

More Alerts, Less Prevention? A Computational Study of Hazard Detection, Response Capacity and Threshold Choice in Occupational Health and Safety

Ashish Mulraj Ved¹, Ravindra Kumar²

¹ PhD Research Scholar, Faculty of Commerce and Management, Asian International University, Manipur, India

² Assistant Professor, Asian International University, Manipur, India

Abstract- AI-enabled hazard detection is usually judged by whether it finds risky conditions. Prevention requires another step: somebody must have the time, authority and competence to respond. This article evaluates how that second step alters the interpretation of predictive safety performance. It reports a factorial computational experiment derived from a doctoral study of AI in occupational health and safety (OHS). Across 54 simulated operating conditions, five alert thresholds and 200 replications per condition, a score-based detector classifies 5,000 opportunities per replication. The design varies hazardous-opportunity prevalence (1%, 5% and 15%), score separation (1, 2 and 3), drift (0 and 0.75) and nominal response probability (0.40, 0.65 and 0.90), with response capacity limited to 500 alerts. At the reference threshold of 1.5, one 5%-prevalence, separation-2, no-drift condition yields mean residual hazardous opportunities of 180.8, 138.2 and 95.7 across the three response assumptions. With the same detector and nominal 65% response, imposing drift raises mean residual opportunities from 138.2 to 185.5. A sensitivity calculation also changes the preferred threshold when the relative weight of unresolved hazards and false alerts changes. The simulation shows why sensitivity, precision, workload and verified action belong in the same OHS decision model. All numerical data are computer-generated. The study does not report a deployed machine-learning system, workplace incidents, observed worker behaviour or injury reduction. Its contribution is a transparent stress test of the management choices that connect detection to prevention.

Keywords— Occupational health and safety; predictive safety; artificial intelligence; simulation; hazard detection; alert fatigue; response capacity; model drift; decision thresholds

I. INTRODUCTION: A WARNING CAN BE RIGHT AND STILL FAIL

Consider a simple possibility. A monitoring system notices many unsafe conditions, but the people who investigate alerts are already overloaded. More warnings arrive; fewer can be closed promptly. The detector may be working as designed, while the prevention process is not. This is not an account of an actual industrial incident. It is the management problem examined in this computational study.

OHS research increasingly describes uses of AI for PPE recognition, exposure monitoring, incident text analysis and predictive risk assessment (Fang et al., 2020; Shah & Mishra, 2024). These applications can increase the speed at which information becomes available. Yet an alert has no preventive effect merely because it appears on a screen. It must be received,

understood, prioritised and converted into a suitable control before exposure occurs.

Existing reviews have cautioned against equating innovation with injury prevention. A mapping review of digital OHS interventions in construction reported limited evidence of better working conditions or lower injury and illness outcomes (Trask & Linderöth, 2023). A systematic review focused on AI-based OHS tools also found that convincing direct outcome evidence remains scarce (Jetha et al., 2025). These findings make it useful to test the logic of a proposed system without pretending that a simulation is a workplace trial.

The present study asks what happens to residual hazardous opportunities when discrimination, alert threshold, prevalence, model drift, nominal response probability and finite response capacity interact.

Instead of beginning with the best possible detector score, the analysis begins with the decision pathway that must follow a detection.

II. RESEARCH QUESTIONS AND CONTRIBUTION

Four questions guide the experiment. First, how do thresholds change the balance between detecting hazardous opportunities and generating unnecessary alerts? Second, does improving the detector continue to improve expected resolution when the response team has a fixed workload limit? Third, how does loss of score separation through drift affect the number of unresolved hazardous opportunities? Fourth, does the preferred alert threshold change when management assigns different relative weights to missed prevention and false alerts?

The article contributes an explicitly specified hazard-to-action model, an interpretable factorial comparison and a consequence-weight sensitivity exercise. It does not propose a new trained machine-learning algorithm. Nor does it interpret simulated response assumptions as estimated training effects. Its novelty is in making the organisational response limit part of the same calculation as predictive detection.

III. DESIGN AND METHODS

1. Unit of analysis and experimental conditions

The unit of analysis is a simulated hazard opportunity, not an accident, worker or organisation. Each replication creates 5,000 opportunities. Hazard prevalence takes values of 0.01, 0.05 or 0.15. Score separation between hazardous and non-hazardous opportunities takes values 1, 2 or 3; drift is either 0 or 0.75; nominal response probability is 0.40, 0.65 or 0.90. These factors produce $3 \times 3 \times 2 \times 3 = 54$ operating conditions. All conditions are evaluated at thresholds 0.5, 1.0, 1.5, 2.0 and 2.5, using 200 Monte Carlo replications for each condition-threshold combination.

Within a replication, the hazardous count is drawn from a binomial distribution with 5,000 trials and the specified prevalence. Non-hazard scores follow a normal distribution with mean 0 and standard deviation 1. Hazard scores follow a normal distribution with mean equal to separation minus drift and standard deviation 1. An alert occurs when the score meets or exceeds the threshold. The generator controls discrimination by construction. No commercial AI system was trained, evaluated or deployed.

The same realised hazard statuses and scores are used across the five thresholds and related response assumptions. This common-random-number design makes the within-grid comparisons less noisy. The thesis specifies the pseudo-random generation sequence, including a fixed seed rule; this article reports results from that declared computational experiment rather than newly collected data (Morris et al., 2019). The design factors are summarised in Table 1.

2. Detection, capacity and expected resolution

True positives (TP) are hazardous opportunities that produce alerts; false positives (FP) are alerts on non-hazardous opportunities. False negatives (FN) are hazardous opportunities without alerts. Recall is $TP/(TP+FN)$; precision is $TP/(TP+FP)$. Neither measure, by itself, states whether the detected hazard was controlled.

Nominal response probability q is a hypothetical combined indicator of competence, authority, timeliness and response effectiveness; it is not a measured probability from a workforce study. A fixed capacity K of 500 alerts can receive full response effort in a replication. If the detector generates more than K alerts, effective response is reduced in proportion to K divided by the alert count. Expected resolved opportunities are TP times effective response, and residual opportunities are the hazardous-opportunity count minus expected resolutions. Since response is calculated as an expectation, these quantities can be fractional and do not represent incident counts.

The default illustrative burden assigns 20 units to each residual hazardous opportunity and one unit to each false alert: $B = 20 \times \text{residual} + \text{FP}$. These are arbitrary scenario weights, not monetary costs, medical severity estimates or observed social preferences. They permit a transparent threshold comparison under a declared assumption. Table 2 sets out the complete decision equations.

3. Repetitions, precision and interpretation

For each condition and threshold, reported figures are means over 200 replications. Monte Carlo standard errors indicate the numerical precision of the generated means; they do not measure uncertainty about industry conditions. No p value is interpreted as evidence that a real intervention has an effect. The simulation does not represent response queues, repeated alerts for the same exposure, risk-dependent dispatch priority or the independent distribution of harm severity. These choices make the first model easy to reproduce, but they also limit transferability.

Table 1. Declared simulation design (all parameters hypothetical).

Factor	Values / method
Hazard prevalence	1%, 5%, 15% of 5,000 opportunities
Score separation	1, 2, 3 normal-score mean units
Drift	0 or 0.75 downward shift in hazardous scores
Nominal response probability, q	0.40, 0.65 or 0.90
Detector thresholds	0.5, 1.0, 1.5, 2.0, 2.5
Response capacity, K	500 alerts per replication
Replications	200 per condition-threshold; 54 conditions $\times$ 5 thresholds = 270 summaries

Table 2. Response and burden definitions.

Measure	Definition / note
Alert count	$\text{TP} + \text{FP}$
Recall	$\text{TP} / (\text{TP} + \text{FN})$
Precision	$\text{TP} / (\text{TP} + \text{FP})$

Measure	Definition / note
Capacity factor	$\min(1, K / \max(\text{TP} + \text{FP}, 1))$
Effective response	$q \times \text{capacity factor}$
Expected resolved	$\text{TP} \times \text{effective response}$
Residual opportunities	$\text{TP} + \text{FN} - \text{expected resolved}$
Illustrative burden	$20 \times \text{residual opportunities} + \text{FP}$; arbitrary weight units

IV. RESULTS

1. Response capability changes the meaning of detection

The reference comparison fixes prevalence at 5%, score separation at 2, drift at zero and threshold at 1.5. Under nominal response assumptions of 40%, 65% and 90%, the simulated mean residual hazardous opportunities are 180.8, 138.2 and 95.7 per 5,000 opportunities, respectively (Table 3). The score generator and detector threshold do not change across these three entries. Only the assumed response capability changes. The difference is therefore a result of the model's response mechanism, not evidence from three actual worker groups.

The same distinction appears when discrimination improves. At 5% prevalence, no drift and 65% nominal response, mean residual opportunities move from 200.4 at separation 1, to 138.2 at separation 2, and 113.2 at separation 3. Better detection helps within the declared model, but the amount of prevention depends on the response that can follow.

2. Drift is an OHS management problem, not just a software metric

At the reference threshold and 65% nominal response, introducing drift of 0.75 into a separation-2, 5%-prevalence condition increases mean residual hazardous opportunities from 138.2 to 185.5. Under the simulation's construction, the background score distribution is held constant while hazardous scores move closer to it. More true hazards fall below the unchanged threshold.

In a workplace, drift might arise because equipment, lighting, procedures, reporting or behaviour changes. The numerical shift here is hypothetical and not an estimate of how much any of those changes would affect a real model. The management implication is nevertheless testable: after deployment, a model needs continuing checks against its original operating context and a process for controlled recalibration or withdrawal.

3. Prevalence and the false-alert problem

When hazardous opportunities are rare, even a detector with encouraging discrimination can create many non-hazardous alerts. That increases review work while doing nothing directly to remove a hazard. The thesis includes a separate 10,000-opportunity score illustration with 5% target prevalence and separation 2. Its ROC area is 0.922 and its average precision is 0.538. These values describe generated score discrimination before applying the capacity rule; they are not the result of validating a deployed model.

This distinction matters because a model's ROC area can look strong even when day-to-day alert handling is difficult. Precision, false-alert counts, available staff time and the consequences of missing a hazard have to be examined together. Overall accuracy is particularly unhelpful if it allows rare hazardous cases to be hidden by many normal observations.

4. Lower thresholds are not automatically safer

Lower thresholds generally increase recall by drawing more hazardous scores into the alerted set. They also classify more normal opportunities as possible hazards. If the resulting workload is above the capacity limit, the model reduces the chance that any alerted hazard is resolved. Raising the threshold does the opposite: the queue becomes more manageable, but some hazardous opportunities are missed.

For condition S26 (5% prevalence, separation 2, zero drift and nominal response 65%), the consequence-weighting exercise in Table 4 makes the trade-off visible. When residual hazards and false alerts are given equal weight, threshold 2.5 has the lowest burden within the tested grid. When a residual

hazard is assigned five times the weight of a false alert, threshold 2.0 is selected. With weights 10, 20, 50 or 100, threshold 1.5 is selected. The detector scores did not change; the decision preference did. These are choices within five tested thresholds, not an optimal operating threshold for any real facility.

5. Numerical precision is not external validity

The thesis reports median Monte Carlo standard errors of the mean residual of 0.429, 0.904 and 1.666 residual opportunities for the 1%, 5% and 15% prevalence groups, respectively (90 summaries in each group). The estimates are fairly precise within the declared simulation. They say nothing about the correctness of the assumed score distributions, response probabilities or workplace workload patterns. A precise answer to an assumed model remains an assumed-model answer.

Table 3. Selected means at threshold 1.5, each per 5,000 simulated opportunities (source: thesis Table 5.6).

Condition	Prevalence	Separation	Drift	Nominal response	Mean residual
S20	5%	1	0	65%	200.4
S25	5%	2	0	40%	180.8
S26	5%	2	0	65%	138.2
S27	5%	2	0	90%	95.7
S29	5%	2	0.75	65%	185.5
S32	5%	3	0	65%	113.2
S44	15%	2	0	65%	540.0

Table 4. Consequence-weight sensitivity for simulated condition S26 (source: thesis Table 5.7).

Residual weight	Best threshold in grid	Mean residual	Mean false alerts	Weighted burden
1	2.5	199.1	29.0	228.0
5	2.0	168.5	106.8	949.4
10	1.5	138.2	315.1	1,697.4
20	1.5	138.2	315.1	3,079.8
50	1.5	138.2	315.1	7,226.9
100	1.5	138.2	315.1	14,138.7

V. DISCUSSION: HOW BETTER SCORES CAN BECOME BETTER SAFETY DECISIONS

The study supports a narrower but more useful way of discussing predictive safety. Detector development asks whether a system identifies relevant patterns. OHS management must also ask what happens next. A false negative can leave a hazard invisible; a false positive can consume scarce attention; and an alert that is correct but unresolved does not remove exposure. Managing these competing consequences is part of safety governance rather than merely a machine-learning tuning exercise.

The response-capacity calculation gives practical meaning to familiar problems such as alert fatigue and weak escalation. A supervisor cannot investigate an unlimited number of alerts with the same care. If workload routinely exceeds capacity, an organisation could improve triage, change how alerts are grouped, strengthen engineering controls, assign more response resources or restrict the technology to better-defined use cases. Buying a more sensitive detector would not, by itself, establish that the system has become safer.

Competency deserves particular attention. The simulated q parameter combines several determinants—knowledge, judgement, authority, available time and control effectiveness. It cannot isolate training, and it certainly does not prove a training intervention raises response from 0.40 to 0.90. In real evaluation these determinants should be measured separately: whether the worker saw the alert, understood it, had authority to act, selected an appropriate control, and verified that the hazard was removed. These observations are more relevant to preventive performance than attendance at a training session alone.

Human oversight remains essential when automated outputs are ambiguous or wrong. Workers and supervisors should be able to challenge an alert, document the reason and continue following the established safe system of work. Such controls also

address the possibility of automation bias, whereby confidence in the technology displaces local judgement (Niehaus et al., 2022). Privacy, fairness and psychosocial consequences must be considered when data come from cameras, wearables or detailed monitoring of worker behaviour (El-Helaly, 2024).

The findings are compatible with a management-system approach. ISO 45001 offers responsibilities for operational control, performance evaluation and continual improvement; ISO/IEC 42001 and NIST AI RMF provide additional AI-governance thinking. A useful assurance report would present recall and precision alongside alert workload, response capacity, time to closure, worker feedback, model changes and independently verified hazard control (International Organization for Standardization, 2018; International Organization for Standardization & International Electrotechnical Commission, 2023; National Institute of Standards and Technology, 2023).

Limitations and research ethics

Every numerical observation in this article is synthetic. The experiment has no real workers, employers, injuries, near misses or completed interventions. No ethics approval or informed consent was necessary for the computational records, but any later study of employees would require its own ethics and data-protection procedures.

The detector is deliberately a simple score generator, not a trained AI model. The capacity penalty is proportional and spreads response effort uniformly; real teams may prioritise serious warnings, encounter queue delays or process duplicate alerts. The three q values are scenario assumptions, not estimated competency effects. The burden weights are illustrative and do not include a model of injury severity or the wider consequences of surveillance. The use of common random numbers strengthens comparison within the experiment but cannot create real-world representativeness.

The reported threshold preference is chosen from the same finite grid whose results are displayed,

rather than from a separately validated policy-selection sample. Therefore the results should not be translated into field deployment thresholds. An organisation would have to calibrate risk, evaluate legal duties and the hierarchy of controls, test its workforce's actual response process and verify that changed decision rules do not create new harm.

VI. A FEASIBLE VALIDATION PATHWAY

A useful next study would focus on one hazard with an established baseline: for example, vehicle-pedestrian separation breaches in a logistics setting. Before any AI tool influences work, the organisation would review current engineering and administrative controls, set measurable objectives, consult workers and map what constitutes a meaningful hazard opportunity. This example is a proposed study design, not a workplace in the existing dataset.

A shadow-mode phase should test the detector under different operating conditions while workers continue using existing controls. Researchers could compare its alerts with independently reviewed observations. Only after the error profile and privacy safeguards are acceptable should a supervised pilot record actual response time, available capacity, escalation, overrides and verified corrective action. A comparative evaluation would then examine exposure and other OHS indicators, controlling as far as possible for changes in staffing, production, reporting and work practices.

These steps would make it possible to replace the assumed q and K parameters with evidence, test non-uniform alert priority, and see whether the management model predicts operational performance. Field evidence may reject parts of the simplified simulation. That would be progress: a responsible model should change when the workplace tells a different story.

VII. CONCLUSION

A predictive warning is useful only if it changes a decision while prevention is still possible. This simulation shows that the answer depends not only on discrimination and thresholds but also on drift,

false-alert burden and a finite ability to respond. It also shows why a threshold cannot be chosen responsibly without discussing the relative consequences of what is missed and what must be investigated.

The reported numbers are computational results, not claims of reduced injuries. The practical message is to design the whole detection-to-action chain. Before asking how many hazards an AI system can identify, a workplace should ask how many meaningful warnings its people can understand, investigate and close safely.

Declarations and data availability

Study design: entirely computational simulation using generated data. No real workers, workplaces, accident records or identifiable personal data were used. Ethics approval: not applicable to the synthetic data. Computational materials: the source thesis describes a generator (`simulate.py`), `virtual_units.csv`, `scenario_results.csv` and `metadata.json`; access to those underlying files must be confirmed before journal submission. Funding and competing interests: to be completed by both authors.

REFERENCES

1. El-Helaly, M. (2024). Artificial intelligence and occupational health and safety, benefits and drawbacks. *La Medicina del Lavoro*, 115(2), e2024014.
<https://doi.org/10.23749/mdl.v115i2.15835>
2. Fang, W., Ding, L., Love, P. E. D., Luo, H., Li, H., Peña-Mora, F., Zhong, B., & Zhou, C. (2020). Computer vision applications in construction safety assurance. *Automation in Construction*, 110, 103013.
<https://doi.org/10.1016/j.autcon.2019.103013>
3. International Organization for Standardization. (2018). ISO 45001:2018 occupational health and safety management systems—Requirements with guidance for use. ISO.
4. International Organization for Standardization & International Electrotechnical Commission. (2023). ISO/IEC 42001:2023 information technology—Artificial intelligence—Management system. ISO.

5. Jetha, A., Bakhtari, H., Irvin, E., et al. (2025). Do occupational health and safety tools that utilize artificial intelligence have a measurable impact on worker injury or illness? Findings from a systematic review. *Systematic Reviews*, 14, 146. <https://doi.org/10.1186/s13643-025-02869-1>
6. Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102. <https://doi.org/10.1002/sim.8086>
7. Niehaus, S., Hartwig, M., Rosen, P. H., & Wischniewski, S. (2022). An occupational safety and health perspective on human in control and AI. *Frontiers in Artificial Intelligence*, 5, 868382. <https://doi.org/10.3389/frai.2022.868382>
8. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). <https://doi.org/10.6028/NIST.AI.100-1>
9. Shah, I. A., & Mishra, S. (2024). Artificial intelligence in advancing occupational health and safety: An encapsulation of developments. *Journal of Occupational Health*, 66(1), uiad017. <https://doi.org/10.1093/joccuh/uiad017>
10. Trask, C. M., & Linderth, H. C. J. (2023). Digital technologies in construction: A systematic mapping review of evidence for improved occupational health and safety. *Journal of Building Engineering*, 80, 108082. <https://doi.org/10.1016/j.jobe.2023.108082>