Interpretable AI in Human-Machine Systems: Insights from Human-in-the-Loop Product Recommendation Engines

Nima Safaei,
Scotiabank

Scotia Plaza, 40 King St W, Toronto, ON
M5H 3Y2, Canada
nima.safaei@scotiabank.com

Pooria Assadi
Associate Professor of Management and
Organizations, College of Business,
California State University
Tahoe Hall 2012, 6000 J Street, Sacramento,
CA 95819-6088, U.S.A
pooria.assadi@csus.edu

Abstract

Recent advances in human-machine systems have renewed a commonly-held expectation that Machine Learning (ML) may be most effectively used in conjunction with human intervention. This expectation is built on the assumption that synthesized human-machine systems which bring humans in the loop by allowing them to provide input, oversight, or supervision outperform humans or machines alone and create a whole that is superior to the sum of its parts. Despite the appeal of such an expectation, what we know about the technical and practical requirements of effective ML utilization has not been applied to carefully consider when human-machine systems might deliver on such expectations: i.e., outperform humans or machines alone. In this paper, we showcase the importance of the recognition and quantification of the false alarms in any technically sound and practically interpretable analysis of the effectiveness of ML systems. Specifically, we propose that the quantification of the costs and risks of the ML-generated false alarms is directly tied to tuning the regularization hyper-parameters, and consequently reducing the complexity and improving the effectiveness of ML systems. Using a series of A/B experiments and simulations, we demonstrate the application of our theory to tease out the effectiveness of two popular human-centric product recommendation engines: Assessment Based Recommendation (ABR) where the customers are primarily filtered by human assessment; and Broad-Spectrum Recommendation (BSR) where the new product or service is introduced to all possible customers. We show that non-existent recognition or incorrect quantification of the false alarms undermines the measurability and interpretability of the economic value of using ML (absent or in conjunction with humans) to the extent that it might render it practically unjustifiable, and postulate conditions under which human-machine systems outperform humans or machines alone. By doing so, we call for researchers to transform how they conceptualize and utilize ML from one that is primarily concerned with accuracy-consistency trade-offs to one that also incorporates the costs and risks of the false alarms in the ML objective function to enhance its interpretability.

1 The Promise of Human-in-the-Loop ML

We examine the assumption that synthesized human-machine systems which bring humans in the loop outperform humans or machines alone [67, 22]. In such systems, “the machine provides “best actions” or “recommended actions” and the human can choose whether to take the actions or not” [85] where the “synthesis of human and machine will provide benefits beyond those achieved by humans or machines alone” and produces a “whole [which] is far greater than the sum of the parts” [74]. We apply what we know about the technical and practical requirements of effective ML utilization [82, 63] to carefully consider when human-machine systems might deliver on such expectations.

2 What Does an Efficient and Effective ML Utilization Require?

The benefits of ML structural estimation techniques and general business applications can only be realized if their underlying ML models perform well. Traditionally, the power of a ML model is considered to depend on how well it can predict or "fit" the unseen data—in addition to the seen data—with a high level of *accuracy* and *consistency*. The learning theory suggests that improving the power of a ML model demands achieving an optimal accuracy-consistency trade-off (i.e., bias-variance trade-off) [36]. In this context, though, on the one hand, on average, less complex linear or parametric models (i.e., underfit models) are consistent (i.e., low variance) but are inaccurate (i.e., high bias). On the other hand, on average, more complex non-linear or non-parametric models (i.e., overfit models) are accurate (i.e., low bias) but are inconsistent (i.e., high variance) [36]. Approaching optimal bias-variance trade-offs in ML models demands reducing underfitting and overfitting at the same time. A technical strategy that is often used to achieve an optimal bias-variance trade-off is *regularization* [93, 39].

In practice, regularization and false alarm quantification are closely related because the lack of knowledge about the cost or risk of false alarms can lead to an incorrect bias-variance trade-off. The key metrics for measuring the economic value of ML projects include cost savings, revenue growth, efficiency gains, improved decision making and customer experiences, and risk mitigation. Regardless of whether the application aims for top-line growth or bottom-line improvements, estimating the economic value of ML projects requires understanding the impact of false alarms on this value. For instance, if we overlook the effects of false alarms but find that personalized product suggestions boost revenue, we might misjudge the true impact of a project without considering lost opportunities or incorrect recommendations. Hence, an important practical question is: if false alarms were to have a negligible impact, would a costly ML pipeline be still beneficial?

3 A Practical Reinterpretation of The Hyper-Parameters

In this article, we offer a new practical interpretation for the regularization hyper-parameter λ to address the conundrum of hyper-parameter optimization; one that does not need determining additional quantities on ad-hoc bases and that has important implications for the performance of human-machine systems. We do so by using a mathematical transformation to transform how ML is conceptualized and used from one that mainly focuses on accuracy-consistency trade-offs to one that captures the costs and risks of the false alarms in the ML loss function. To explain our interpretation, consider a generic supervised model $Y = f(X; W)$ where W represents the vector of learning coefficients and where $\Gamma = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_t, Y_t), \dots (X_T, Y_T)\}$ represents the historical observations with X as independent variable set (i.e., feature set) and Y as dependent variable set (i.e., label set). Regularization is applied using the penalty function $g(W)$ and the regularization hyper-parameter λ.

Since the regularization hyper-parameter λ is a positive constant, without loss of generality, we rewrite the learning objective function as $min\, \gamma \ell(f(X; W), Y) + g(W)$, where $\gamma = \frac{1}{\lambda}$. We define γ as the unit cost or risk of a model's error or *false alarm* [47, 64, 26]. In other words, γ is the unit cost or risk of a misclassification, a wrong prediction, or an incorrect decision [67, 66]. We define γ in this way as prior research on the theory of the optimal classifier in decision-making in the fields of economic control chart design and optimal Bayesian risks suggests that a careful consideration of the "parameter that specifies how "dangerous" it is to misclassify" is important in using ML. One illustrative example would be misclassifying important emails in an ML-based spam filtering [81].

In this context, γ may represent a single entity or a matrix of false components. In fact, hyper-parameter optimization represents a conundrum because at its core it involves quantifying a matrix of more than one false component [48, 40, 62, 17, 88]. For example, in a binary classification setting, $\gamma = (\gamma_{FP}, \gamma_{FN})$ represents the unit cost or risk of both *false positive* (i.e., *FP or Type I error*) [84, 91, 77, 28, 33) and *false negative* (i.e., *FN or Type II error*) [10, 67] components. In a multi-class classification setting, however, γ can be represented in matrix form $[\gamma_{ij}]$—similar to the confusion matrix—such that the entity γ_{ij} is the unit cost or risk when an observation known to be in group i is predicted to be in group j where $\gamma_{ii} = 0\ \forall i$. We propose extending classification models' loss function as the following generic form

$$\sum_{i,j} \gamma_{ij} \ell_{ij}(f(X;W), Y), \text{ subject to } g(W) \leq c,$$

to be able to measure the economic value of ML projects. ℓ_{ij} counts the number of observations labelled to be in group i but predicted to be in group j and c is a complexity control parameter.

To conceptually illustrate our proposed reinterpretation, let $Y = f(X, Y; W)$ represent a recommendation engine that is designed to determine a set of targeted customers to whom a new product should be recommended [32]. In this context, then, the unit cost of a false positive γ_{FP} is the operational and dissatisfaction costs associated with a customer declination even though the trained model classified that target as a potential customer, and the unit cost of a false negative γ_{FN} is the opportunity costs of missing a potential customer. In this context, γ_{FP} and γ_{FN} capture the unit cost of incorrect detection because of the gap between the training and implementation phases of ML modeling; a gap that is often caused by the ignorance on prior knowledge and/or potential missing features. This example, and its underlying reinterpretation, can be extended to other settings beyond recommendation engines.

4 When Does ML Practically Require Human Intervention?

With our proposed interpretation, we argue that ML alone is not particularly useful for applications where either the unit cost or risk of error cannot be practically quantified (i.e., $\gamma \rightarrow \infty\ or\ \lambda \rightarrow 0$) or the unit cost or risk of error is negligible (i.e., $\gamma \rightarrow 0\ or\ \lambda \rightarrow \infty$). In the former, the risk or cost is practically intolerable (i.e., reaching infinity), and in the latter the risk or cost is irrelevant (i.e., so small that makes no difference). As an example of a case where $\gamma \rightarrow \infty$, consider an unsupervised ML model responsible for detecting the potential failures in aircraft engines [21]. Even though such model's accurate performance is highly desirable, it comes with a high risk of false alarms. In particular, any false positive is very costly due the consequences associated with aircraft downtime [72]. In addition, the risk of false negatives is extremely high due to any likely catastrophic failures or fatal crashes resulting in loss of life [73]. In such a scenario, decision makers cannot solely rely on the performance of ML models, and instead must consult with expert humans to ensure that the model outcomes are consistent with safety and reliability standards and regulations. In fact, in scenarios like this, where $\gamma \rightarrow \infty$, there is often no real practical opportunities for reducing human intervention and speeding up the decision-making processes using current ML techniques and applications. Where the unit cost or risk of false alarms is small for practical purposes (i.e., $\gamma \rightarrow 0$), it is often not useful to use ML at all, with or without human intervention. As an example of such a case, consider movie suggestion platforms (e.g., used in Netflix and Hulu). In applications like this, the least complex approach will prove to be more practically useful; that is, recommend the new product (e.g., new movie) to all customers (i.e., do not use ML) rather than expending ML deployment resources to target specific customers.

5 When Can ML Practically Go It Alone?

We posit that ML alone (i.e., without human intervention) is particularly useful for applications where the unit cost or risk of error or false alarm is practically quantified and nonnegligible (i.e., $0 < \gamma \ll \infty$). In the earlier scenario involving recommendation engines, suppose that $\gamma = (\$A, \$B)$ where $\$A$ is the unit opportunity cost of each false negative per customer and $B is the unit cost of each false positive per customer (i.e., fixed cost for personalized advertisement, marketing campaign, skilled labor, and information technology infrastructure). After training a hypothetical ML model given in-sample training and labelled data, the probability of a false positive (i.e., probability of targeting the wrong customer) is α^{M}% and the probability of a false negative (i.e., probability of ignoring a potential customer) is β^{M}%. Further suppose that the product selling price (i.e., revenue earned per acquired customer) is *$C* where *$A* = *$C* - *$B.*

Absent ML, various strategies could be applied to make the recommendation decisions. These strategies highly rely on human assessment and expertise (human-centric). To investigate when ML alone can practically be useful, we try to examine profitability, with ML versus absent ML. To do so, we examine two major common practices: (1) Assessment-based Recommendation and (2) Broad-Spectrum Recommendation. While in the former strategy, the customers will be primarily filtered by human assessment, in the latter strategy the new product/service will be introduced to all possible customers.

5.1 Assessment Based Recommendation (AbR)

This strategy relies on the human domain knowledge – including non-ML tools/methods – to make an initial assessment based on the customer's available information. Figure 1 summarized this process.

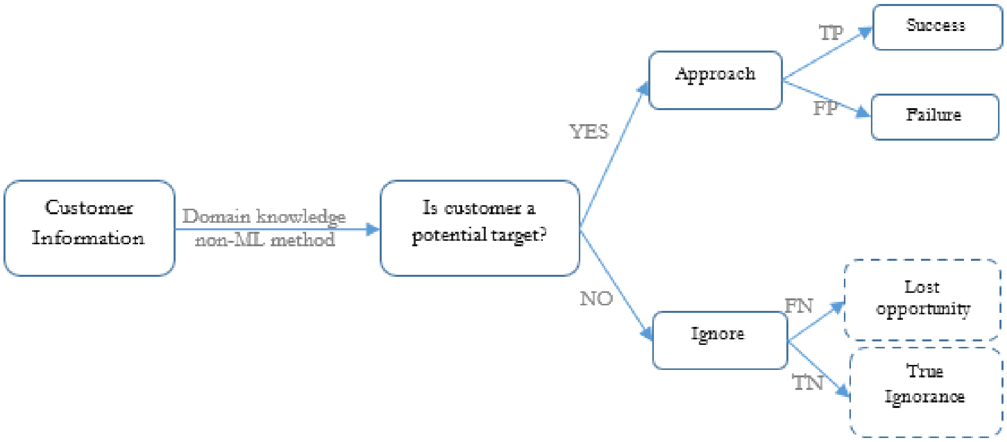

Figure 1 Assessment-based Recommendation

Similar to the ML scenario, assume a FP rate of α^{H}% and FN rate of β^{H}% associated with the process shown in Figure 1. In practice, the probability of a lost opportunity, β^{H}, cannot be easily estimated due to the lack of information when the customer is ignored because of the initial assessment. Some external knowledge such as having similar products from competitors might be useful to estimate β^{H}. The expected net profit that the AbR strategy generates, then, is:

$$(1-\beta^{\mathrm{H}})(C-B)-\alpha^{\mathrm{H}}B-\beta^{\mathrm{H}}(C-B) \qquad (1.1)$$

The above term consists of the expected net profit of a TP decision minus the expected loss of a FP decision and the expected opportunity cost of a FN decision. Notice that $1-\beta^{\mathrm{H}}$ is the probability of true guess (or true positive) which is also known as *Precision* in the ML field. It is the probability that the approached customer is successfully acquired. The expected net profit that the AbR strategy generates can be rewritten in terms of accuracy as:

$$(1-\beta^{\mathrm{H}})C-\vartheta^{\mathrm{H}}B-\beta^{\mathrm{H}}(C-B) \qquad (1.2)$$

where $\vartheta^{H}=1-(\alpha^{\mathrm{H}}+\beta^{\mathrm{H}})$ is the *accuracy* of the AbR strategy extracted from the confusion matrix.

Similar to term (1.2), the expected net profit associate to the ML scenario would be:

$$(1-\beta^{\mathrm{M}})C-\vartheta^{\mathrm{M}}B-\beta^{\mathrm{M}}(C-B) \qquad (2)$$

where $\vartheta^{M}=1-(\alpha^{\mathrm{M}}+\beta^{\mathrm{M}})$ is the machine accuracy. Having the alternative of an AbR strategy, a ML model alone would be useful (i.e., it is economically justifiable for practical purposes) only if the expected net profit earned because of using ML, term (1.2), is greater than the expected net profit earned absent ML, term (2). That is:

$$\frac{C-B}{C}>\frac{(\vartheta^{M}-\vartheta^{H})+(\beta^{M}-\beta^{H})}{(\vartheta^{M}-\vartheta^{H})-(\beta^{M}-\beta^{H})} \quad (3)$$

The left-hand side of Inequality (3) represents the *Net Profit Margin* (NPM) in a percentage form. The net profit margin provides insights into how effectively a business is managing costs relative to its selling prices. For context, companies with a NPM of 20% generally show strong financial health. If this metric drops to around 5% or lower, most businesses will need to make changes to remain sustainable [45]. Note that the significance of the NPM value can vary depending on the industry and the specific circumstances of the business. The right-hand side of Inequality (3) provides a lower bound on NPM (LB-NPM) beyond which the ML would not be economically profitable. In fact, the bigger the LB-NPM, the harder would be the case where the ML becomes profitable. For example, assuming that ϑ^{M} and ϑ^{H} are given, the possibility of ML profitability grows by decreasing the difference $\beta^{M}-\beta^{H}$.

For numerical illustration, we consider a major electronics retailer in the U.S which posted an average quarterly net profit margin of 5%. Figure 2 shows the scatter plots per pair values of $\vartheta^{M}$and ϑ^{H} where $\vartheta^{M},\vartheta^{H}\in\{0.7,0.75,0.8,0.85,0.9,0.95\}$. Each scatter plot represents the standardized LB-NPM values under various β^{M} and β^{H} where $\beta^{M}\in(0.01,1-\vartheta^{M}]$ and $\beta^{H}\in(0.01,1-\vartheta^{H}]$. While green areas represent ML profitability, the red points represent the areas where ML is not economically justifiable. The size of each point represents the magnitude of LB-NPM. The size of red points is in fact the expected loss per customer. Figure 2 provides valuable insights on ML profitability. For instance, when ML's accuracy is significantly higher than AbR's accuracy (i.e., the far right-top plot), ML would be profitable only if the greatest portion of AbR's inaccuracy, $1-\vartheta^{H}$, is because of the false negative, i.e., β^{H} (not false positive), regardless of the value of β^{M}. For another instance, when ML and AbR have nearly same accuracy (i.e., diagonal plots in Figure 2), ML is profitable everywhere except where $\beta^{H}\approx\beta^{M}$.

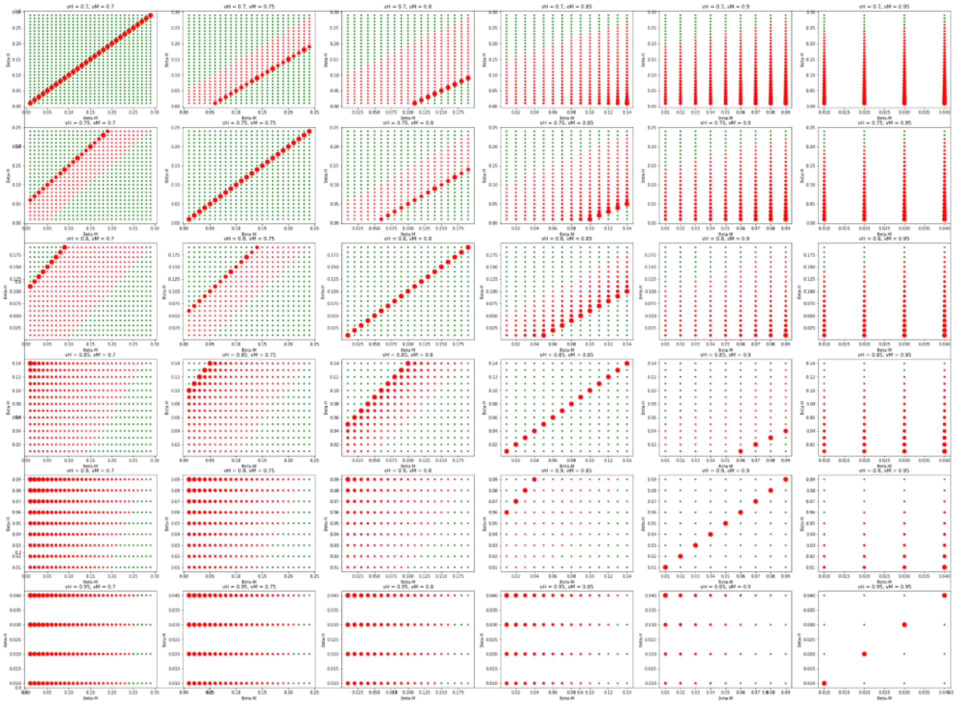

Figure 2: LB-NPM scatter plots assuming the AbR strategy and NPM = 5%

5.2 Broad-Spectrum Recommendation (BSR)

In this strategy the new product/service will be introduced to all existing customers without initial assessment. The expected net profit for BSR would be $(1-\beta^{\mathrm{H}})C - B$ where $1-\beta^{\mathrm{H}}$ is same as precision or equivalently the success rate for the customer acquisition. Having this strategy as an alternative, a ML model alone would be economically justifiable, only if Term (2) is strictly greater than $(1-\beta^{\mathrm{H}})C - B$, i.e.:

$$(1-\beta^{\mathrm{M}})C - \vartheta^{\mathrm{M}}B - \beta^{\mathrm{M}}(C-B) > (1-\beta^{\mathrm{H}})C - B,$$

or equivalently

$$\frac{C-B}{C} > \frac{\vartheta^{M}+\beta^{M}-\beta^{H}-1}{\vartheta^{M}-\beta^{M}-1} \qquad (4)$$

Inequality (4) reveals that under the BSR strategy, a ML could be justifiable only if the company has a significant high NPM (i.e., a strong financial situation). The comparison among Figures 3a (i.e., a weak financial condition) and 5b (i.e., a strong financial condition) reveals that the higher the NPM, the larger the possibility of ML justifiability (i.e., larger green areas) assuming parameter settings $\vartheta^{M} \in \{0.7, 0.75, 0.8, 0.85, 0.9, 0.95\}$, $\beta^{M} \in (0.01, 1-\vartheta^{M}]$ and $\beta^{H} \in (0.01, 0.3]$.

It is worthwhile to highlight an important assumption here. In practice, the quality of labelled data for the ML models may depend on the human performance metrices such as ϑ^{H} or μ^{H}. However, this dependency is beyond the scope of this research and is ignored for the sake of simplicity. Future work can consider this possibility.

Taken together, we argue that analysis of this kind using our proposed reinterpretation of γ that involves a careful quantification of the false alarms has the potential to uncover when using a ML technique or application alone is economically justifiable for practical purposes.

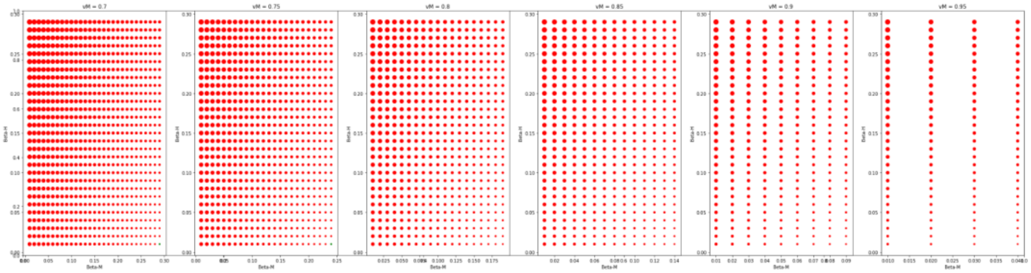

Figure 3a: LB-NPM scatter plots assuming the BSR strategy and NPM = 5%

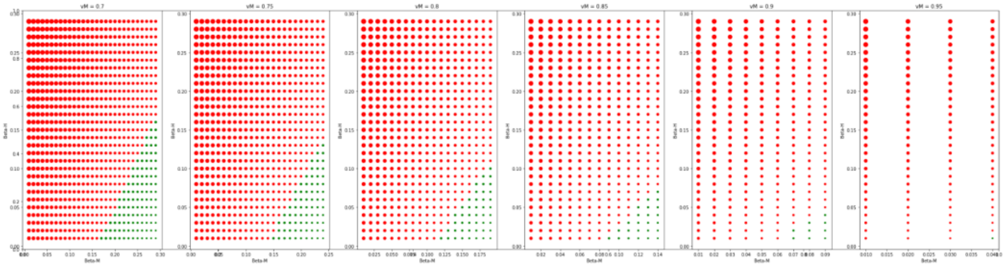

Figure 3b: LB-NPM scatter plots assuming the BSR strategy and NPM = 30%

6 Conclusion

Much of ML structural estimation techniques and applications in research and practice lack a meaningful cost-benefit analysis of their usefulness and interpretability. In this article, we identify the root cause of this limitation in the general lack of recognition and quantification of the false alarms [18, 70]. We particularly show that, at its core, the quantification of false alarms at the business level is directly tied to tuning the regularization hyper-parameters, reducing complexity, and enhancing the usefulness of ML. By offering a clear link between hyper-parameter regularization and false alarm quantification, we advance the theoretical and practical aspects of ML to highlight scenarios under which ML is practically useful versus not. In one concrete application, we specify conditions under which human-machine systems outperform humans.

References

[1] Abbass, H. (2021). What is Artificial Intelligence?, *IEEE Transactions on Artificial Intelligence*, 2(2), 94-95.

[2] Agarwal, A., Gans, J., and Goldfarb, A. (2020). How to Win with Machine Learning, *Harvard Business Review*.

[3] Amengual, D., Carrasco, M., & Sentana, E. (2020). Testing distributional assumptions using a continuum of moments. *Journal of Econometrics*, 218(2), 655-689.

[4] Amos, B. and Kolter, J.Z. (2017) OptNet: Differentiable Optimization as a Layer in Neural Networks, *International Conference on Machine Learning (ICML)*, Sydney, Australia.

[5] Ban, G. Y., El Karoui, N., & Lim, A. E. (2018). Machine learning and portfolio optimization. *Management Science*, 64(3), 1136-1154.

[6] Bartok, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J.; Csanyi, G.; Ceriotti, M. (2017) Machine Learning Unifies the Modeling of Materials and Molecules. *Science Advances* 3(12).

[7] Bergstra , J., Bardenet, R., Bengio, Y. and Kegl, B. (2011). *NIPS'11: Proceedings of the 24th International Conference on Neural Information Processing Systems*. 2546–2554.

[8] Bergstra , J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13, 281-305.

[9] Bertsimas D., King, A. (2016) OR Forum—An Algorithmic Approach to Linear Regression, *Operations Research*, 64(1), pp.2-16.

[10] Bish, D. R., Bish, E. K., Xie, R. S., & Stramer, S. L. (2014). Going beyond "same-for-all" testing of infectious agents in donated blood. *IIE Transactions*, 46(11), 1147-1168.

[11] Bishop, M. A., & Trout, J. D. (2005). *Epistemology and the psychology of human judgment*. Oxford University Press on Demand.

[12] Brown, D. (2019). Tesla stock drops after a report that Autopilot was engaged during a deadly crash in Florida, *USA TODAY*.

[13] Capizzi, G., & Masarotto, G. (2009). Bootstrap-based design of residual control charts. *IIE Transactions*, 41(4), 275-286.

[14] Carneiro, P., Lee, S., & Wilhelm, D. (2020). Optimal data collection for randomized control trials. *The Econometrics Journal*, 23(1), 1-31.

[15] Chang, N. C. (2020). Double/debiased machine learning for difference-in-differences models. *The Econometrics Journal*, 23(2), 177-191.

[16] Chen, A., & Chen, Y. K. (2007). Design of EWMA and CUSUM control charts subject to random shift sizes and quality impacts. *IIE Transactions*, 39(12), 1127-1141.

[17] Chen, S., & Nembhard, H. B. (2011). Multivariate Cuscore control charts for monitoring the mean vector in autocorrelated processes. *IIE Transactions*, 43(4), 291-307.

[18] Cheng, A. Y., Liu, R. Y., & Luxhøj, J. T. (2000). Monitoring multivariate aviation safety data by data depth: control charts and threshold systems. *IIE Transactions*, 32(9), 861-872.

[19] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*., p1-68.

[20] Choudhury, P., Allen, R. T., & Endres, M. G. (2021). Machine learning for pattern discovery in management research. *Strategic Management Journal*, 42(1), 30-57.

[21] Cottrell, M., Gaubert, P., Eloy, C., François, D., Hallaux, G., Lacaille, J., & Verleysen, M. (2009). Fault prediction in aircraft engines using self-organizing maps. *In International workshop on self-organizing maps* (pp. 37-44). Springer, Berlin, Heidelberg.

[22] Cowgill, B., & Tucker, C. E. (2020). Algorithmic fairness and economics. *The Journal of Economic Perspectives*.

[23] Chen, S., & Nembhard, H. B. (2011). Multivariate Cuscore control charts for monitoring the mean vector in autocorrelated processes. *IIE Transactions*, 43(4), 291-307.

[24] Cui, Q., Xu, Y., Zhang, Z., & Chan, V. (2021). Max-linear regression models with regularization. *Journal of Econometrics*, 222(1), 579-600.

[25] Elmousalami, H. H. (2020). Comparison of artificial intelligence techniques for project conceptual cost prediction: a case study and comparative analysis. *IEEE Transactions on Engineering Management*, 68(1), 183-196.

[26] Epprecht, E. K., Costa, A. F., & Mendes, F. C. (2003). Adaptive control charts for attributes. *IIE Transactions*, 35(6), 567-582.

[27] Escanciano, J. C., & Li, W. (2021). Optimal linear instrumental variables approximations. *Journal of Econometrics*, 221(1), 223-246.

[28] Fan, J., Feng, Y., & Xia, L. (2020). A projection-based conditional dependence measure with applications to high-dimensional undirected graphical models. *Journal of Econometrics*, 218(1), 119-139.

[29] Fang, C. and Liao, S. (2017). Scalable Gaussian Kernel Support Vector Machines with Sublinear Training Time Complexity, *Information Sciences*, 418-419, 480-494.

[30] Fève, F., & Florens, J. P. (2010). The practice of non-parametric estimation by solving inverse problems: the example of transformation models. *The Econometrics Journal*, 13(3), S1-S27.

[31] Florens, J. P., Johannes, J., & Van Bellegem, S. (2012). Instrumental regression in partially linear models. *The Econometrics Journal*, 15(2), 304-324.

[32] Galuzzi, B. G., Giordani, I., Candelieri, A., Perego, R., & Archetti, F. (2020). Hyperparameter optimization for recommender systems through Bayesian optimization. *Computational Management Science*, 17(4), 495-515.

[33] Giarratana, M. S., Mariani, M., & Weller, I. (2018). Rewards for patents and inventor behaviors in industrial research and development. *Academy of Management Journal*, 61(1), 264-292.

[34] Gupta, V., Han, B. R., Kim, S. H., & Paek, H. (2020). Maximizing intervention effectiveness. *Management Science*, 66(12), 5576-5598.

[35] Gupta, V., & Rusmevichientong, P. (2021). Small-data, large-scale linear optimization with uncertain objectives. *Management Science*, 67(1), 220-241.

[36] Rajnarayan, D., Wolpert, D. (2011). Bias-Variance Trade-offs: Novel Applications. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_75.

[37] Hazimeh, H. and Mazumder, R. (2020). Fast Best Subset Selection: Coordinate Descent and Local Combinatorial Optimization Algorithms, *Operations Research*, 68(5).

[38] Heiler, P., & Mareckova, J. (2021). Shrinkage for categorical regressors. *Journal of Econometrics*, 223(1), 161-189.

[39] Huang, S., Li, J., Chen, K., Wu, T., Ye, J., Wu, X., & Yao, L. (2012). A transfer learning approach for network modeling. *IIE Transactions*, 44(11), 915-931.

[40] Huang, W., Shu, L., Woodall, W. H., & Tsui, K. L. (2016). CUSUM procedures with probability control limits for monitoring processes with variable sample sizes. *IIE Transactions*, 48(8), 759-771.

[41] Hui, Z. and Trevor H. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society*. Series B (statistical Methodology). Wiley. 67(2): 301–20.

[42] Hutchinson, P. (2020). Reinventing Innovation Management: The Impact of Self-Innovating Artificial Intelligence. *IEEE Transactions on Engineering Management*, 68(2), 628-639.

[43] Hutter, F. (2009). *Automated configuration of algorithms for solving hard computational problems* (Doctoral dissertation, University of British Columbia).

[44] Igami, M. (2020). Artificial intelligence as structural estimation: Deep Blue, Bonanza, and AlphaGo. *The Econometrics Journal*, 23(3), S1-S24.

[45] Instapage (2024). Retrieved from https://instapage.com/en/template/pitch-page-template-for-cosmetic-stores

[46] Iskhakov, F., Rust, J., & Schjerning, B. (2020). Machine learning and structural econometrics: contrasts and synergies. *The Econometrics Journal*, 23(3), S81-S124.

[47] Jacobson, S. H., Kobza, J. E., & Easterling, A. S. (2001). A detection theoretic approach to modeling aviation security problems using the knapsack problem. *IIE Transactions*, 33(9), 747-759.

[48] Jardim, F. S., Chakraborti, S., & Epprecht, E. K. (2019). Chart with estimated parameters: the conditional ARL distribution and new insights. *Production and Operations Management*, 28(6), 1545-1557.

[49] Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., & Lindén, K. (2019). Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65, 675-782.

[50] Khalaf, G. and Shukur, G. (2005). Choosing Ridge Parameter for Regression Problems. *Communications in Statistics – Theory and Methods*. 34 (5): 1177–1182.

[51] Knaus, M. C., Lechner, M., & Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*, 24(1), 134-161.

[52] Lever, J., Krzywinski, M. & Altman, N. (2016). Regularization. *Nature Methods* 13, 803–804.

[53] Li, J., Netessine, S., & Koulayev, S. (2018). Price to compete… with many: how to identify price competition in high-dimensional space. *Management Science*, 64(9), 4118-4136.

[54] Lin, J., & Wu, X. (2017). A sequential test for the specification of predictive densities. *The Econometrics Journal*, 20(2), 190-220.

[55] Lyell D, Coiera E, Chen J, Shah P, Magrabi F. (2021). How machine learning is embedded to support clinician decision making: an analysis of FDA-approved medical devices. *BMJ Health & Care Informatics*. 28(1).

[56] Makai, A. (2019). Toward principled regularization of deep networks—From weight decay to feature contraction. *Science Robotics*, 4(30).

[57] Marvasti, A.E., Marvasti, E E., Bagci, U., & Foroosh, H. (2021). Maximum Probability Theorem: A Framework for Probabilistic Machine Learning, *IEEE Transactions on Artificial Intelligence*, 2(3), 214-227.

[58] Matteucci, G. & Zoccolan, D. (2020). Unsupervised experience with temporal continuity of the visual environment is causally involved in the development of V1 complex cells. *Science Advances* 6(22).

[59] Mardani, R.A. (2020). Bayesian Hyper-Parameter Optimization: Neural Networks, TensorFlow, Facies Prediction Example. *Towards Data Science*.

[60] Mišić, V.V. (2020). Optimization of Tree Ensembles. *Operations Research*. 68(5).

[61] Narasimhan, K., Barzilay, R., & Jaakkola, T. (2018). Grounding language for transfer in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 63, 849-874.

[62] Ning, X., & Tsung, F. (2013). Improved design of kernel distance–based charts using support vector methods. *IIE transactions*, 45(4), 464-476.

[63] Paleja, R., Ghuy, M., Ranawaka Arachchige, N., Jensen, R., & Gombolay, M. (2021). The utility of explainable ai in ad hoc human-machine teaming. Advances in neural information processing systems, 34, 610-623.

[64] Park, C., Lee, J., & Kim, Y. (2004). Economic design of a variable sampling rate EWMA chart. *IIE Transactions*, 36(5), 387-399.

[65] Poggio, T., Liao, Q. & Banburski, A. (2020). Complexity control by gradient descent in deep networks. *Nature Communcations.* 11, 1027.

[66] Ransbotham, S. (2014). Analytical Value From Data That Cries Wolf, *MIT Sloan Management Review*.

[67] Ransbotham, S. (2017). Justifying Human Involvement in the AI Decision-Making Loop, *MIT Sloan Management Review*.

[68] Rathje, J. M., & Katila, R. (2021). Enabling technologies and the role of private firms: A machine learning matching analysis. *Strategy Science*, 6(1), 5-21.

[69] Reeb, D.M. & Zhao, W. (2020). Disregarding the Shoulders of the Giants: Evidence from Innovation Research. *Journal of Economic Literature*.

[70] Reynolds, M. R., & Stoumbos, Z. G. (1998). The SPRT chart for monitoring a proportion. *IIE Transactions*, 30(6), 545-561.

[71] Scaillet, O. (2016). On ill-posedness of nonparametric instrumental variable regression with convexity constraints. *The Econometrics Journal*, 19(2), 232-236.

[72] Safaei, N. and Jardine, K.S. (2018). Aircraft routing with generalized maintenance constraints. *Omega*, 80, 111-122.

[73] Safaei, N. (2019). Premature Aircraft Maintenance: A Matter of Cost or Risk? *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 1-11.

[74] Sanders, N., & Wood, J. (2021). Combining Humans and Machines in an Emerging Form of Enterprise: the Humachine. Foresight: *The International Journal of Applied Forecasting*, (61).

[75] Saratiano, A. (2021). Europe Proposes Strict Rules for Artificial Intelligence, *The New York Times*.

[76] Shapley, L. S. (1953). *A value for n-person games.* Contributions to the Theory of Games, 2(28), 307-317.

[77] Sun, Y., & Yang, J. (2020). Testing-optimal kernel choice in HAR inference. *Journal of Econometrics*, 219(1), 123-136.

[78] Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, 61(4), 15-42.

[79] Tang, L. C., & Cheong, W. T. (2004). Cumulative conformance count chart with sequentially updated parameters. *IIE Transactions*, 36(9), 841-853.

[80] Tidhar, R., & Eisenhardt, K. M. (2020). Get rich or die trying… finding revenue model fit using machine learning and multiple cases. *Strategic Management Journal*, 41(7), 1245-1273.

[81] Tretyakov, K. (2004). Machine learning techniques in spam filtering. In *Data Mining Problem-oriented Seminar, MTAT*, 3(177), pp. 60-79.

[82] Villasevil, M. T., Pamies, M. B. I., Wang, Z., Desai, S., Chen, T., Agrawal, P., & Gupta, A. (2023, November). Breadcrumbs to the Goal: Goal-Conditioned Exploration from Human-in-the-Loop Feedback. In Thirty-seventh Conference on Neural Information Processing Systems.

[83] Vu, V.M., Bibal, A., & Frénay, B. (2021). Constraint Preserving Score for Automatic Hyperparameter Tuning of Dimensionality Reduction Methods for Visualization, *IEEE Transactions on Artificial Intelligence*, 2(3), 269-282.

[84] Wang, D., Chen, B., & Chen, J. (2019). Credit card fraud detection strategies with consumer incentives. *Omega*, 88, 179-195.

[85] Woods, D. (2012). The Man-Machine Framework: How to Build Machine-Learning Applications the Right Way, *Forbes*.

[86] Xu, H., Caramanis, C. and Mannor, S. (2016). Statistical Optimization in High Dimensions, *Operations Research*, 64(4), 958 – 979.

[87] Yang, J. C., Chuang, H. C., & Kuan, C. M. (2020). Double machine learning with gradient boosting and its application to the Big N audit quality effect. *Journal of Econometrics*, 216(1), 268-283.

[88] Yuan, Y., Zhou, S., Sievenpiper, C., Mannar, K., & Zheng, Y. (2011). Event log modeling and analysis for system failure prediction. *IIE Transactions*, 43(9), 647-660.

[89] Zhang, C. W., Xie, M., & Goh, T. N. (2006). Design of exponential control charts using a sequential sampling scheme. *IIE Transactions*, 38(12), 1105-1116.

[90] Zhang, H., Li, S. J., Zhang, H., Yang, Z. Y., Ren, Y. Q., Xia, L. Y., & Liang, Y. (2020). Meta-Analysis Based on Nonconvex Regularization. *Scientific Reports*, 10(1), 1-16.

[91] Zhang, N., Kannan, K., & Shanthikumar, G. (2021). Nudging a Slow-Moving High-Margin Product in a Supply Chain with Constrained Capacity. *Production and Operations Management*, 30(1), 11-27.

[92] Zhang, M., Zhang, Y., & Fu, G. (2018). Transition-based neural word segmentation using word-level features. *Journal of Artificial Intelligence Research*, 63, 923-953.

[93] Zou, C., Tseng, S. T., & Wang, Z. (2014). Outlier detection in general profiles using penalized regression method. *IIE Transactions*, 46(2), 106-117.

[94] Chung, Y-A., Yang S-W., & Lin, H-T (2020), Cost-Sensitive Deep Learning with Layer-Wise Cost Estimation. *IEEE - International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, Taipei, Taiwan.

[95] Yanka Aleksandrova; Mariya Armianova, Evaluation of cost-sensitive machine learning methods for default credit prediction. *IEEE - 2022 International Conference Automatics and Informatics (ICAI)*, Varna, Bulgaria.

Appendix 1: When Do Human-Machine Systems Outperform Humans?

Artificial intelligence may be construed as a "social and cognitive phenomena that enable a machine to socially integrate with a society to perform competitive tasks requiring cognitive processes" [1]. This aspiration for integration has given rise to the popularity of human-machine interactions. Our proposed approach to assessing the usefulness of ML structural estimation techniques and general business applications has important implications for assessing the performance of human-machine systems—aka, human-in-the-loop or weak supervision systems [22]. In these systems, a ML model provides recommendations and a human chooses whether to go along with the recommendation or not [85]. For example, consider a ML-based predictive or proactive maintenance application in which a ML model—such as a supervised model SVM, logistic regression, deep model, or XGBoost—is employed. This ML application requires inputs such as vibration sensor data, environmental/climate information, maintenance data, system health condition, and historical failures, and subsequently predicts the failure risk in a binary or probabilistic form.

In this example, and in many similar others, because of the high risk of the false alarms, the ML application alone is often not trusted to detect the potential faults of the mechanical equipment, but instead is designed to assist and complement humans' expertise [22]. In other words, a human-machine combination is used to make the decisions, such that the predictions made by ML will be reviewed by expert humans for final decisions such as continuation, stoppage, maintenance, or replacement. The role of ML in this example is to remove the scenarios with low failure risks from human intervention and flag the scenarios with high failure risks for human consideration. However, depending on the ML model's accuracy, some high-risk scenarios may be missed due to false negatives.

In this context, a commonly held assumption in the field is that human-machine systems outperform humans. For instance, research recently argues for an intuitive reaction that human-machine systems, what they coin as "Humachines," can make use of "the strengths of humans and machines in a symbiotic relationship, which can achieve a "superintelligence" that outstrips performance achieved by either humans or machines alone" [74]. However, we argue that a systematic consideration and quantification of false alarms in human-machine systems may challenge this widely held assumption and in doing so offer important implications for assessing the performance and economic value of ML techniques and applications.

To formally evaluate the economic usefulness of a mixed human-machine system, suppose that the performance of a human can be mapped into a confusion matrix similar to a binary classification model, as Figure 4 shows. In maintenance and reliability use cases, for example, this confusion matrix can be extracted from the historical failures and the computerized maintenance management system (CMMS) [73].

1.1 A Real-World Application

Similar to the application provided in Section 5, consider α^{M}, β^{M}, α^{H} and β^{H} with the same definitions associated to the entities in the ML's and human's confusion matrixes. The only difference is the interpretation of false negatives. In this application, β^{H} and β^{M} are the probability of missing the risky scenarios by human or machine.

In this context, the probability of the human error is $E^{\mathrm{H}} = \alpha^{\mathrm{H}} + \beta^{\mathrm{H}}$ and the probability of a human-machine error is $E^{\mathrm{HM}} = \alpha^{\mathrm{M}}\alpha^{\mathrm{H}} + (1 - \beta^{\mathrm{M}})\beta^{\mathrm{H}} + \beta^{\mathrm{M}}$. The latter probability of a human-machine error, E^{HM}, is calculated based on the decision tree depicted in Figure 5 to calculate the confusion matrix of a human-machine system. The key assumption is that the human expert is not biased to or influenced by the machine; that is, the human expert evaluates the machine flagged (i.e., true positives or false negatives) scenarios regardless of their falsity or truth because the machine cannot explain why such predictions are made (i.e., the machine explainability is low). Of course, a potential avenue for extension of our work involves examining the implications of cases where the machine provides some level of explainability, for example, through Shapley values [76] or causality analysis, such that human expert could be biased to or influenced by the machine.

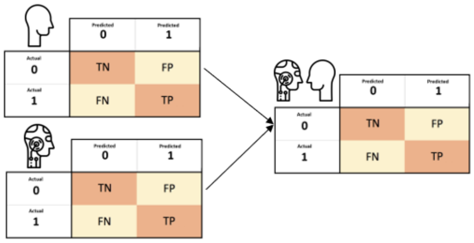

Figure 4: Human-machine performance evaluation in terms of the confusion matrix

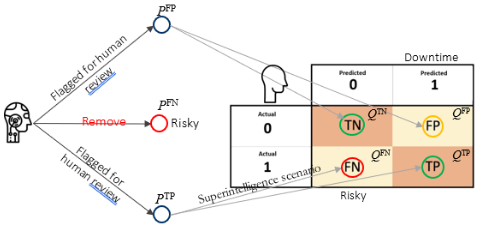

Figure 5: Human-machine confusion matrix based on machine to human decision tree

Our proposed approach on the quantification of the false alarms indicates that the commonly held assumption which asserts human-machine systems universally outperform humans in terms of the risk, that is, $E^{\mathrm{HM}} < E^{\mathrm{H}}$ [74] may not always be true. Indeed, human-machine systems only outperform humans if the expected cost or risk associated with human-machine systems is lower than that of humans. After rewriting E^{H} and E^{HM} and adding the cost/risk coefficients, human-machine systems only outperform humans if:

$$B\alpha^{\mathrm{H}} + A\beta^{\mathrm{H}} > B\alpha^{\mathrm{M}}\alpha^{\mathrm{H}} + A(1-\beta^{\mathrm{M}})\beta^{\mathrm{H}} + A\beta^{\mathrm{M}},$$

or equivalently:

$$\boldsymbol{\tau} = \frac{A}{B} < \frac{\alpha^{\mathrm{H}}-(\alpha^{\mathrm{M}}\alpha^{\mathrm{H}})}{(\beta^{\mathrm{M}}-\beta^{\mathrm{H}})+(1-\beta^{\mathrm{M}})\beta^{\mathrm{H}}} \qquad (5)$$

where B is the opportunity cost of a false positive—e.g., unnecessary equipment downtime in reliability and maintenance fields—and A is the cost of a false negative in the form of ignored hazardous scenarios that may result in catastrophic or secondary failure in such applications. $\boldsymbol{\tau} = \frac{A}{B}$ represents the *Risk Factor* of the associated use case—the ratio of the cost of a FN to the cost of a FP. In the current ML field – except a few research works focusing on cost-sensitive learning allowing the model to be aware of costs [94, 95] – an untested assumption is that $\boldsymbol{\tau} = \mathbf{1}$. However, for high-risk applications, $\boldsymbol{\tau}$ could be large or even approach infinity.

For numerical illustration, Inequality (5) is investigated under two scenarios: first, Conventional ML practice where there is no difference between risk/cost of FN versus FP (Risk Factor = 1), and, second, High-risk applications where the risk/cost of an FN is much higher than the risk/cost of an FP., e.g., Risk Factor = 10. These two scenarios are simulated in Figures 6a and 6b with parameter settings $\vartheta^{M} = \vartheta^{H} = 0.8$, $\alpha^{M} \in (0.01, 1-\vartheta^{M}]$, $\alpha^{H} \in (0.01, 1-\vartheta^{H}]$, $\beta^{M} \in (0.01, 1-\alpha^{\mathrm{M}}-\vartheta^{\mathrm{M}}]$ and $\beta^{\mathrm{H}} \in (0.01, 1-\alpha^{\mathrm{H}}-\vartheta^{\mathrm{H}}]$.

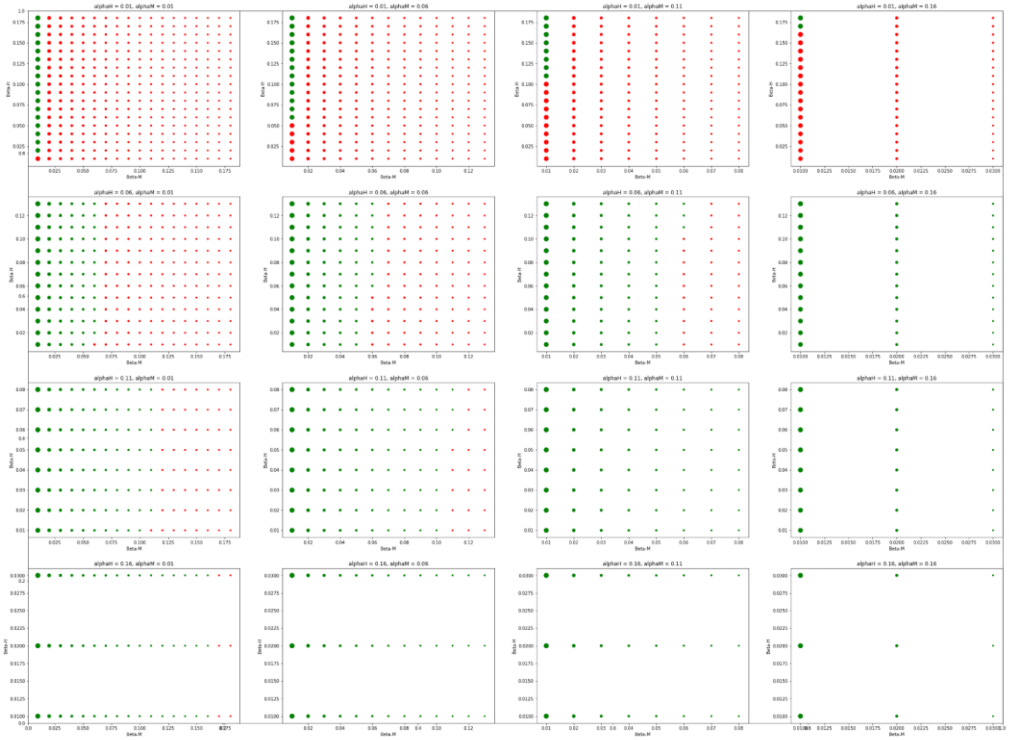

Figure 6a: Human-machine justifiability under ML conventional practice (*Risk Factor* = 1)

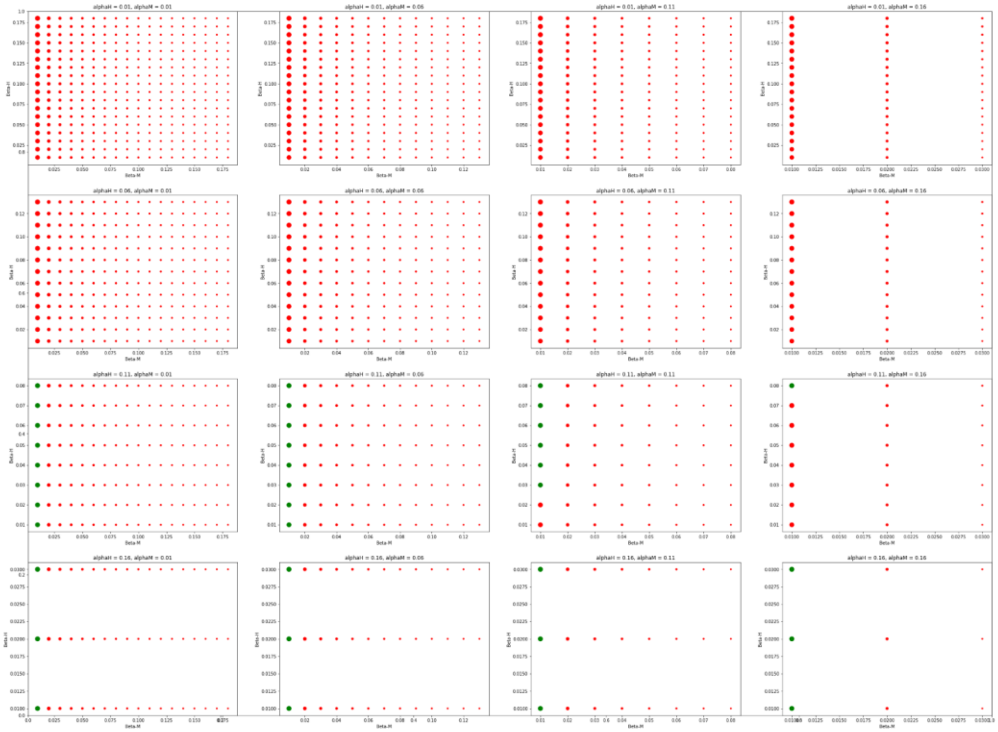

Figure 6b: Human-machine justifiability in High-risk applications (e.g., *Risk Factor* = 10)

1.2 Conceptual Interpretation

The term $(1-\beta^{\mathrm{M}})\beta^{\mathrm{H}}$ in Inequality (5) is the probability of a human-machine false negative (*superintelligence)* scenario where the machine can truly detect a positive scenario but the human cannot. One potential reason for this scenario would be the lack of explain-ability by ML model. Likewise, the term $\alpha^{\mathrm{M}}\alpha^{\mathrm{H}}$ is the probability of human-machine false positive (*ignorance)* scenario where neither of the machine or the human can truly detect a negative scenario. , Since $\boldsymbol{\tau}$ cannot be negative, we should have $\beta^{\mathrm{H}} < \beta^{\mathrm{M}} + (1-\beta^{\mathrm{M}})\beta^{\mathrm{H}}$ in the denominator where $\beta^{\mathrm{M}} + (1-\beta^{\mathrm{M}})\beta^{\mathrm{H}}$ is the machine's *surprise* probability—i.e., the probability that either the machine cannot detect a positive scenario or it outperforms the human. With these definitions, Inequality (5) can be rewritten as:

$$\boldsymbol{\tau} < \frac{\alpha^{\mathrm{H}} - \Pr(Ignorance)}{(\beta^{\mathrm{M}} - \beta^{\mathrm{H}}) + \Pr(Superintelligence)} \equiv \frac{\alpha^{\mathrm{H}} - \Pr(Ignorance)}{\Pr(Surprise) - \beta^{\mathrm{H}}}. \qquad (6)$$

Without the loss of generality, suppose that the probabilities of human false alarms (i.e., false positives and false negatives) are almost equal ($\alpha^{\mathrm{H}} \approx \beta^{\mathrm{H}}$)—henceforth, the probability of human error. Likewise, suppose that ($\alpha^{\mathrm{M}} \approx \beta^{\mathrm{M}}$)—henceforth, the probability of machine error. To held Inequality (6) and since $\boldsymbol{\tau}$ cannot be negative, then we should have:

$$\Pr(Ignorance) \leq \Pr(Human\ error) \leq \Pr(Surprise), \quad (7.1)$$

and

$$\Pr(Human\ error) - \Pr(Machine\ error) \leq \Pr(Superintelligence); \qquad (7.2)$$

Otherwise, Inequality (6) will flip and thereby human-machine systems will not outperform humans. By considering the costs and risks of false alarms which is directly tied to tuning the regularization hyper-parameters, Inequality (7.1) states that ML is practically useful and that human-machine systems outperform humans if the human error rate is greater than human-machine ignorance rate but smaller than the machine's surprise rate. When these conditions are met, "defections" from machine recommendations would represent an irrationality because "if we have two strategies for solving a problem and one is more reliable, it is folly to use the less reliable strategy to correct the more reliable one" [11]. At the same time, Inequality (7.2) imposes that ML is practically useless if the delta between human and machine error rates is greater than the machine superintelligence rate.