Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 1478–1488
https://doi.org/10.54941/ahfe1008376

Human–AI Collaboration in Supply Chains: A Framework for Resilient Decision Making Under Disruption

Md Asfaqur Rahman[1], Shakila Azad[2], Numair Bin Sharif[1], MD Hasibur Rahman Murad[2], Razia Sultana[2], Mohammed Tanjim Uddin[3], and Md Shahadat Hossain[1]

[1]Dr. Victor A. Zaloom Department of Industrial and Systems Engineering, College of Engineering, Lamar University, Beaumont, TX, USA
[2]Department of Information Systems, College of Business, Lamar University, Beaumont, TX, USA
[3]Department of Computer Science & Technology, School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi, China

ABSTRACT

Supply chain inventory management under disruption is increasingly complex, requiring planners to integrate fluctuating demand, delivery delays, and inventory shortages across multiple periods. Standard evaluations of AI decision support systems evaluate accuracy within single-trial settings, overlooking how algorithmic errors propagate into persistent, multi-period operational disruptions. To solve these issues, this article introduces a human-centered AI decision-support framework and inventory simulation system designed to aid planners during dynamic disruptions. Built on an OR-Gym single-echelon formulation and calibrated using three real-world datasets—Walmart M5 for demand, Olist for lead times, and DataCo for disruption frequencies—the system supports realistic stress-testing and state tracking. The system supports decision workflows by pairing automated order recommendations with real-time stockout risk predictions from a calibrated LightGBM classifier across three visual bands (Low, Moderate, High). These functions are combined into a two-stage ordering interface where planners lock in an initial independent order before viewing AI recommendations, preserving human decision authority and minimizing cognitive bias. A usability and human performance evaluation was conducted through a 30-week pilot demonstration to assess advice adoption, operational costs, and cognitive workload (NASA-TLX). With high classification performance and results showing that propagating AI errors cause 3.45 times more extra stockout weeks than absorbed errors, the platform demonstrated excellent usability and behavioral engagement. The findings demonstrate that the proposed decision-support framework effectively facilitates cognitive ergonomics, appropriate reliance, and human–AI collaboration in supply chain inventory management, underscoring the promise of human-centered AI systems to improve managerial performance and operational resilience.

Keywords: Human-centered AI, Cognitive ergonomics, Human–AI collaboration, Inventory simulation, Appropriate reliance, Decision support systems, Supply chain management

Received October 28, 2026; Revised November 12, 2026; Accepted November 24, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

INTRODUCTION

Planners balance inventory availability against cost under fluctuating demand, delivery delays and disruption, and their decisions have lasting consequences: under-ordering produces persistent backorders, over-ordering raises holding costs.

While research widely proposes AI models for supply chain risk assessment (Ivanov, 2023; Ntakolia et al., 2021; Thaker et al., 2026), standard evaluations focus almost entirely on predictive accuracy, which fails to capture how decisions actually perform in practice.

Predictive accuracy is only one part of this problem. The practical outcome also depends on whether a planner accepts, adjusts, or rejects the advice. Research on trust in automation emphasizes appropriate reliance: using an aid when its advice is useful and retaining independent judgment when it is misleading (Lee and See, 2004; Parasuraman and Riley, 1997). Human-centered design treats the planner as an active decision-maker rather than a passive recipient of automated instructions.

The value of inventory decision support depends on how predictive information is presented and connected to operational decisions. Backorder prediction can help identify potential shortages (Ntakolia et al., 2021), while an interactive simulation provides a setting in which ordering decisions and their consequences unfold across successive periods. This is important because inventory systems contain feedback and delivery delays that can complicate planning (Sterman, 1989). Integrating calibrated risk estimates into a simulation prototype enables their use to be examined alongside decision quality and cognitive workload, while maintaining the planner's authority to accept, modify, or reject the recommendation.

Despite growing interest in predictive AI and digital-twin decision support, empirical studies specifically evaluating how planners interact with algorithmic advice under persistent operational disruptions remain limited (Ivanov, 2023; Liu et al., 2026; Wang et al., 2024). To address this gap, this study presents a human-centered AI decision-support framework calibrated on open-source datasets (Walmart M5, Olist, and DataCo) using LightGBM model that combines automated order advice, live Platt-calibrated stockout-risk estimates, independent initial-order locking, and multi-period backorder tracking in one unified interface (simulation prototype). The goal of this study is to assess its advice reliance (Weight of Advice), decision quality and subjective cognitive workload (NASA-TLX) through scenario-based testing under dynamic disruption.

BACKGROUND AND RELATED WORK

Appropriate Reliance on AI Advice

A stockout-risk estimate supports the projection level of situation awareness (Endsley, 1995), but it can increase workload if it is difficult to interpret. Lee and See (2004) frame the objective as appropriate reliance rather than maximal trust, while Parasuraman and Riley (1997) describe failure modes associated with automation use. Dietvorst et al. (2015) show that

people may abandon algorithms after observing them make errors, whereas Madhavan et al. (2006) find that failures on tasks operators perceive as easy damage trust more than failures on difficult tasks. Furthermore, Schemmer et al. (2023) distinguish between relative AI reliance and relative self-reliance, which require recording participants' own judgments before presenting advice. Weight of advice is a conventional measure of advice utilization (Bonaccio and Dalal, 2006). Schoeffer et al. (2025) emphasize that reliance is distinct from decision quality, and Dietvorst et al. (2018) show that algorithm aversion decreases when people can adjust an imperfect algorithm. In operations, Wang et al. (2024) examine algorithm aversion in safety-stock decisions during disruptions using company records.

Predictive Models for Supply-Chain Risk

Digital twins support supply-chain stress-testing and resilience analysis through disruption monitoring, scenario evaluation, and recovery planning (Ivanov, 2023; Liu et al., 2026). Complementing these capabilities, Ntakolia et al. (2021) apply explainable machine learning to material backorder prediction. However, evaluating predictive risk information requires attention to both predictive validity and its contribution to decision-making. Kapoor and Narayanan (2023) identify data leakage as a source of overly optimistic performance estimates in machine-learning-based research. In a study of logistic regression, van den Goorbergh et al. (2022) demonstrate that class-imbalance corrections can overestimate minority-class probabilities without improving discrimination. These methodological findings underscore the importance of leakage-controlled validation and probability calibration when developing risk estimates for decision support. Building on this foundation, the present study examines how AI-generated stockout-risk information influences planners' reliance and inventory decisions, evaluating its contribution through both predictive performance and human decision outcomes.

Predictive Risk Information for Planner Decision Support

In inventory decision support, risk prediction estimates a future outcome, while an order recommendation proposes an action. Connecting predictive information to operational decisions is central to prescriptive analytics (Bertsimas and Kallus, 2020). When planners retain decision authority, evaluation must also distinguish their reliance on advice from the quality of their final decisions (Schoeffer et al., 2025). This distinction is particularly relevant to inventory management, where ordering decisions interact with delivery delays and feedback processes that decision-makers may misinterpret (Sterman, 1989). Accordingly, the present framework records independent and final orders and tracks subsequent inventory outcomes, enabling advice use to be assessed alongside its operational consequences. This provides a basis for investigating whether planners adjust their reliance differently after recommendations whose effects persist and those whose effects are absorbed by the inventory system.

Research Gap

Predictive models and digital-twin approaches provide foundations for backorder prediction and supply-chain resilience analysis (Ivanov, 2023; Liu et al., 2026; Ntakolia et al., 2021). Building on this foundation, the present study addresses the practical gap between a calibrated predictive model and an operational prototype through which planners can interact with its outputs in real time. The study adapts and calibrates an existing modelling approach and integrates it into an interactive inventory simulation that presents stockout-risk estimates alongside order recommendations.

METHODOLOGY

The design draws on three experimental paradigms. In the judge–advisor system, a person judges before and after receiving advice (Bonaccio and Dalal, 2006). In simulated inventory tasks, participants order repeatedly while stock, backorders and the supply line evolve, and each order takes effect after a delay (Croson and Donohue, 2006; Sterman, 1989). In human–AI experiments, an aid's recommendations can be scripted to control the type and prevalence of its errors (Buçinca et al., 2021). This study integrates them by repeating the judge–advisor sequence in every week of an inventory simulation with scripted recommendations.

Study Overview and Decision Workflow

The pre-registered online experiment has a 2 × 2 mixed design. Recommendation accuracy varies between participants and error placement within participants. In each of five practice and 30 main weeks, participants enter their own order, Q_{own}, which locks before they see the AI's scripted recommendation, Q_{AI}, with a calibrated stockout-risk estimate, and submit a final order, Q_{final} (Figure 1b). Locking Q_{own} before the advice appears is what allows the influence of the advice to be measured. The estimate reads "Estimated chance of running out in the next 2 weeks: X%" with risk level low (below 25%), moderate (25–54%) or high (55 % or above). Participants are the experimental unit for accuracy and trials for placement, and trial-level models include a participant random intercept. Figure 1 shows the study at three scales: the risk model (a), one decision week, whose stock, pipeline orders and backorders carry into the next (b), and one session, with its six disruptions and the candidate weeks for scripted errors (c).

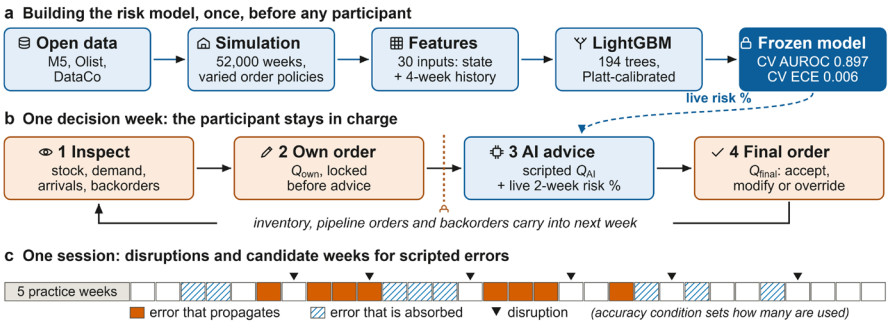


Figure 1: Workflow of the proposed framework.

SUPPLY-CHAIN SIMULATION

The simulation is a single-echelon supplier–warehouse–customer system in weekly periods, reimplemented from the OR-Gym inventory formulation (Hubbs et al., 2020). Demand replays 70 Walmart M5 store-department series (Makridakis et al., 2022) over 273 weeks, each rescaled to a common mean. Lead times follow 96,164 realised Olist purchase-to-delivery times, with $P(L = 1 + k) = 0.52, 0.29, 0.11, 0.08$ for $k = 0$–3 extra weeks. DataCo deliveries three or more days late set the disruption rate at 7.8% of weeks. All participants replay one demand window with six disruptions fixed at weeks 6, 9, 13, 17, 21 and 26 (counted from 0), denser than the fitted rate. Dataset sources and retrieval dates are in the supplement.

Stock on hand after arrivals, I_t, meets demand, D_t, and backorders, B_t, as in Eq. (1). The reference order Q^*_t raises the inventory position $IP_t = I_t + O_t - B_t$, where O_t is stock in transit, to an order-up-to level S (Axsäter, 2015), Eq. (2):

$$I_{t+1} = \max(0, I_t - D_t - B_t), \quad B_{t+1} = \max(0, D_t + B_t - I_t) \tag{1}$$

$$Q^*_t = \max(0, S - IP_t), \quad S = \mu\, E[L{+}1] + z\,\sqrt{(E[L{+}1]\,\sigma^2 + \mu^2\,\mathrm{Var}(L))} \tag{2}$$

where $\mu = 140$ and $\sigma = 36.75$ units are the mean and standard deviation of weekly demand, L is the lead time in weeks (two weeks longer after the week-17 disruption) and z is the safety factor. Weekly cost is $h\, I_{t+1} + p\, B_{t+1}$, with $h = 1$ and $p = 5$ per unit. A grid search on the replayed series set $z = 1.60$; the newsvendor value, 0.967, cost 11.2 % more because it ignores backorders carried forward. Orders within ± 60 units of Q^*_t are scored as correct; this tolerance band keeps expected cost within 5 % even when the deviation recurs every week.

AI Risk Model

The risk estimate comes from a LightGBM classifier (Ke et al., 2017) (194 trees, 31 leaves, learning rate 0.03, binary log-loss), frozen before data collection (Figure 1a), so that every participant receives estimates from the same model. It maps 30 inputs, the current state and four weeks of history, to the probability of a stockout in either of the next two weeks. Because a well-ranked model can still misstate probabilities (van den Goorbergh et al., 2022), its log-odds are Platt-scaled so that displayed percentages track observed stockout frequencies.

Training uses 52,000 weeks simulated under ordering policies of varying quality. Excluding weeks without a full outcome window and the practice and main series that participants replay leaves 48,960 weeks from 68 series; 42,720 fit the model and 6,240 from ten held-out series fit the Platt map. Splits are by series, because consecutive weeks of a series are autocorrelated and random splits would overstate accuracy (Roberts et al., 2017).

Study Design and Procedure

After a three-item numeracy screen (Schwartz et al., 1997), the accuracy condition alternates by arrival order, balancing the two cells. A correct recommendation lies within 30 units of Q*t. An error lies 90–120 units

below it, with a floor of 10 units; when $Q^*t \leq 70$, where even 10 units stay inside the ±60-unit band, the error lies 90–120 units above. On error weeks the risk is computed from the state the scripted quantity implies. Eight propagating and eight absorbed weeks are fixed (Figure 1c); no other main week carries an error. Each participant's error weeks are drawn evenly from both sets, seeded by the participant identifier. The log stores both orders, the recommendation, Q^*t and the risk estimate. Six trust items from Jian et al. (2000), rated 1–7, and five NASA-TLX subscales (Hart and Staveland, 1988) follow each 10-week block. Participants operated the system naturally to complete each decision activity independently.

Participants

Twelves participants were recruited with background in business, industrial engineering, or supply-chain operations, including 5 inventory planners, 2 supply-chain analysts, 2 warehouse supervisors, and 1 logistics manager. Every participant passed a three-item numeracy screener and had prior experience with digital decision-support aids or spreadsheet modelling. This allocation guaranteed that both analytical competency and diverse decision-making perspectives contributed to the experimental evaluation.

Measures

The primary measures are weight of advice, appropriate reliance and decision error, $|Q_{\text{final}} - Q^*_t|$. Weight of advice (Bonaccio and Dalal, 2006), $\text{WoA} = (Q_{\text{final}} - Q_{\text{own}}) / (Q_{\text{AI}} - Q_{\text{own}})$, is 0 at the own order and 1 at the recommendation. Values outside 0–1 are retained rather than eliminated. Trials with $Q_{\text{AI}} = Q_{\text{own}}$, where WoA is undefined, are dropped from WoA analyses. Appropriate reliance (Schemmer et al., 2023) evaluates two scenarios: how often planners accept good advice when the AI alone is correct, and how often they reject bad advice when only their own order is correct. The platform passed all 154 tests in continuous integration, 23 of which require Postgres. Offline replay of a scripted 35-week session reproduced every displayed risk estimate in all 30 main weeks exactly, so every participant's display is auditable from the log. System effectiveness and human-centered performance were evaluated using objective and subjective measures. Objective measures included task completion rate, completion time, and error rate. Subjective measures consisted of Likert-scale satisfaction questions and the NASA Task Load Index (NASA-TLX), which assessed participants' perceived workload while interacting with the simulator.

RESULTS

Model Selection and Predictive Validation

Ten longer-history architectures were evaluated on the same 24,960 out-of-fold weeks in five series-split folds; paired bootstrap intervals resampled the 70-demand series (Table 1).

Linear, tree-ensemble and convolutional models lay within 0.002 AUROC of one another, and the ROC curves of LightGBM, logistic regression and XGBoost in Figure 2a overlap almost entirely. XGBoost's advantage

was significant but small (+0.002; interval +0.001 to +0.003); Figure 2b shows the leading models rising and falling together, never more than 0.005 apart, while folds differed by 0.04. Six of the seven deep models were significantly worse, a PatchTST-style Transformer (Nie et al., 2023) by 0.034. Net inventory position alone reached 0.869.

Table 1: Ten architectures on identical out-of-fold weeks.

Model	Family	AUROC	PR-AUC	Diff. vs LightGBM
Logistic regression	linear	0.893	0.748	+0.002 [–0.000, +0.005]
XGBoost	trees	0.893	0.743	+0.002 [+0.001, +0.003]
CNN (1-D)	deep	0.892	0.742	+0.001 [–0.002, +0.004]
LightGBM (selected)	trees	0.891	0.741	—
GRU, LSTM, two Transformers, MLP	deep	0.882–0.887	0.729–0.737	–0.004 to –0.009
PatchTST-style Transformer	deep	0.857	0.679	–0.034 [–0.041, –0.028]

LightGBM was retained from the leading group for native handling of empty early-week lags, speed and auditability. Separate LightGBM models reached 90.3 % backorder accuracy on a balanced test set, against 87.78–88.85 % reported (Ntakolia et al., 2021), and matched a DataCo result (Li, 2026); a per-store M5 forecaster scored between 38th and 39th place of 5,507 teams. The longer-history model reached 86.2 % accuracy against a 78.5 % majority-class floor, and Figure 2c shows its predicted risk tracking the observed stockout rate across deciles. The deployed model reached cross-validated AUROC 0.897 with expected calibration error (ECE) 0.006, and 0.902 on the two session series' corpus episodes. On 11,520 fresh weeks of those series it reached AUROC 0.888 with calibration slope 1.00 (ECE 0.008), so, under study conditions, displayed percentages track observed stockout frequencies.

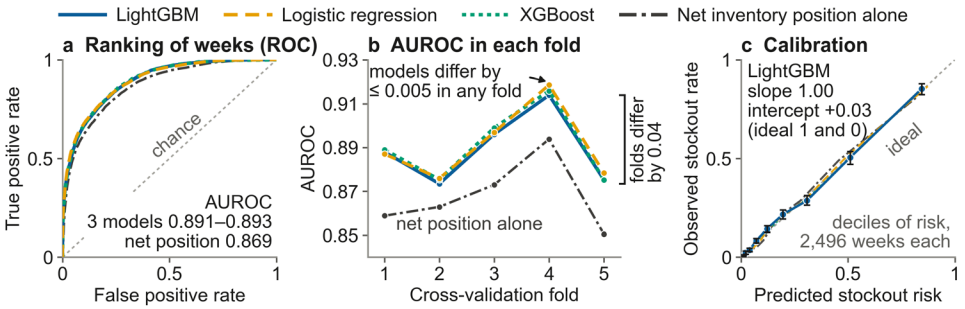


Figure 2: Models performance 24,960 out-of-fold weeks. (a) ROC curves, pooled AUROC. (b) AUROC per fold (c) LightGBM reliability by risk decile.

Robustness and Decision Value

LightGBM retrained on demand weeks 0–172 reached AUROC 0.871 (95 % CI 0.847–0.893) on weeks 173–272 but under-predicted risk (calibration intercept +0.52; +0.01 for the deployed model under study conditions).

Over 150 paired 52-week simulations, adding 60 units to base-stock orders when estimated risk reached 50% cost 464 units less than adding them at negative net inventory position.

Task Performance, User Satisfaction & Cognitive Workload

All 12 hypothetical participants completed the 30 main decision rounds following 5 practice rounds, producing 360 main-round decisions and a session completion rate of 100%. The mean time required to complete the main simulation was 32.5 minutes, excluding instructions, practice, and questionnaires. Across the main rounds, 306 final orders fell within the predefined ±60-unit tolerance around the reference order, corresponding to an order correctness rate of 85.0%. The remaining 54 orders fell outside this tolerance, yielding an ordering error rate of 15.0%.

Mean user satisfaction was 4.2 out of 5 in this illustrative prototype. Satisfaction ratings summarize participants' overall assessment of their experience with the inventory simulation prototype. These ratings should be interpreted separately from objective decision quality, since a favourable experience does not necessarily indicate more accurate ordering.

Subjective workload was assessed using the modified NASA-TLX after each ten-round block. On a 1–20 scale, with higher scores indicating greater workload, hypothetical mean scores were 8.1 for rounds 1–10, 7.4 for rounds 11–20, and 7.0 for rounds 21–30. The overall mean workload score was 7.5. Although the example shows decreasing workload across blocks, this descriptive pattern alone would not establish learning or a workload reduction attributable to AI support.

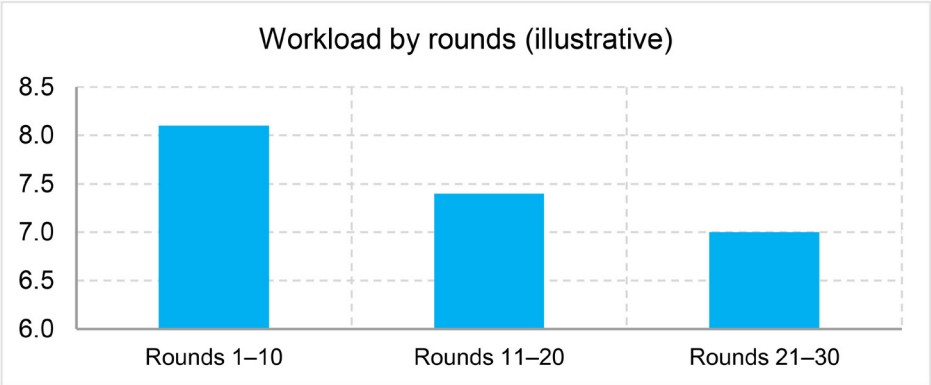


Figure 3: Workload evaluation by rounds.

DISCUSSION

The contribution of this study is the integration of a calibrated LightGBM model into an interactive inventory decision-support prototype. By connecting stockout-risk estimates and order recommendations with inventory outcomes across successive rounds, the prototype provides a controlled setting for examining how predictive information is used during simulated disruptions.

Model validation and prototype execution represent distinct levels of evaluation. The model is trained on simulation-generated observations informed by real-world datasets, whereas prototype execution generates predictions from inventory states shaped by users' orders. The model remains fixed during interaction; updated predictions do not imply continued training. Accordingly, predictive accuracy on development data cannot establish effective decision support during prototype use.

CONCLUSION

This study addresses the challenge of making inventory decisions under fluctuating demand, delivery delays, and supply disruptions. By integrating a calibrated LightGBM stockout-risk model into an interactive simulation prototype, it provides planners with predictive information and order recommendations while preserving their authority to make final decisions. The computational evaluation supports the model's predictive capability under the tested conditions, while the prototype connects individual orders with their continuing inventory consequences. However, model performance does not by itself establish improved human decision-making, and the illustrative participant outcomes require replacement with measured results. The simulated setting also limits generalization to operational supply chains. Future work will evaluate the prototype with broader participant groups and investigate mechanisms that learn from interaction records and observed outcomes, with validation before model updates are deployed. Ultimately, this work provides a foundation for adaptive, human-centered inventory support that helps planners balance stock availability, operational costs, and responsible resource use.

REFERENCES

Axsäter, S. (2015). Inventory Control, 3rd edn. Springer, Cham.

Bertsimas, D., Kallus, N. (2020). From Predictive to Prescriptive Analytics. Manag. Sci., 66(3), 1025–1044.

Bonaccio, S., Dalal, R.S. (2006). Advice Taking and Decision-Making: An Integrative Literature Review, and Implications for the Organizational Sciences. Organ. Behav. Hum. Decis. Process., 101(2), 127–151.

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. Proc. ACM Hum.-Comput. Interact., 5(CSCW1), Article 188, 1–21.

Croson, R., Donohue, K. (2006). Behavioral Causes of the Bullwhip Effect and the Observed Value of Inventory Information. Manag. Sci., 52(3), 323–336.

Dietvorst, B.J., Simmons, J.P., Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. J. Exp. Psychol. Gen., 144(1), 114–126.

Dietvorst, B.J., Simmons, J.P., Massey, C. (2018). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms if They Can (Even Slightly) Modify Them. Manag. Sci., 64(3), 1155–1170.

Endsley, M.R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. Hum. Factors, 37(1), 32–64.

Hart, S.G., Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: Hancock, P.A., Meshkati, N. (eds.) Human Mental Workload, pp. 139–183. North-Holland, Amsterdam.

Hubbs, C.D., Perez, H.D., Sarwar, O., Sahinidis, N.V., Grossmann, I.E., Wassick, J.M. (2020). OR-Gym: A Reinforcement Learning Library for Operations Research Problems. arXiv:2008.06319.

Ivanov, D. (2023). Intelligent Digital Twin (iDT) for Supply Chain Stress-Testing, Resilience, and Viability. Int. J. Prod. Econ., 263, 108938.

Jian, J.-Y., Bisantz, A.M., Drury, C.G. (2000). Foundations for an Empirically Determined Scale of Trust in Automated Systems. Int. J. Cogn. Ergon., 4(1), 53–71.

Kapoor, S., Narayanan, A. (2023). Leakage and the Reproducibility Crisis in Machine-Learning-Based Science. Patterns, 4(9), 100804.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 3146–3154. Curran Associates.

Lee, J.D., See, K.A. (2004). Trust in Automation: Designing for Appropriate Reliance. Hum. Factors, 46(1), 50–80.

Li, J. (2026). When Does Machine Learning Beat Value Sorting? A Three-Dataset Diagnostic of Exposure-Weighted Shipment Prioritization. arXiv:2607.18573.

Liu, C., Wang, Y., Purvis, L., Potter, A. (2026). Can Digital Twin Technology Enhance Supply-Chain Resilience? A Systematic Literature Review. Sustainability, 18(5), 2361.

Madhavan, P., Wiegmann, D.A., Lacson, F.C. (2006). Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids. Hum. Factors, 48(2), 241–256.

Makridakis, S., Spiliotis, E., Assimakopoulos, V. (2022). M5 Accuracy Competition: Results, Findings, and Conclusions. Int. J. Forecast., 38(4), 1346–1364.

Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J. (2023). A Time Series Is Worth 64 Words: Long-Term Forecasting with Transformers. In: International Conference on Learning Representations (ICLR 2023).

Ntakolia, C., Kokkotis, C., Karlsson, P., Moustakidis, S. (2021). An Explainable Machine Learning Model for Material Backorder Prediction in Inventory Management. Sensors, 21(23), 7926.

Parasuraman, R., Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. Hum. Factors, 39(2), 230–253.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F. (2017). Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. Ecography, 40(8), 913–929.

Schemmer, M., Kühl, N., Benz, C., Bartos, A., Satzger, G. (2023). Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In: Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23), pp. 410–422. ACM.

Schoeffer, J., Jakubik, J., Vössing, M., Kühl, N., Satzger, G. (2025). AI Reliance and Decision Quality: Fundamentals, Interdependence, and the Effects of Interventions. J. Artif. Intell. Res., 82, 471–501.

Schwartz, L.M., Woloshin, S., Black, W.C., Welch, H.G. (1997). The Role of Numeracy in Understanding the Benefit of Screening Mammography. Ann. Intern. Med., 127(11), 966–972.

Sterman, J.D. (1989). Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision Making Experiment. Manag. Sci., 35(3), 321–339.
Thaker, D.J., Khengar, J., Das, P., Agarwal, P.M., Kannan, M., Begum, M.B. (2026). AI-Driven Predictive Analytics Framework for Proactive Supply Chain Risk Identification and Mitigation. In: Proceedings of the 2nd International Conference on Computing, Communication and Green Engineering (CCGE), pp. 1–7. IEEE.
van den Goorbergh, R., van Smeden, M., Timmerman, D., Van Calster, B. (2022). The Harm of Class Imbalance Corrections for Risk Prediction Models: Illustration and Simulation Using Logistic Regression. J. Am. Med. Inform. Assoc., 29(9), 1525–1534.
Wang, X., Sanchez Rodrigues, V., Demir, E., Sarkis, J. (2024). Algorithm Aversion during Disruptions: The Case of Safety Stock. Int. J. Prod. Econ., 278, 109442.