An Empirical Study on Automation Transparency (i.e., seeing-into) of an Automated Decision Aid System for Condition-based Maintenance

Fahimeh Rajabiyazdi, Greg A. Jamieson and David A. Quispe G.

May 28, 2021

An Empirical Study on Automation Transparency (i.e., seeing-into) of an Automated Decision Aid System for Condition-based Maintenance

Fahimeh Rajabiyazdi [1], Greg A. Jamieson[1], and David A. Quispe G. [1]

[1] University of Toronto, Department of Mechanical and Industrial Engineering, M5S 3G8

Abstract. Prior studies have shown conflicting results about the impact of information disclosure on human performance– often referred to as transparency (i.e., seeing-into) studies. We conducted an experiment to investigate whether transparency manipulations predicted whether participants could identify whether features and their relative weights of a decision aid guided by a Machine Learning model were consistent with stated best practices for making maintenance decisions. We had insignificant results on state estimation, automation reliance, trust, workload, and self-confidence. This study shows that disclosing information about the decision aid rationale does not necessarily impact operator performance.

Keywords: Automation Transparency, Decision aids, Maintenance

1 Introduction

In Machine Learning (ML)-based decision aid systems, human oversight may be required to check that the ML rationale aligns with end-user goals and metrics. Furthermore, the end-user may need to verify that the training and validation data are representative of real-world conditions. Thus, the ML rationale may have to be disclosed to the end-user. However, the effective presentation of this rationale for these end-user tasks is still an ongoing research question.

Doshi-Velez and Kim [1] distinguished between *local* and *global* explanations to end-users for ML algorithms. A global explanation is one that offers information on the logic of an ML algorithm as a whole. A local explanation discloses the logic of an ML algorithm that led to a specific decision. Human Factors researchers have evaluated the impacts of information disclosure about automation logic on end-user performance under the notion of *transparency*. For instance, Seong and Bisantz [2] and Mercado et al. [3] reported that disclosing information about automation had a positive impact on human task performance and trust calibration. In contrast, Adhikari et al. [4] reported that participants objectively performed the worst when presented with any amount and type of information. However, participants self-reported a better understanding of the ML-based Decision Support System rationale with greater information disclosure. Similarly, Skraaning and Jamieson [5] reported that participants performed worst but calibrated trust with information disclosure as automation capabilities increased. Given these inconsistent and, at times, conflicting results, there is a need to conduct more empirical studies on transparency to establish the effective type and amount of information disclosure that positively impacts human performance.

As suggested by [1], we anticipate that local explanations will support participants to assess the correctness of a specific decision made by the ML algorithm. Thus, our research question is: What are the effects of disclosing the rationale that led to an automated decision through Feature Weight (also known as Feature Importance) and the Decision Rules on human performance (including reliance decisions, trust, task efficacy, and workload)?

2 Method

2.1 Participants

We recruited 24 (14 female, 10 male) chemical engineering undergraduate and graduate students from the University of Toronto who had completed courses in process engineering and statistics. Participants were between the ages of 18 and 30 (*M* = 25, *SD* = 3.17). Ten participants indicated prior work experience in the process operation industry (*M* = 23 months, *SD* = 25). Eleven participants stated moderate familiarity with ML (i.e., had completed ML courses or self-taught ML concepts). Participants were paid $15/hour rounded to the nearest 20 minutes. To incentivize participants, they were entered into a draw for an extra $25 after study completion. To motivate participants to follow the instructions, we told them to imagine that a company hired them as a consultant to perform this task.

2.2 Apparatus

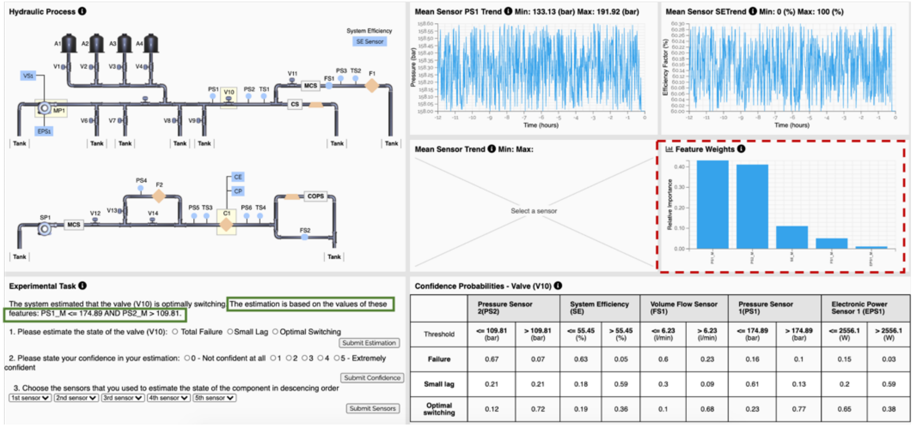
A machine learning-based micro-world platform for condition-based maintenance named Automated Reliability Decision Aid System (ARDAS) was used for this experiment [6]. The architecture of ARDAS comprises the ML algorithm and the user interface. In ARDAS, a supervised ML algorithm is trained to generate models that predict the states of four hydraulic components. The models were trained using multi-sensor time-series and historical event data to classify sensor measurements [7].

The original ML model in ARDAS used a random forest algorithm trained with sixty-eight features and composed of thousands of trees. These features are extracted from the statistical moments (including mean, kurtosis, skewness, and variance) of seventeen sensors data. Due to the complexity of the ML model, it was necessary to reduce the number of features to employ this ML model while fulfilling the constraints of a controlled experiment. The constraint for our experiment was that participants must be able to complete the experimental task (i.e., estimate the hydraulic component's state) in every experimental condition while avoiding learning effects in trials.

To meet this constraint, we simplified the model. First, we obtained the global feature weights of each components using sixty-eight features. Then, we selected five features for each component imposing the following constraints: First, select features that deemed most influential by the original ML model. Second, select only features that represent the mean of the sensors data. Third, select only features that adhered to process engineering principles about the hydraulic system.

This process resulted in five different global influential features in determining each hydraulic component's states. At this stage we employed Wizard of Oz technique [8]. We used the same five features specific for each component in each trial but assigned a unique local weight to each feature. Using engineering principles, the feature's weights were calculated based on how far the value of the feature is into the threshold as a function of the threshold's size. Finally, we estimated a component's state based on the weighted average of the estimated probabilities. These probabilities were guided by the thresholds of the different decision trees within the ML model.

The user interface of ARDAS (Fig.1) included a hydraulic process diagram (top-left), three mean sensor data graphs (top-right), and a confidence probabilities table (bottom-right). Depending on the experimental condition, the user interface also included the Local Feature Weights (Fig.1, dashed red box), the Decision Rules (Fig.1, solid green box), or both. The Feature Weights and Decision Rules presented the automation rationale that led to a particular state estimation.



Experimental Task

The system estimated that the valve (V10) is optimally switching. The estimation is based on the values of these features: PS1_M <= 174.89 AND PS2_M > 109.81.

1. Please estimate the state of the valve (V10): ○ Total Failure ○ Small Lag ○ Optimal Switching

2. Please state your confidence in your estimation: ○ 0 - Not confident at all ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 - Extremely confident

3. Choose the sensors that you used to estimate the state of the component in descencing order

Confidence Probabilities - Valve (V10)

	Pressure Sensor 2(PS2)		System Efficiency (SE)		Volume Flow Sensor (FS1)		Pressure Sensor 1(PS1)		Electronic Power Sensor 1 (EPS1)	
Threshold	<= 109.81 (bar)	> 109.81 (bar)	<= 55.45 (%)	> 55.45 (%)	<= 6.23 (l/min)	> 6.23 (l/min)	<= 174.89 (bar)	> 174.89 (bar)	<= 2556.1 (W)	> 2556.1 (W)
Failure	0.67	0.07	0.63	0.05	0.6	0.23	0.16	0.1	0.15	0.03
Small lag	0.21	0.21	0.18	0.59	0.3	0.09	0.61	0.13	0.2	0.59
Optimal switching	0.12	0.72	0.19	0.36	0.1	0.68	0.23	0.77	0.65	0.38

Figure 1: The user interface of ARDAS

Hydraulic Process Diagram

The top left section in Fig. 1 shows the hydraulic process diagram that includes the positions of the components in the yellow rectangles (Valve (V10), Pump (MP1), Cooler (C1)), and seventeen sensors in blue squares and circles. Each hydraulic component has three states (normal functioning, minor malfunctioning, failure).

Mean Sensor Value Graphs

The top right section in Fig.1 displays up to three-line graphs of mean sensor values over a 12-hour period. The minimum and maximum mean sensor values are written at the top of each graph. If a fourth sensor is selected, the earliest line graph will be replaced by the newly selected sensor's mean graph. In this experiment, the point of interest is the sensor value at the current time.

Confidence Probabilities Table

The bottom right section in Fig.1 shows a confidence probabilities table. For each component, probabilities are shown for each state, given the thresholds for each feature. Each feature's value is compared against a threshold for determining each state's probability for given component.

2.3 Experimental design

A 2x2 within-subject design was employed, with Local Feature Weights (displayed, not displayed) and Decision Rules (displayed, not displayed) as factors. The result is four levels of Display: none [baseline], Local Feature Weights, Decision Rules, and Combined (Local Feature Weights and Decision Rules). Each level was presented in a separate block of twelve independent trials. Each participant completed all four blocks.

We used the method proposed by Zeelenberg and Pecher [9] to counterbalance the order of the four experimental conditions, and the assignment of trial sets 1 – 4 to these conditions. This method systematically identifies 8 block-condition orders (a pair of Latin squares) that balance the immediate and remote sequence effects. The order of trials within a set was not randomized.

We created a total of 48 independent trial stimuli by selecting three unique hydraulic components (Valve (V10), Pump (MP1), Cooler (C1)), each with three states. The state of each hydraulic component was generated based on five features. Each feature was assigned a unique value in each trial. Furthermore, in the conditions wherein the feature weights were disclosed, the feature weights and their orders were randomized. The 48 stimuli were divided into four sets of 12 trials, labelled 1 – 4, each having the three hydraulic components appear four times but with unique values for each feature.

The automated decisions were manipulated to produce incorrect estimates in 8 out of 48 trials (83.3% reliability) We challenged participants to identify whether the decision aid's features, and their relative weights were consistent with the company's best practices. Thus, we simulated incorrect automated decisions due to incorrect features (4 trials) and incorrect weights (4 trials). There were two incorrect trials in each block. The order of these eight incorrect trials was randomized throughout the blocks.

The company's best practice was to estimate the state of each component using the five designated features. If the decision aid system used any other features, then its decision should deem incorrect (i.e., incorrect features error trial). Furthermore, there is a chance of noisy measurement readings in virtual sensors (i.e., cooling efficiency and system efficiency) since their values are computed based on multiple sensors. Thus, the company practice is that if these virtual sensors' feature values are at or within ± 0.5% higher or lower than the threshold, the least weight should be associated with it. If that is not the case, then the automated decision is incorrect (i.e., incorrect weights error trial).

2.4 Experimental Task

Participants estimated the state of a hydraulic component given three possible states. In all trials, the state estimation for that component was given. To estimate the state of a hydraulic component and assess whether they agreed with the decision aid system's state estimation, the participants were expected to compare the automated estimation against their domain knowledge. However, participants' domain knowledge may differ. To minimize the effect of prior domain knowledge on subjects' assessment of the automation rationale, were trained participants on what information to incorporate in their decisions. Participants were trained to complete the task in the following steps: First, determine the value for each of the five designated features at the current time using the mean sensor value graphs. Second, compare the values against the designated threshold given in the confidence probabilities table to determine each state's probability. Finally, calculate the average of the probabilities for each of the three states. The state with the highest probability is the company's desired estimation. Participants were told to use ML's rationale as appropriate.

Transparency conditions

Local Feature Weight Graph

The feature weights represent the amount that each feature contributed to the final prediction. The local feature weight graph (Fig.1, dashed red box) presents the weights of the five most prominent features used to estimate the hydraulic component state. In each trial, each feature of a component is assigned a weight by the automation. The sum of the five weights is one.

Decision Rules

The decision rules present the IF-THEN rules (solid green box, Fig.1) guided by the decision trees within the ML algorithm to estimate the hydraulic component condition. The decision rule condition statement included the features weighted more than or equal to 0.2 and the threshold that the feature value was compared against.

Combined (Local Feature Weight Graph + Decision Rules)

In this condition, both the local feature weight and the decision rules were presented.

2.5 Procedure

The experiment was conducted online with the experimenter present. The protocol was executed over two consecutive days, and on average, participants took 4 and a half hours to complete the study. At the start of the first day, participants signed a consent form and completed a demographic questionnaire. They then watched a video explaining the platform and experiment task. Afterwards, participants were given three practice trials where they were encouraged to think-aloud before proceeding with the experiment.

In each trial, after submitting their state estimation, participants were asked to state their confidence on a scale of whole numbers between 0 and 5 inclusively with 0 = "not confident at all" and 5 = "extremely confident". After each trial, the participants were asked to indicate the sensor data that they used to arrive at a state estimation for the component (in descending order of influence). This question was asked to ensure that participants were not randomly choosing a state but rather were using the sensors to estimate the components' states. After each block, participants rated their mental workload on the NASA-TLX questionnaire modified to a seven-point scale [10]. They rated their trust in the automated estimation on a modified trust questionnaire designed by [11]. After the final block, participants completed the relative weighting portion of the NASA-TLX questionnaire as it applied to all the experimental tasks throughout the blocks. Finally, we conducted a semi-structured interview asking participants about their experience using the platform and their strategies in estimating component states.

3 Results

For the analysis below, trial data was aggregated by block (N = 96). State estimation, correct automation usage, correct automation rejection, were treated as proportion data. Respectively, that is the number of correct estimations out of 12 trials, the number of correct automation usage event out of 10 trials, and the number of correct automation rejections out of 2 error trials in each block.

We used the *glmer*() function from the lme4 package to build a generalized linear mixed model with a binomial distribution for state estimation, correct automation usage, and correct automation rejection. First, we built a baseline model from only the intercept. Then, we added Display as a predictor to our model. We specified a random part to our model. The random effect was specified as the Display nested within participant ID to account for our data dependency. For each of these dependent variables, we compared the baseline to the main model with the Display predictor.

3.1 State Estimation

The state estimation is categorical with either correct or incorrect estimation. The correct estimation is the most probable state among three states. Display was not a significant predictor of correctly estimating a state, $X^2(3) = 5.27$, Pr (> *Chisq*) = 0.15. Non-orthogonal contrasts revealed that state estimations were not significantly more correct for Feature Weight compared to Baseline, B(*SE*) = 0.42(0.22), z = 1.91, *p* = 0.06, odds ratio = 1.52, or between Decision Rules or Baseline, B(*SE*) = -0.03(0.21), z = -0.16, *p* = 0.88, odds ratio = 0.96, or between Combined or Baseline, B(*SE*) = 0.21(0.22), z = 0.97, *p* = 0.33, odds ratio = 1.23.

3.2 Correct Automation Usage

The correct automation usage is categorical with two categories of yes or no. Display was not a significant predictor of correctly using automation, $X^2(3) = 5.90$, Pr (>

Chisq) = 0.12. Non-orthogonal contrasts revealed that automation usage was not significantly more correct for Feature Weight compared to Baseline, B(*SE*) = 0.47(0.26), z = 1.83, *p* = 0.07, odds ratio = 1.61, or between Decision Rules or Baseline, B(*SE*) = -0.12(0.24), z = -0.5, *p* = 0.60, odds ratio = 0.88, or between Combined or Baseline, B(*SE*) = 0.17(0.25), z = 0.70, *p* = 0.48, odds ratio = 0.17.

3.3 Correct Automation Rejection

The correct automation rejection is categorical with two categories of yes or no. Display was not a significant predictor of correctly rejecting automation, $X^2(3) = 0.88$, Pr (> *Chisq*) = 0.83. Non-orthogonal contrasts revealed that automation usage was not significantly more correct for Feature Weight compared to Baseline, B(*SE*) = 0.35(0.44), z = 0.80, *p* = 0.42, odds ratio = 1.43, or between Decision Rules or Baseline, B(*SE*) = 0.26(0.44), z = 0.60, *p* = 0.55, odds ratio = 1.30, or between Combined or Baseline, B(*SE*) = 0.35(0.44), z = 0.80, *p* = 0.42, odds ratio = 1.43.

3.4 Mean Confidence, Mean Trust, Workload, Mean Response Time

We calculated workload scores using the method advised by [3]. We scaled the workload scores and the first 10 questions in the modified trust questionnaire to 0 to 100 (inclusive) for the analysis.

According to Levene's test, the homogeneity of variance assumption was met for mean confidence F (3,92) = 0.11, Pr(>F) = 0.96, mean trust F (3,92) = 2.21, Pr(>F) = 0.09, and workload F (3,92) = 0.36, Pr(>F) = 0.78. According to Shapiro-Wilk test, mean confidence (*W* = 0.99, *p* = 0.50), mean trust (*W* = 0.97, *p* = 0.06), and workload (*W* = 0.98, *p* = 0.40) met the assumption of normality.

We conducted ezANOVA on these variables with orthogonal contrasts. Mauchly's test indicated that the assumption of sphericity had been met for mean confidence (*W* = 0.91, *p* = 0.85), mean trust (*W* = 0.72, *p* = 0.21), and workload (*W* = 0.73, *p* = 0.24). The results showed that mean confidence F (3,69) = 2.56, p = 0.66, η^2= 0.02, mean trust F (3,69) = 0.43, *p* = 0.73, η^2 = 0.01, and workload F (3,69) = 0.40, *p* = 0.75, η^2 = 0.01were not significantly affected by the type of Display.

According to Levene's test, the homogeneity of variance assumption was met for mean response time, F (3,92) = 0.13, Pr (>F) = 0.94. However, according to Shapiro-Wilk test, the assumption of normality has been violated for mean response time (*W* = 0.94, *p* < 0.05). Log transformation was used after which it met the assumption of normality (*W* = 0.99, *p* = 0.82). We conducted ezANOVA on mean response time. Mauchly's test indicated that the assumption of sphericity had been violated for mean response time (*W* = 0.33, *p* <.05), therefore, degrees of freedom were corrected using Huynh-Feld estimates of sphericity (ε = 0.74). The results showed that mean response time was not significantly affected by the type of display, F (2.21,50.85) = 0.67, *p* = 0.53, η^2 = 0.01.

4 Discussion and Conclusion

We found no evidence to corroborate the common belief that presenting a rationale for a decision aid's conclusion will positively impact automation reliance and efficacy. Disclosing information about the decision process in the form of local feature weights, decision rules or both combined did not predict performance on state estimation, automation reliance, trust or self-confidence, workload, or mean response time.

One of many limitations is the constraints that we imposed on the ML features (see 2.2). It is possible that, had we not restricted the ML features in the described manner, participants may have benefitted more from the rationale disclosure. However, it is also possible that the complex nature of ML models may not lend itself to controlled experimental traditions that characterize human factors and ergonomics research.

References

1. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017).
2. Seong, Y., Bisantz, A. M.: The impact of cognitive feedback on judgment performance and trust with decision aids. International Journal of Industrial Ergonomics 38(7-8), 608–625 (2008).
3. Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., Procci, K: Intelligent agent transparency in human-agent teaming for multi-UxV management. Human Factors 58(3), 401–415 (2016).
4. Adhikari, A., Tax, D. M.: LEAFAGE: Example-based and Feature importance-based Explanations for Black-box ML models. In: 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–7. IEEE (2019).
5. Skraaning, G., Jamieson, G. A.: Human Performance Benefits of The Automation Transparency Design Principle: Validation and Variation. Human Factors, 1–23 (2019).
6. Quispe, D., Rajabiyazdi, F., Jamieson, G. A.: A Machine Learning-Based Micro-World Platform for Condition-Based Maintenance. In: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 288-295. IEEE (2020).
7. Helwig, N., Pignanelli, E., Schutze, A.: Condition monitoring of a complex hydraulic system using multivariate statistics. In: IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings, pp. 210–215. IEEE (2015).
8. Stanton, N. A., Salmon, P. M., Rafferty, L. A., Walker, G. H., Baber, C., Jenkins, D. P.: Human Factors Methods. In Human Factors Methods. Ashgate Publishing, Ltd. (2013).
9. Zeelenberg, R., Pecher, D.: A method for simultaneously counterbalancing condition order and assignment of stimulus materials to conditions. Behavior research methods 47(1), 127–133 (2015).
10. Hart, S. G., Staveland, L. E.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Advances in psychology, 53, 139–183 (1988).
11. Jian, J. Y., Bisantz, A. M., Dury C. G.: Foundations for an empirically determined scale of trust in automated systems. International journal of cognitive ergonomics 4(1), 53–71(2000).