
| RESEARCH ARTICLE

Human-in-the-Loop AI for Exposure Detection in Loan Portfolio Monitoring

Ankur Mehra

Independent Researcher, New Jersey, United States

Corresponding Author: Ankur Mehra, Independent Researcher, **E-mail:** reachankurm@gmail.com

| ABSTRACT

Automated detection also allows different aspects of what a bank perceives about its loans to change. Alerts are provided continuously about concentration build-ups, early indicators of credit deterioration, weakening collateral and covenant positions, and anomalies in the data sent by servicers of loan tape data. Despite this, the next binding constraint to be solved is whether an alert is received by a person when it is generated. Research in Human-Computer Interaction found that reviewers accept automated feedback, ignore it when it makes an error, and accept it when it confirms their prior beliefs. Explanatory output does not eliminate biases and may even reinforce them. As such, having a human reviewer at the end of the automated pipeline ends up being oversight in name only. The remainder of the article argues the human reviewer should be folded into triage, not be assigned at the end. Combining the literature on detection, human-AI decision studies, financial alert reviews and supervisory expectations leads to: a taxonomy of loan portfolio exposure signals and human roles; a seven-stage human-in-the-loop reference architecture with an analyst reaching an independent conclusion before seeing the model evidence; an alert decision contract document describing all the information the alert needs to convey to the human decision-maker; and a set of human-oversight governance metrics implemented in line with the model risk guidance, the NIST AI Risk Management Framework and the human-oversight obligations in the EU AI Act. An example using a commercial real estate concentration cycle illustrates the principles. From this, it appears that the quality of monitoring depends at least as much on how alerts are reviewed as on how they are detected.

| KEYWORDS

human-in-the-loop, anomaly detection, monitoring the loan portfolio, automation bias, explainable AI, model risk governance

| ARTICLE INFORMATION

ACCEPTED: 4 September 2026

PUBLISHED: 10 October 2026

DOI: 10.32996/jcsts.2026.8.10.8

1. Introduction

Exposure in a loan portfolio typically occurs below each of these thresholds, as a sector approaches the limit on concentrations a few basis points at a time, as the ages of appraisals in the portfolio necessarily increase, and the collateral values decrease. The covenant is no longer tested against borrower behavior and the servicer tape shows the same occupancy figure quarter after quarter. Manual review (often periodic) struggles with such slow build-up, while automated detection can find it more easily. In large portfolios, ML models outperform benchmarks in predicting loan and corporate default (Barbaglia et al., 2023; Moscatelli et al., 2020), and unsupervised methods can find anomalies that might be unexpected to a rule-based model (e.g., Pang et al., 2021).

The real problem for governance comes after something is detected: every detected exposure is converted into a notification, competing for a limited number of human reviewers. The natural governance response is to keep a human in the loop, such as an analyst who reviews and signs off on model outputs. The result appears reassuring, but evidence shows that reviewing decisions with decision aids often results in omission and commission errors when the decision aid itself is flawed. (Parasuraman & Manzey, 2010). Models may simply be abandoned upon discovery of issues (Dietvorst et al., 2015). Moreover, many approaches to oversight effectively leave the human in a role where they have little more to do than approve the output (Green, 2022).

Supervisors have also recognized this risk: the EU Artificial Intelligence Act requires personnel overseeing automated systems with high risk must understand the temptation to over-rely on automated output (European Parliament & Council of the European Union, 2024). Nonetheless, the need for effective challenge of model outputs was described over a decade earlier in model risk guidance issued by American regulators (Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, 2011).

We argue the reviewer and the alert form the proper unit of design. We define a taxonomy of exposure signals with corresponding human roles, a human-in-the-loop reference architecture, an alert decision contract, and governance metrics ensuring the oversight quality of human reviewers. Future work may build on our contributions.

2. Literature Review

2.1 Detection capability and its limits

In general, anomaly detection has consistent terms, e.g. Chandola et al. (2009) defined point, contextual and collective anomalies, and these translated naturally into loan monitoring. A point anomaly is one loan with an improbable value. A contextual anomaly is a sector exposure that is normal nationally, but excessive in a single area. A cluster of simultaneous covenant waivers is collective. Pang et al. (2021) organize deep anomaly detection into 3 high-level categories and 11 fine-grained categories, including weakly supervised methods that deal with the situation where there are only a few labeled cases to learn from, which is the case for analyst dispositions.

Alternatively, ML can track deteriorating credit. Barbaglia et al. (2023) predict the default of 12 million residential mortgages, concluding ML consistently outperforms logistic benchmarks. This is also what Moscatelli et al. (2020) find with 300,000 Italian non-financial firms in 2011-2017, wherein the advantage vanishes when controlling for behavioral credit data. This qualification is useful in the sense that more data may matter as much as a better model. However, Chen et al. (2024) point out that post-hoc explanations such as LIME and SHAP may also become less reliable under class imbalance. The infrequent occurrence of credit deterioration means that, where they are most needed, explanations for an exposure alert are least reliable.

2.2 How people use automated advice

Observational studies of human-AI decision making question the assumption that the reviewer is an effective protective layer. Parasuraman and Manzey (2010) build on existing research and combine the two phenomena into one, positing that overloaded reviewers reallocate attentional resources away from monitoring the aid. Indeed, Dietvorst et al. (2015) document a reversed failure: if people find that the algorithm fails, they prefer a human forecaster to an algorithm, even an inferior one.

While these explanations were expected to resolve this tension, Bansal et al. (2021) experimentally showed that, whether the recommendation was correct or incorrect, they increased the likelihood that subjects would adopt it. Bućinca et al. (2021) counter that cognitive forcing functions are AI designs that force a person to think before they can see the AI's answer. In an experiment with 199 participants, cognitive forcing functions reduced over-reliance, compared with simple explainable-AI conditions. Vasconcelos et al. (2023) further qualify the effect of explanations: across five studies with 731 participants, explanations only reduced over-reliance when they made verifying the AI's answer easier.

Two further studies clarify: Alon-Barkat and Busuioc (2023) find no evidence of blanket automation bias with 605 participants in their first study; however, reviewers follow advice that corresponds to their own beliefs, indicating selectivity in following algorithmic recommendations. Dratsch et al. (2023) show that expertise offers only so much protection, as both inexperienced and very experienced radiologists misclassify cases when presented with erroneous recommendations.

2.3 Alert tone in finance applications

As financial institutions already have operational experience mitigating the problem through AML transaction monitoring, an interview study by Oztas et al. (2024) with eight AML experts found that false positives and providing explainable outputs were two persistent challenges in AML supervision. Automatic suppression is one such option, but Bakry et al. (2024) show that a suppression model that scores 86% F-beta would still misclassify 11% of money-laundering customers in test data. Jesus et al. (2021) tested these claims in real-world fraud investigations and found that explainers did not help fraud analysts to make better decisions. Instead, showing only the data gave analysts the highest accuracy.

2.4 Governance expectations

Governance frameworks tend to address oversight but rarely provide specifications for how oversight should be designed. In SR 11-7 model risk is referred to in terms of improper use and misuse of outputs; appropriately skilled challenge is required (Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, 2011). The NIST AI Risk Management Framework outlines the functions of govern, map, measure and manage (National Institute of Standards and Technology, 2023). Model risk and data governance are outlined as vulnerabilities in finance by the Financial Stability Board (2024). According to Green (2022), many of the 41 oversight policies surveyed have unrealistic assumptions about human reviewer performance,

although there is some design advice. Amershi et al. (2019) offer 18 design principles for human-AI collaboration. Mosqueira-Rey et al. (2023) view human-in-the-loop ML as a continuum of strategies from active learning to machine teaching.

The literatures rarely intersect. Detection rarely asks about reading its alerts. Human-AI studies rarely use credit data. And supervisorial papers rarely specify how to achieve the desired outcome.

3. Methodology

Therefore, the study takes a three-step design-oriented synthesis approach: First, failure modes from the human review of automated advice and associated evidence-based design countermeasures are synthesized in Section 2. Second, the countermeasures are integrated into a reference architecture and alert specification on the basis of the author's experience in building analytics platforms for loan portfolio quality, exposure and risk forecasting, and spearheading AI- and ML-assisted workflow automation in a regulated banking setting. Thirdly, the design is exercised against a stylized example, a commercial real estate concentration alert, to ensure that every step has a defined input, a defined human action and an auditable output. This design was adjusted, however, because the counts of the model score and exposure information were shown to the participants from the start. Eventually, it was decided that separating the two be only after weighing the forcing-function and over-reliance evidence against the time the extra review would take.

However, two caveats should be kept in mind with these findings: first, most experiments on reviewer behavior have been on radiology, public administration, and financial-crime review, not credit monitoring, so extrapolating from those experiments to loan portfolios is tenuous. No proprietary alert or disposition data were available; however, those whose logs were accessible could provide this data if needed.

4. Results

4.1 A taxonomy of exposure signals

Table 1 organizes the types of signals that a loan-portfolio monitoring system can generate into four families. The distinguishing feature of these families is not so much how the signal is detected but what a human reviewer can add to the signal, for concentration signals that is a judgment of materiality and strategic context. For data-integrity anomalies, it is the ability to tell a data problem from a real change in exposure.

Table 1. Exposure-signal families in loan portfolio monitoring

Signal family	Typical detection approach	Characteristic failure	Human reviewer's distinctive role
Concentration build-up (sector, region, obligor)	Aggregation with trend and limit-trajectory thresholds	Slow drift that never breaches a single limit	Judge materiality against strategy and appetite
Credit deterioration (early warning)	Supervised ML default classifiers	False positives in stable segments; unstable explanations on rare events	Corroborate with obligor knowledge
Collateral and covenant weakness	Rules on covenant tests, valuation age, loan-to-value refresh	Stale inputs make weak positions look healthy	Request refresh; assess waivers
Data-integrity anomalies in loan tapes	Unsupervised methods (isolation forest, autoencoders)	Benign changes, such as servicing transfers, flagged as anomalies	Separate data error from genuine exposure change

Source: Author's synthesis drawing on Chandola et al. (2009), Pang et al. (2021), Barbaglia et al. (2023), Moscatelli et al. (2020) and Chen et al. (2024).

4.2 How AI-assisted review fails

The six failure modes and their countermeasures are shown in Table 2. Two comments are made. The failure modes pull in opposite directions (both over-trust and under-trust can occur), and some countermeasures are about changing the order of the workflow, making a universal best countermeasure unlikely.

Table 2. Failure modes in AI-assisted exposure review and design countermeasures

Failure mode	Evidence	Design countermeasure
Automation bias (omission and commission errors)	Parasuraman & Manzey (2010); Dratsch et al. (2023)	Independent first look before model evidence for high-severity alerts
Selective adherence	Alon-Barkat & Busuioc (2023)	Log every disposition; audit agreement patterns by reviewer; double-review a sample
Algorithm aversion	Dietvorst et al. (2015)	Show the model's track record for the alert type, not just its score
Explanation-induced over-trust	Bansal et al. (2021); Jesus et al. (2021)	Present verifiable evidence first; test explanation formats with analysts
Alert fatigue	Oztas et al. (2024); Bakry et al. (2024)	Severity tiering; suppression only with sampled human audit
Oversight as ratification	Green (2022)	Give reviewers authority to challenge thresholds and escalate, with that authority recorded

Figure 1 shows why the first countermeasure matters. In the Dratsch et al. (2023) experiment, readers' accuracy with correct AI suggestions stayed high, but it collapsed when suggestions were wrong: mean accuracy for inexperienced readers fell from 79.7% to 19.8%, and for very experienced readers from 82.3% to 45.5%. The domain differs from credit, but the task structure is similar: an expert confirming or rejecting a flagged case under time pressure.

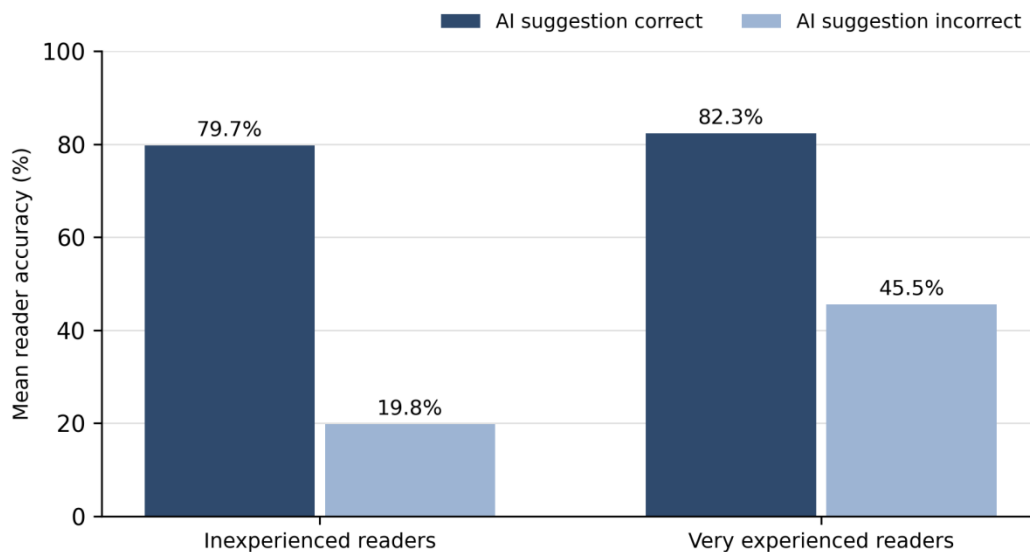


Figure 1. Mean reader accuracy with correct versus incorrect AI suggestions, by reader experience. *Source: data from Dratsch et al. (2023); evidence from radiology, shown for its analogous task structure.*

4.3 Seven-stage reference architecture model

The architecture is outlined in Figure 2.

1. Detection. Rules, statistical monitors and ML models run every portfolio refresh.
2. Alert assembly. The raw detections are deduplicated, resolved at obligor and portfolio segment level, assigned a severity score, and packaged per the decision contract (Section 4.4).
3. Independent first look. For high-severity alerts, the analyst reviews just the exposure facts and records their initially selected judgment, before getting the model score or explanation. This provides a cognitive forcing step (Bućinca et al., 2021). Lower tiers skip this step, being evaluated in a batch, so that the cost is only incurred where it might make a difference.
4. Evidence reveal. Model score, calibrated confidence, principal drivers, comparable past alerts and data quality flags are shown. Even in cases where evidence and feature attributions compete, evidence attribution is generally preferred (Jesus et al., 2021; Vasconcelos et al., 2023).
5. Disposition: The analyst escalates the alert, monitors it, closes it as a false positive, or reclassifies it as a data issue, using a set of pre-defined and finite reason codes. A brief rationale must be given for any departure from this view.

6. Feedback. Dispositions become labels. So-called active-learning queues route the most uncertain cases for review (Mosqueira-Rey et al., 2023), and the retraining process follows model validation rather than skipping over it.

7. Governance monitoring. We continuously compute metrics about the interaction between models and reviewers (Section 5.1). Because severity tiering accounts for much of the economic value of the design, careful consideration must be given to its rules. Tier assignment must depend only on the size and direction of the exposure, not on the model's confidence, else the model itself decides which of its outputs are to be scrutinized. In rough terms, a spike goes straight to the high end if it breaches a materiality limit, the limit is flagged for possible breach in the planning horizon, or a data integrity flag touches a sufficiently large obligor. Otherwise, it goes into batch review. In all cases, even a random sample is fed through the full first-look chain, so that none of these low tiers are ever unobserved.

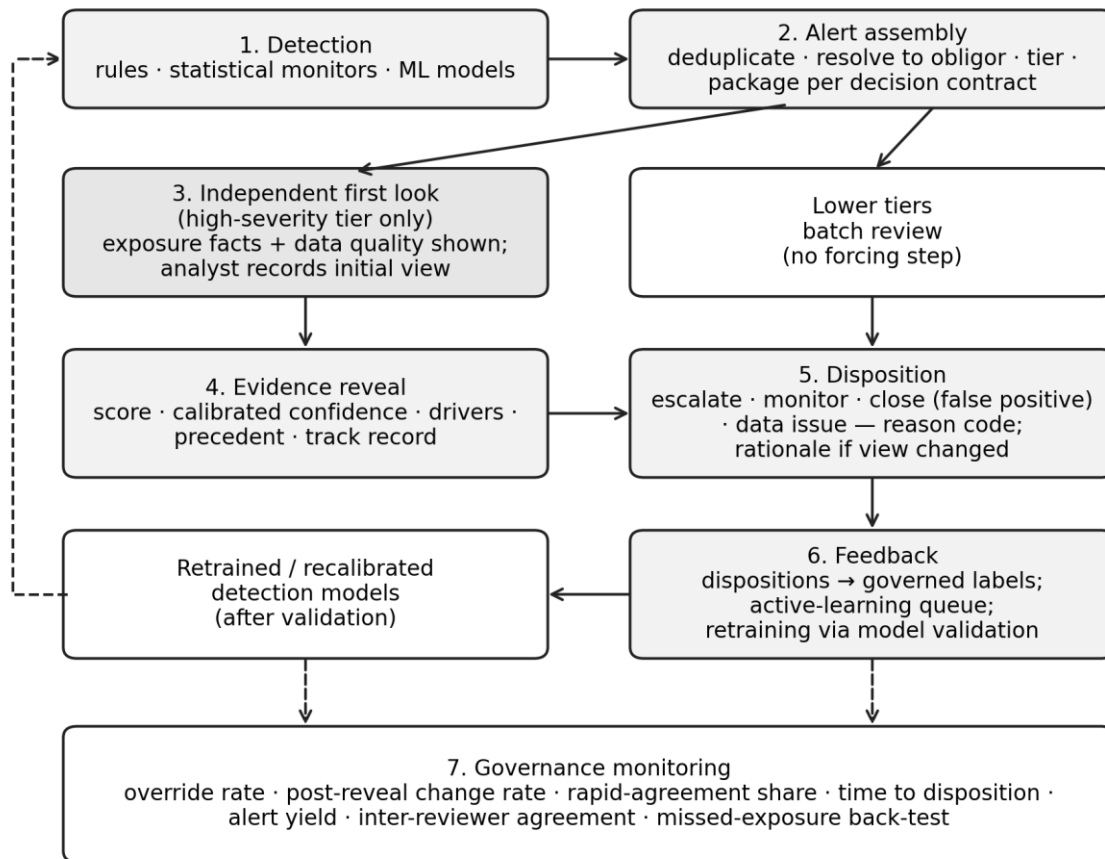


Figure 2. Seven-stage human-in-the-loop architecture for exposure detection in loan portfolio monitoring. *Source: Author's design.*

4.4 The alert decision contract

An alert is only as good as what the alert says. Table 3 defines the minimum contract. The one absolute is data quality: input freshness and known data issues are never suppressed, since the most pernicious alert in a reviewer's inbox is a confident score based on stale collateral values. Two other fields address the opposite failure: precedent and track record give the analyst some reason to believe or disbelieve the model, regardless of the current case. This could simply be because an analyst who sees that the alert type is generally accurate is less likely to dismiss an alert after one visible miss (Dietvorst et al., 2015), which is consistent with the theory of algorithm dislike.

Table 3. The alert decision contract: minimum content of every exposure alert

Field	Content	When shown
Exposure facts	What, where, how much, and the trend	Always first
Trigger	Rule or threshold crossed and why now	Always first
Data quality	Input freshness, missing fields, known issues	Always; never suppressed
Model evidence	Score, calibrated confidence, principal drivers	After the first look (high tier)
Precedent	Similar past alerts and their dispositions	With the model evidence
Track record	Historical precision of this alert type	With the model evidence

Dispositions	Allowed outcomes and reason codes	At decision
Accountability	Owner, service-level target, escalation path	Always
Lineage	Model version, data snapshot, rule set	Recorded; available on request

4.5 A worked example

A nightly run may find that office exposure in one metropolitan area is on track to breach an internal limit in two quarters' time. The unsupervised monitor may flag the same area for anomaly detection and alerts, if the occupancy field for several large loans has been unchanged in multiple servicer report cycles. This alert is high severity. The analyst will see the balances, limit trajectory and the collateral valuation dates, and returns an immediate "monitor" view since exposure is close to but not exceeded (model evidence follows). The main drivers are stale appraisals on large loans, and the data-quality field connects the occupancy anomaly back to the same obligors. When the analyst changes the disposition to "escalate", reason code "stale collateral and suspect servicer data", and requests new appraisals, we receive back the confirmed concentration alert and the confirmed data-integrity issue. The governance layer now knows that revealing the alert led to a different decision. That's exactly the sort of event you want to count.

5. Discussion

5.1 Making oversight measurable

An observable indication of oversight quality is needed if its role is more than a signature. Table 4 proposes indicators of review and model interactions, each mapped to one of the frameworks of interest for comparative evaluation. However, a high rate of rapid agreement with the model is not evidence that the model is right, but rather that it is a rubber stamp.

Table 4. Governance metrics for human-in-the-loop exposure review

Metric	What it reveals	Framework alignment
Override rate (both directions)	Whether reviewers exercise independent judgment	SR 11-7 effective challenge; EU AI Act Art. 14
Post-reveal change rate	How often model evidence changes the initial view, and in which direction	NIST AI RMF Measure; EU AI Act Art. 14
Rapid-agreement share	Dispositions made very quickly and in agreement with the model (a rubber-stamp signal)	NIST AI RMF Manage
Time to disposition by tier	Whether review capacity matches alert load	NIST AI RMF Manage
Alert yield	Share of alerts escalated, by signal family	NIST AI RMF Measure
Inter-reviewer agreement	Label quality on a double-reviewed sample	SR 11-7 data and outcomes analysis
Missed-exposure back-test	Later downgrades or defaults that were closed or never alerted	SR 11-7 outcomes analysis; NIST AI RMF Govern

5.2 Objections

Two objections to the argument above are direct responses: in principle, could better detection models render human review unnecessary? Although accuracy will improve, the current regulatory environment and the rare-event nature of credit deterioration mean that the most important cases are more likely to be those that are least stable. We cannot transfer responsibility to a model under current use norms around accountability for lending decisions (Chen et al., 2024). Cost is the second objection: because structured review costs analyst time and generates latency, independent first look is only conducted for alerts of higher severity. The cost of not investigating an alert that causes a material exposure is greater than the time the design requires of analysts, in most portfolios. Suppressing without audit is not a cheaper alternative; it simply moves the error somewhere less visible (Bakry et al., 2024).

5.3 Limits

Findings have limitations due to reliance on studies in adjacent disciplines: labels also reflect bias of reviewers in retraining (selective adherence is used as data unless audited). And metrics in Table 4 need to be calibrated against actual alert log histories; "rapid" agreement will vary by signal family and institution.

A few final points about the label problem: it connects the human and machine halves of the design. If analysts' dispositions mirror their prior beliefs (Alon-Barkat & Busuioc, 2023), a model retrained on their dispositions can learn not only their expertise but their blind spots over time. The inter-reviewer agreement metric and the missed-exposure back-test exist in part to detect this drift, but no one metric is sufficient to catch this drift on its own. Such institutions may opt in, but would need to revisit how labels are weighed when there is a real disposition history.

6. Conclusion

Automated detection has solved the problem of seeing more exposure, but has not solved the problem of acting on it well. The evidence about how people use automated advice, a reviewer at the end of the pipeline tends to override or ignore the model and explanations don't change that. Here, we change the role of the human: an independent first look on material alerts, a contract that any alert has a data quality context, a feedback loop treating dispositions as governed labels, and measures to show metrics that make the quality of oversight itself visible. In practice, this is often to test whether an institution can point to someone, rather than a computer program or model, making the decision after the analyst closes an exposure alert.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

- Alon-Barkat, S., & Busuioc, M. (2023). Human–AI interactions in public sector decision making: "Automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory*, 33(1), 153–169. <https://doi.org/10.1093/jopart/muac007>
- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery. <https://doi.org/10.1145/3290605.3300233>
- Bakry, A. N., Alsharkawy, A. S., Farag, M. S., & Raslan, K. R. (2024). Automatic suppression of false positive alerts in anti-money laundering systems using machine learning. *The Journal of Supercomputing*, 80(5), 6264–6284. <https://doi.org/10.1007/s11227-023-05708-z>
- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. <https://doi.org/10.1145/3411764.3445717>
- Barbaglia, L., Manzan, S., & Tosetti, E. (2023). Forecasting loan default in Europe with machine learning. *Journal of Financial Econometrics*, 21(2), 569–596. <https://doi.org/10.1093/jfinc/nbab010>
- Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency. (2011). *SR 11-7: Guidance on model risk management*. <https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm>
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–21. <https://doi.org/10.1145/3449287>
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 1–58. <https://doi.org/10.1145/1541880.1541882>
- Chen, Y., Calabrese, R., & Martin-Barragan, B. (2024). Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*, 312(1), 357–372. <https://doi.org/10.1016/j.ejor.2023.06.036>
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126. <https://doi.org/10.1037/xge0000033>
- Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., Mähringer-Kunz, A., Püsken, M., Baessler, B., Sauer, S., Maintz, D., & Pinto dos Santos, D. (2023). Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology*, 307(4), Article e222176. <https://doi.org/10.1037/xge0000033>
- European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- Financial Stability Board. (2024). *The financial stability implications of artificial intelligence*. <https://www.fsb.org/2024/11/the-financial-stability-implications-of-artificial-intelligence/>
- Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45, Article 105681. <https://doi.org/10.1016/j.clsr.2022.105681>
- Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., & Gama, J. (2021). How can I choose an explainer? An application-grounded evaluation of post-hoc explanations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 805–815). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445941>

- Moscatelli, M., Parlapiano, F., Narizzano, S., & Viggiano, G. (2020). Corporate default forecasting with machine learning. *Expert Systems with Applications*, 161, Article 113567. <https://doi.org/10.1016/j.eswa.2020.113567>
- Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 56(4), 3005–3054. <https://doi.org/10.1007/s10462-022-10246-w>
- National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>
- Oztas, B., Cetinkaya, D., Adedoyin, F., Budka, M., Aksu, G., & Dogan, H. (2024). Transaction monitoring in anti-money laundering: A qualitative analysis and points of view from industry. *Future Generation Computer Systems*, 159, 161–171. <https://doi.org/10.1016/j.future.2024.05.027>
- Pang, G., Shen, C., Cao, L., & van den Hengel, A. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2), 1–38. <https://doi.org/10.1145/3439950>
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381–410. <https://doi.org/10.1177/0018720810376055>
- Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), 1–38. <https://doi.org/10.1145/3579605>