Evidence Before Action: When AI Systems Should Decide, Ask, or Abstain

Andrew Koski

6 November 2026 | Working paper · Version 1.0 | Draft prepared 6 November 2026

Abstract

An AI system can produce a plausible answer before it has enough evidence to justify the action attached to that answer. I argue that acceptance should depend on the specific claim, the evidence supporting it, and the consequence of being wrong. This working paper proposes an evidence-to-action contract that separates candidate generation, claim validation, and action authorization. The framework applies to visual recognition, document processing, equipment identification, and information retrieval. It builds on selective prediction and confidence calibration without claiming to invent abstention. Its proposed contribution is an operational evaluation framework for claim-level evidence requirements, targeted requests for additional information, and proportionate escalation. A controlled evaluation is specified, including domain shift, incomplete inputs, and false acceptance costs. No new cross-domain experiment is reported.

Keywords: selective prediction; evidence provenance; abstention; decision systems; uncertainty

Explore the paper

1. The position
2. Three separate contracts
3. A decision rule with explicit costs
4. A reusable workflow
5. Proposed evaluation
6. Objections and limitations
7. Conclusion
References

1. The position

A system that always answers is easy to demonstrate. A system that knows what its answer actually establishes is harder to build. I argue that this distinction should sit at the center of practical AI design. The relevant question is not simply whether an answer looks correct. It is whether the available evidence supports the claim strongly enough for the next action.

Consider a photograph of an industrial component. Its appearance may support a product-family match while failing to establish its voltage rating, revision, or compatibility with a particular installation. A document may resemble an invoice while providing no evidence that its payment instructions are authorized. Recognizing a pattern and authorizing an action are different decisions. These examples are illustrative, not reported deployments or measured outcomes.

Selective classification already provides a formal setting in which a predictor can reject some inputs rather than return a classification for every case. Geifman and El-Yaniv study the trade-off between coverage and risk for deep neural networks. [1] The position developed here extends that practical question to the interface between a claim, its supporting evidence, and the action that a product permits.

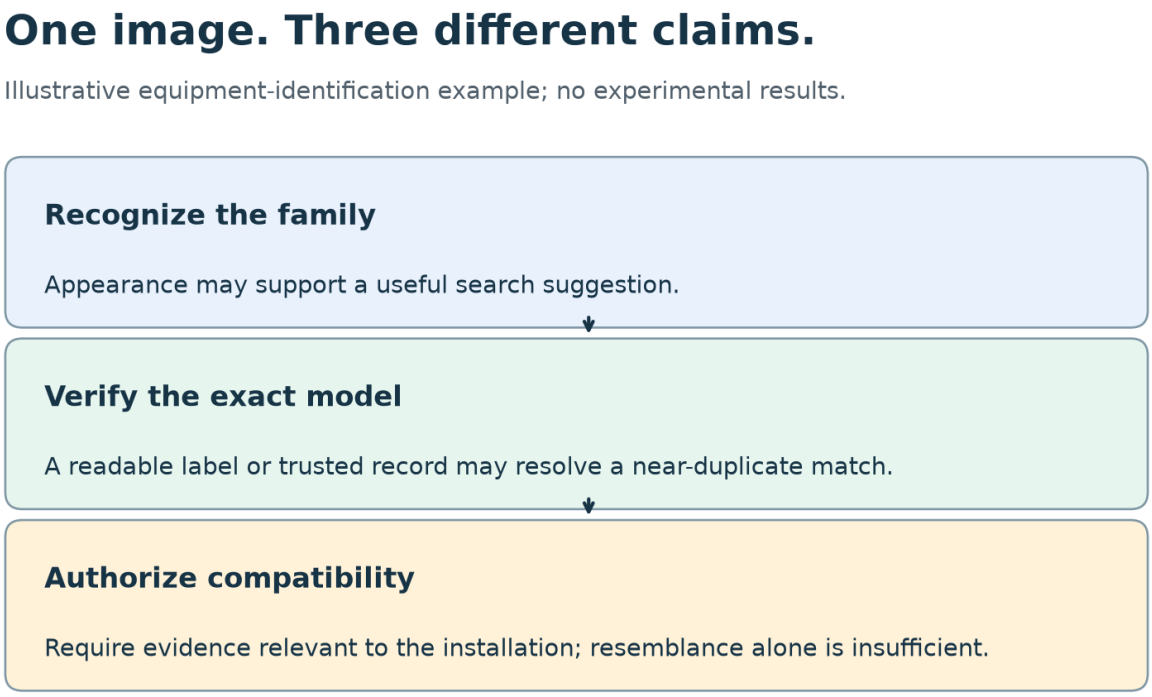


Figure 1. Illustrative claims require different evidence. A useful family match does not authorize compatibility.

2. Three separate contracts

The recognition contract describes what the system is trying to establish. It should identify a target claim at a fixed level of specificity: a document type, a product family, a revision, or an individual object. Changing that level after seeing a wrong answer makes evaluation unreliable. Returning a broad category must not silently count as success on an exact-identity task.

The evidence contract defines what observations can support that claim. A visible model number, an authenticated record, a second view, or a trusted measurement may have different roles. Repeated model answers based on the same photograph are not independent observations. A confidence score also needs an interpretation. Research on neural-network calibration shows why apparent confidence cannot automatically be treated as the frequency of correctness. [2]

The action contract defines what follows from acceptance. Displaying a tentative suggestion, updating a catalog record, and triggering an irreversible operation have different consequences. The framework therefore permits a candidate to remain useful while withholding authority to act. It does not assume that a human reviewer is infallible or that every uncertainty requires review.

3. A decision rule with explicit costs

Let c denote a candidate claim, e its evidence, and a a proposed action. Let q(c,e) estimate the probability that c is correct in the relevant operating population. In a simplified binary loss model, acceptance has expected error loss $(1 - q) \times L_{wrong}$. Deferral has cost L_{defer}. Acceptance is preferable when its expected loss is lower, subject to mandatory evidence requirements. This is an illustrative decision rule, not a universal probability guarantee.

Mandatory requirements matter because some evidence gaps cannot be repaired by increasing a model score. If the action requires authorization, visual resemblance is the wrong kind of evidence. If the requested property cannot be observed from the input, confidence in a neighboring property is insufficient. A policy should state these constraints before deployment and record the reason for every deferral.

An additional-information step should request evidence likely to distinguish the leading alternatives. Asking for a second photograph of the same uninformative surface adds effort without resolving uncertainty. A useful request might instead seek a serial label, a different angle, or an authoritative record. When the expected value of further information is lower than its cost, the system should stop requesting it and provide a clear unresolved outcome.

4. A reusable workflow

Figure 1 separates candidate generation from permission to act. The system first defines the requested claim and available evidence. It generates candidates, tests required evidence, and evaluates the remaining uncertainty against the action policy. It may accept, ask for targeted information, route to review, or abstain. Each result retains provenance and the policy version used to reach it.

The same structure can serve several domains without assuming that their thresholds transfer. In document extraction, a field can remain unresolved even when the document class is known. In equipment identification, a family match can support search while exact compatibility remains blocked. In catalog recognition, shared artwork may support a group of products while failing to identify the printing. The reusable part is the contract; the evidence and costs remain domain-specific.



Figure 2. Proposed evidence-to-action workflow. Additional evidence returns to validation; acceptance is specific to the claim and action.

5. Proposed evaluation

The proposed study compares three conditions: a forced-answer predictor, a confidence-threshold predictor, and the evidence-to-action workflow. All conditions should use the same underlying model

and candidate pool where possible. A separate ablation should remove targeted evidence requests to test whether that feature contributes beyond abstention alone.

Construct held-out sets for at least three domains, including incomplete inputs, unfamiliar devices or document layouts, and near-duplicate candidates. Split by underlying entity and acquisition source so that alternate photographs of the same object do not leak between calibration and testing. Freeze claim definitions, policy costs, and thresholds before scoring. Label both identity correctness and whether the supplied evidence actually establishes the requested property.

Report coverage and error among accepted cases together, with uncertainty intervals and domain-specific breakdowns. Also measure false acceptance at the action boundary, unnecessary deferrals, request completion rate, review burden, latency, and cost. Comparing systems at matched coverage and matched action loss prevents an always-abstain system from appearing useful merely because it makes few mistakes.

The proposal would lose support if the evidence workflow fails to improve accepted-case error or action loss over a calibrated threshold baseline at comparable coverage and cost. A gain confined to familiar images would not establish cross-domain value. Where labels are disputed, report adjudication procedures and disagreement rather than silently selecting favorable answers.

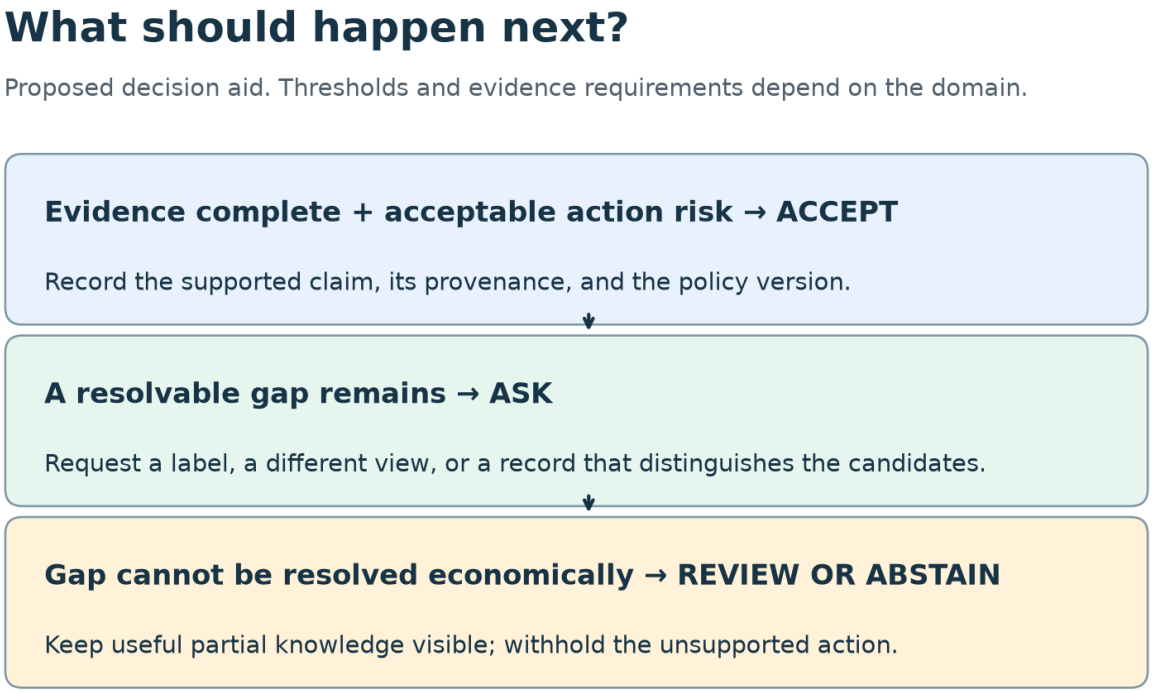


Figure 3. Proposed outcomes: accept, ask for discriminating evidence, review, or abstain.

6. Objections and limitations

Evidence requirements can become brittle rules that reject legitimate cases. They can also encode a designer's assumptions about which observations matter. The evaluation must therefore include valid but unusual inputs, and policies need explicit revision procedures. A rigid checklist that blocks unfamiliar evidence without improving outcomes is not a successful implementation.

There is also a risk of moving uncertainty into the reviewer or the evidence source. Human review needs measured error rates; an authoritative-looking record can be outdated or compromised. Calibration may deteriorate after deployment. None of these problems is solved by attaching an evidence label to a prediction. The framework supplies places to inspect them, not a guarantee that they disappear.

7. Conclusion

I expect reliable AI products to make the boundary between suggestion and action increasingly explicit. The value of abstention is not the refusal itself. It is the ability to preserve useful partial knowledge while avoiding a claim that the evidence cannot support. This paper proposes a way to evaluate that boundary across domains. The next step is a reproducible study, not a broader accuracy claim.

References

[1] Geifman, Y., and El-Yaniv, R. (2017). Selective Classification for Deep Neural Networks. Advances in Neural Information Processing Systems 30. https://arxiv.org/abs/1705.08500

[2] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 1321–1330. https://proceedings.mlr.press/v70/guo17a.html