Prove Itself Before It Judges Us

Kostakis Bouzoukas

Breakthrough Pursuit; kostasbuzukas@gmail.com

Abstract

Artificial intelligence systems increasingly score, sort, and advise people in welfare, policing, education, and employment. Many of these systems are trusted on the basis of thin evidence such as benchmark scores, internal tests on historical data, or polished demonstrations rather than robust evaluation in the real world. This paper argues that such deployments invert the burden of proof, because people must show that they were harmed while vendors rarely have to show that their tools work fairly and reliably for the communities they affect. Drawing on documented cases in child welfare, online proctoring, and facial recognition, I propose a simple evidence ladder for AI that ranges from basic lab tests to independent audits, monitored field trials, and formal certification with ongoing review. The novelty is a cross domain, five level scaffold that any team can use to state its current proof and to plan concrete steps toward stronger evidence. I link these levels to familiar engineering practices and to current policy frameworks including the OECD AI Principles, the NIST AI Risk Management Framework, and the European Union AI Act. The central claim is that demanding evidence scaled to the stakes of the decision is a basic form of respect for the people whose lives AI systems judge.:

Keywords: artificial intelligence; algorithmic accountability; evaluation; governance; public policy

---

Executive Summary

1. Problem. High impact AI often relies on lab benchmarks and internal validations that do not reflect lived conditions, which shifts risk onto the public.
2. Contribution. A practical evidence ladder with five levels clarifies how strong the proof is for any given AI use and aligns required evidence with real world stakes.
3. Engineering link. The ladder extends verification and validation, independent audits, canary style pilots, and continuous monitoring to the social effects of AI.
4. Counterargument. The ladder does not stifle innovation. It builds trust, reduces backlash, and de risks deployment.
5. Policy fit. The approach complements the OECD AI Principles, NIST AI RMF, the White House Blueprint for an AI Bill of Rights, the EU AI Act, and ISO 23894.
6. Bottom line. If an AI system will judge people, it should earn that authority by climbing the evidence ladder first.

1. Introduction

Robert Williams was at home with his family in Detroit when police officers arrived and arrested him on his front lawn. The primary trigger was an automated facial recognition match. There were no witnesses who knew him and no physical evidence that tied him to the scene. He was detained and later released when investigators admitted that the system had made a mistake. His case became a widely reported example of what happens when an AI output is treated as fact [1]. It became a signal moment for policy and engineering because a score achieved in the lab translated into real harm in the world.

 doi:10.20944/preprints202512.1615.v1

Not all harms look like this. Many are quiet and repeated. Consider Miriam, a composite character based on public reports from welfare systems. She is a single parent whose benefits are reviewed by an algorithm that generates a risk score. When that score crosses a threshold, a letter arrives that tells her payments will be reduced or stopped. She is now asked to prove that she is not defrauding anyone because a model she never met has labeled her as risky.

The common structure is clear. An AI system is treated as authoritative without ever having to prove, in the world, that it deserves that authority. Recent work has explored where AI should not be used and how to pilot generative tools in classrooms, and has examined how a small number of actors control much of the global AI talent and infrastructure [6][8][9]. This paper proposes the Evidence Ladder, a cross domain, five level scaffold that states how strong an AI system's proof really is, from lab tests to certified and monitored deployments. It links each level to familiar engineering practice and to current policy frameworks. The claim is proportional. High stakes uses require higher rungs. The aim is a shared language of proof that builders, buyers, regulators, and the public can use together.

2. Where Current Trust Comes From

Today many AI deployments lean on three thin forms of proof. First, benchmark scores on standard datasets. A model is declared to be highly accurate on a well known collection of examples. Second, small gains on leaderboards. One algorithm beats another by a fraction of a percentage point in a shared test. Third, polished demonstrations that show the system at its best.

These signals can help teams compare designs and detect obvious defects. They do not answer the questions that matter to the people who live with the system after launch. Benchmarks say little about how a model behaves for groups that are under represented in the training data. Leaderboards do not say what happens when a system runs for months in changing environments. Demos avoid awkward edge cases.

The gap between test performance and lived impact is visible across domains. In child welfare, a screening tool that relied on administrative data correlated poverty with risk and sent more low income families into intrusive investigations [10]. In policing, face analysis products that passed vendor tests have misidentified Black men at higher rates [2][19]. In remote education, proctoring systems have failed to detect darker skin tones and have struggled with unstable networks, pushing some students into impossible trade offs between unreliable exams and accusations of cheating [3].

In each case someone could point to testing of some kind. Numbers existed. The missing piece was a shared sense of what level of proof should be required before the system was allowed to judge real people.

3. The Evidence Ladder

A plain way to address this is to treat proof as a ladder. Each step represents a stronger answer to a basic question. How do you know this system works, and for whom. Level 5 will usually be domain specific. Medical tools, financial decision systems, and safety critical applications already have pathways for formal approval, while other domains may rely on standards bodies or public procurement rules.

Table 1. The Evidence Ladder for AI Systems.

Level	What it means	What it does not tell us
1. Lab tests and anecdotes	Tested by creators on curated or historical data. Success stories exist.	Whether it works for real people in varied conditions.

 doi:10.20944/preprints202512.1615.v1

2. Simulations and historical validation	Held out tests or synthetic runs including retrospective checks on past cases.	Long term behavior, adaptation by users, and side effects.
3. Independent audit or peer review	External experts evaluate performance, fairness, robustness, and claims.	How it behaves inside real institutions.
4. Field trial with monitoring	Limited deployment with monitoring, transparency, and a real option to roll back.	Full generalization across regions, time, and communities.
5. Certification and ongoing review	Assessment by a regulator or standards body and continuous monitoring after approval.	Not applicable. Ongoing review is part of the obligation.

Alt text. Table 1 lists five levels of evidence from lab tests to certification and ongoing review, and states what each level confirms and what it does not.

Related Work

The Evidence Ladder aligns with established hierarchies of evidence in medicine and program evaluation, and it extends verification and validation in software engineering to the social effects of AI. It complements documentation and accountability work such as model cards for model reporting and datasheets for datasets [13][14]. It also fits proposals for independent audits of algorithmic systems and is consistent with critiques of abstraction that caution against over reliance on lab performance when systems enter institutions [15][16]. Texts on fairness in machine learning summarize the pitfalls of evaluation without external validity [17]. Work on large language models has warned that scale and fluency can hide new failure modes that benchmarks do not reveal [18]. The ladder addresses the deployment gap between lab performance and in the wild behavior by foregrounding external validity and long term observation, and it complements ISO and NIST risk management by making evidence levels legible to non specialists [4][21]. Prior efforts like model cards and datasheets document how models and datasets are built and where they work. Audit scholarship defines what independent evaluation should include and why it matters for accountability. Work on sociotechnical fairness warns that lab performance can hide harms once systems enter institutions. The Evidence Ladder adds a shared, cross domain way for builders, buyers, and communities to state where a system stands on proof today and what steps would raise that proof tomorrow.

The novelty lies in adding an explicit coverage test that asks whose lives are in the evidence and in mapping each level to socio technical practices rather than only to model metrics.

4. Does the Ladder Stifle Innovation

A common concern is that higher proof will slow or stop progress. If AI tools must pass audits, pilots, and regulatory reviews, will society lose out on useful systems. Three points help place this worry. First, other fields already use staged evidence where risk is high. New medicines do not go straight from bench to pharmacy. New aircraft do not go straight from wind tunnel to commercial service. These processes take effort, yet we accept them because we have seen the cost of skipping them. Second, the ladder is proportional. Not every system needs Level 5. The level of proof should match the stakes. A recommendation widget may remain at Level 1 or 2. A tool that screens job applicants or prioritizes cases for investigation should climb higher before it shapes outcomes. Third, evidence protects innovation. Systems rushed into the world on thin proof can spark scandals and

 doi:10.20944/preprints202512.1615.v1

backlash that make the public sceptical of AI more broadly. By contrast, systems that can point to independent audits, monitored pilots, and regulatory clearance are more likely to gain the trust of professionals and the public. In practice, staged rollouts with clear reversal criteria sustain adoption better than launches that trigger public controversy.

Practical constraints and risks

Cost for small teams. Evidence work can burden small developers. Shared templates, pooled test suites, and open challenge sets can reduce cost and raise consistency.

Goodhart effects. Audits and metrics can be gamed. Surprise audits, rotating challenge sets, and evaluation by independent parties reduce this risk.

Data access and privacy. Independent evaluation may require sensitive data. Secure enclaves, differential privacy, and synthetic testbeds can enable third party checks without exposing personal information.

Distribution shift. Passing one audit is not enough. Drift monitoring, scheduled reassessment, and clear triggers for retraining or rollback belong at Level 5.

5. Whose Lives Become the Evidence

Evidence also has a coverage problem. When we say that a system has been tested, whose lives were used to generate that proof. Pilots often begin in well resourced settings. A large hospital hosts a clinical trial of a diagnostic tool. A high ranked university trials a teaching assistant. A global platform tests a content ranking change on a small share of its base. Deployment then extends to very different places. Public health services with staff shortages. Community colleges and local agencies. Regions with unstable networks. People encounter the system with different languages, work patterns, and constraints, and with fewer ways to resist or complain.

Computed grading systems in the United Kingdom during the pandemic provide a clear example. Algorithms that used school level historical performance to adjust teacher predictions ended up cutting grades for many students from less advantaged schools. Public anger forced a reversal after real opportunities were lost [20]. The lesson is simple. Evidence gathered in privileged settings can be used to justify systems that fall hardest on communities with the fewest buffers.

Cross border deployment magnifies this problem. Evidence gathered in the Global North is often used to justify systems in the Global South without local pilots or community review.

The ladder is incomplete unless we ask about coverage at every stage. Who was included. Who was missing. Who decided that this was acceptable.

6. Engineering Practices that Implement the Ladder

Teams can implement the ladder within normal engineering practice.

At Levels 1 and 2, treat benchmark and offline results as necessary but not sufficient. Build challenge sets that focus on known failure modes such as under represented dialects, low bandwidth environments, and historically disadvantaged groups. Document what your tests do not cover.

At Level 3, invite external auditors or academic partners. Provide controlled access to models and data under privacy respecting agreements. Publish high level results so claims can be scrutinized. Independent audits give Level 3 its backbone by bringing an outside view of performance, fairness, and robustness into the engineering loop [15].

At Level 4, design pilots with clear gates. Specify who is included. Define metrics that matter in context. Set thresholds that trigger rollback. Log predictions, outcomes, human overrides, and user feedback. Treat the pilot as a socio technical test, not only a software test.

At Level 5, extend site reliability practices to model behavior. Track drift and error patterns across groups. Classify incidents where model outputs cause harm and treat them as first class operational events. Participate in sector specific standards for documentation and recertification.

Checklist A. What counts as Level 3 and Level 4 evidence

 doi:10.20944/preprints202512.1615.v1

For Level 3. Independent audit or peer review

1. Scope stated in writing. Which decisions, which populations, which failure modes.
2. Fresh data held out from developers.
3. Grouped results. Accuracy, error types, and calibration by relevant groups.
4. Harm review. Document known harms and near misses from similar systems.
5. Public summary. Methods and results explained in plain language.

For Level 4. Field trial with monitoring

1. Clear gate. Who is included, duration, opt out and rollback rules.
2. Metrics. Outcomes that matter to people, not only throughput.
3. Logging. Predictions, overrides, appeals, and complaints.
4. Listening loop. Routine meetings with affected staff and communities.
5. Stop rule. Criteria to pause or remove the system if harm appears.

Table 2. Stakes to evidence map.

Decision stakes	Examples	Minimum ladder level
Low stakes and reversible	Content or product recommendations that users can ignore	Level 1 or Level 2
Medium stakes and reversible	Customer service triage or study aids with easy human override	Level 2 or Level 3
High stakes and partly reversible	Hiring screening, loan pre screening, clinic triage with clear override paths	Level 3 and monitored Level 4 pilot
High stakes and hard to reverse	Welfare eligibility, grading, policing alerts, credit limits	Level 4 pilot followed by Level 5 with ongoing review

Alt text. Table 2 maps decision stakes to a minimum evidence level and gives examples for each cell.

Worked example. Applying the ladder to a hiring screen

Decision and stakes. An employer wants to use an AI screen to shortlist applicants for entry level roles. The decision affects access to work and income, and errors can compound historical disadvantage. The stakes are high and partly reversible through human review.

Level 3 audit plan. Scope. Screening for role fit across three job families. Populations. Early career candidates from community colleges, four year universities, and non traditional backgrounds. Failure modes. Penalizing non standard career paths, penalizing caregiving gaps, and discounting resume formats from non elite schools. Data. Fresh holdout data that developers have not seen. Metrics. Selection rate by group, false positive and false negative rates by group, positive predictive value by group, area under the ROC curve, precision, recall, and calibration error. Thresholds. If subgroup selection rates fall below agreed ranges, or if error asymmetries exceed agreed limits for two consecutive reporting periods, the tool fails the audit. Reporting. A public summary in plain language that explains scope, methods, and findings.

Level 4 pilot plan. Scope. Two business units for six months, with explicit opt out for hiring managers and candidates. Monitoring. Weekly dashboards for subgroup selection rates, manager overrides, candidate complaints, and time to fill. Listening loop. Bi weekly meetings with hiring managers and a monthly forum with candidates and community college partners. Rollback rule. If subgroup selection rates fall below thresholds for two consecutive weeks, pause the tool in the affected job family and investigate.

 doi:10.20944/preprints202512.1615.v1

Decision on Level 5. If the tool meets thresholds over the pilot, shows stability across hiring cycles, and demonstrates that managers can and do override when needed, seek certification or attestation under an accepted audit standard. Maintain ongoing review and publish a yearly plain language summary of performance and incidents.

Box C. Common anti patterns

Announcing accuracy without subgroup performance.
Scaling from a pilot to production without opt out or rollback.
Treating vendor tests as independent audits.
Logging model outputs but not human overrides, appeals, or complaints.
No defined stop rule for harm.

7. Policy landscape

Policy and standards are moving in the same direction. The OECD AI Principles and the NIST AI Risk Management Framework call for trustworthy and human centered AI and encourage systematic testing, documentation, and monitoring [5][4]. The White House Blueprint for an AI Bill of Rights begins with the principle of safe and effective systems and recommends pre deployment testing, risk mitigation, and ongoing monitoring [7]. The European Union AI Act, enacted in 2024, introduces a risk based approach. It imposes stricter obligations on high risk systems in areas such as employment, education, law enforcement, and critical infrastructure, including requirements for quality data, human oversight, logging, and post market monitoring [11]. At the city level, New York City Local Law 144 is a concrete example in employment. Employers may not use automated employment decision tools unless those tools undergo an annual bias audit by an independent reviewer and notices are provided to candidates. Summaries of audits must be made public [12]. ISO 23894 provides complementary guidance on AI risk management that can be mapped to the ladder for operational practice [21].

8. Method Note and Limitations

This is a practical scaffold, not a domain protocol. Teams select outcome measures with affected communities. Secure evaluation setups protect sensitive data. The ladder describes levels of proof. It does not prescribe one metric for every sector.

The ladder is a simple tool. It cannot capture every nuance of context or harm. It requires complementary work on data quality, privacy, and human oversight. It assumes institutions have some capacity to run pilots and collect feedback. Even so, the ladder improves on the current default, which is to rely on thin lab evidence and polished demonstrations.

9. Conclusions

AI systems will always involve uncertainty. There will always be surprises and trade offs. That is not a reason to avoid them. It is a reason to be explicit about what we know before we hand them authority over people.

If an AI system will grade essays, rank applicants, flag families for investigation, score citizens as high or low risk, or shape what information people see, we should expect more than a lab score and a confident presentation. We should expect to know where on the evidence ladder it stands and who was included in that proof. For any AI that touches rights or life chances, make one simple rule. Publish the current ladder level and the plan to reach the next one.

To ask for that is not hostility to AI. It is a way to show respect for the people who live with these systems. Our lives are not edge cases. Our communities are not disposable test sets. If an AI system is going to judge us, it should prove itself first.

Show us the evidence.

 doi:10.20944/preprints202512.1615.v1

And show us the people behind it.

Appendix A. Printable ladder and stakes map

This page repeats the two tables so readers can print a single sheet for meetings and reviews.

Table A1. The Evidence Ladder for AI Systems.

Level	What it means	What it does not tell us
1. Lab tests and anecdotes	Tested by creators on curated or historical data. Success stories exist.	Whether it works for real people in varied conditions.
2. Simulations and historical validation	Held out tests or synthetic runs including retrospective checks on past cases.	Long term behavior, adaptation by users, and side effects.
3. Independent audit or peer review	External experts evaluate performance, fairness, robustness, and claims.	How it behaves inside real institutions.
4. Field trial with monitoring	Limited deployment with monitoring, transparency, and a real option to roll back.	Full generalization across regions, time, and communities.
5. Certification and ongoing review	Assessment by a regulator or standards body and continuous monitoring after approval.	Not applicable. Ongoing review is part of the obligation.

Alt text. Table A1 repeats the five ladder levels for easy printing.

Table A2. Stakes to evidence map.

Decision stakes	Examples	Minimum ladder level
Low stakes and reversible	Content or product recommendations that users can ignore	Level 1 or Level 2
Medium stakes and reversible	Customer service triage or study aids with easy human override	Level 2 or Level 3
High stakes and partly reversible	Hiring screening, loan pre screening, clinic triage with clear override paths	Level 3 and monitored Level 4 pilot
High stakes and hard to reverse	Welfare eligibility, grading, policing alerts, credit limits	Level 4 pilot followed by Level 5 with ongoing review

Alt text. Table A2 repeats the stakes to evidence mapping for meetings.

References

1. Bhuiyan, J. 2023. First man wrongfully arrested because of facial recognition testifies as California weighs new bills. The Guardian. April 27, 2023.

 doi:10.20944/preprints202512.1615.v1

2. Buolamwini, J., and Gebru, T. 2018. Gender Shades. Intersectional accuracy disparities in commercial gender classification. Proceedings of the Conference on Fairness, Accountability, and Transparency.
3. Coghlan, S., Miller, R., and Paterson, J. 2021. Good proctor or big brother. Ethics of online exam proctoring. Philosophy and Technology 34, 1581 to 1605.
4. National Institute of Standards and Technology. 2023. Artificial Intelligence Risk Management Framework. AI RMF 1.0.
5. Organisation for Economic Co operation and Development. 2019. OECD Principles on Artificial Intelligence.
6. Schmitt, K. 2024. When not to use AI. IEEE Technology and Society Magazine 43, 2, 7 to 15.
7. White House Office of Science and Technology Policy. 2022. Blueprint for an AI Bill of Rights. Safe and effective systems.
8. Wu, Y., and Lin, C. 2024. The hidden multiplier. Unraveling the true cost of the global AI skills gap. IEEE Technology and Society Magazine 43, 4, 15 to 23.
9. Orchard, A., Behrens, J. T., and Dhaliwal, R. S. 2025. Two Years In the Wild. A multidisciplinary approach to teaching generative AI. IEEE Technology and Society Magazine 44, 3, 23 to 29.
10. Eubanks, V. 2018. A child abuse prediction model fails poor families. Wired. January 2018.
11. European Union. 2024. Artificial Intelligence Act.
12. New York City Department of Consumer and Worker Protection. 2023. Automated Employment Decision Tools. Local Law 144 of 2021.
13. Mitchell, M., Wu, S., Zaldivar, A., et al. 2019. Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency.
14. Gebru, T., Morgenstern, J., Vecchione, B., et al. 2021. Datasheets for datasets. Communications of the ACM 64, 12, 86 to 92.
15. Raji, I. D., Smart, A., White, R., et al. 2020. Closing the AI accountability gap. Defining auditing and audit studies for algorithmic systems. Proceedings of the Conference on Fairness, Accountability, and Transparency.
16. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. 2019. Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency.
17. Barocas, S., Hardt, M., and Narayanan, A. 2019. Fairness and machine learning. Online book.
18. Bender, E. M., Gebru, T., McMillan Major, A., and Shmitchell, S. 2021. On the dangers of stochastic parrots. Proceedings of the Conference on Fairness, Accountability, and Transparency.
19. Grother, P., Ngan, M., and Hanaoka, K. 2019. Face recognition vendor test. Part 3. Demographic effects in facial recognition algorithms. NIST Interagency Report 8280.
20. Kelly, A. 2021. A Tale of Two Algorithms. The Appeal and Repeal of Calculated Grades Systems in England and Ireland in 2020. British Educational Research Journal.
21. ISO/IEC 23894. 2023. Information technology. Artificial intelligence. Risk management.