Int. J. Human–Computer Studies 196 (2025) 103434

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Human performance effects of combining counterfactual explanations with normative and contrastive explanations in supervised machine learning for automated decision assistance

Davide Gentile *, Birsen Donmez, Greg A. Jamieson

Department of Mechanical and Industrial Engineering, University of Toronto, 5 King's College Road, Toronto, ON M5S 3G8, Canada

ARTICLE INFO

Keywords:
Explainable AI
Automation transparency
Reliance
Example-based explanations
Counterfactuals

ABSTRACT

Counterfactual explanations have emerged as a popular solution for elucidating the reasons behind machine learning predictions due to their contribution in supporting people's understanding of causality. Despite psychological research suggesting potential burdens associated with counterfactuals, empirical data on the influence of counterfactual explanations on human decisions is limited, especially in comparison with other more traditional explanation methods in machine learning for decision assistance. We present an experiment to examine the human performance effects of counterfactual explanations combined with normative and contrastive explanations in the context of condition-based manteinance. Twenty-four participants provided their diagnosis of the conditions of a hydraulic system with the assistance of a simulated decision aid based on machine learning, under four experimental conditions (baseline with no explanations, normative plus contrastive explanations, normative plus counterfactual explanations, and normative plus contrastive plus counterfactual explanations). The results indicate a lack of significant performance differences between explanation conditions. However, we found a reduction in false alarm rate in the condition with all three explanations, and a potential reduction in decision time and workload in the two conditions that included counterfactual explanations. These findings highlight the potential of counterfactuals to reduce decision time and workload, but they also caution against overestimating their benefits in supporting decision performance within digital work environments.

1. Introduction

One of the primary challenges in automation design is determining the type of information that automated system should provide when advising end users, with the objective of promoting appropriate reliance behaviors (Riley, 2018). Traditionally, this research area falls under the umbrella of automation transparency (van de Merwe et al., 2022; Rajabiyazdi and Jamieson, 2020). However, the increasing integration of machine learning (ML) models in automated systems has given rise to "explainable artificial intelligence" (XAI) or "explainability" (Hoffman et al., 2018). Field-specific terminology and research scopes of transparency and explainability may differ (Warden et al., 2019). Nonetheless, when automation leveraging ML models is used to support human decisions, both design principles of transparency and explainability share the common objective of building systems that empower human operators to identify crucial information and intervene effectively, especially in the event of automation failure (Skraaning and Jamieson, 2023).

Researchers in both fields are increasingly highlighting how the design of automated systems should support the mitigation of risks such as unwarranted trust and reliance. This objective requires an interdisciplinary approach to understand and effectively manage the interaction between humans and algorithms, particularly given the growing societal implications of using algorithms for decision-making (Shin 2023; 2024). Such an interdisciplinary perspective can inform how transparency or explanation methods can support human decision-making and promote effective reliance on automated decision aids. This is particularly relevant in high-stakes contexts which require traceable and interpretable processes, such as healthcare (Holzinger et al., 2019) and nuclear power generation (Hall et al., 2024).

When automated decision aids leverage ML models, and such models are not inherently explainable (in other words, they are "black-boxes"), stakeholders often turn to post-hoc explanations to understand the reasons behind the automated advice (Kenny et al., 2021). Post-hoc

* Corresponding author.
E-mail address: dgentile@mie.utoronto.ca (D. Gentile).

https://doi.org/10.1016/j.ijhcs.2024.103434
Received 21 April 2024; Received in revised form 2 December 2024; Accepted 9 December 2024
Available online 10 December 2024

explanations elucidate the inner workings of systems *after the fact* (i.e., the prediction), and are meant to help users understand why a certain result is generated (or not). They can be global when they elucidate the inner workings of the whole model, or local, when they provide a rationale for a single prediction (Setzu et al., 2021). Among the many types of post-hoc local explanations (Liao et al., 2020), *counterfactual explanations* have gained traction in explainable AI. In the context of XAI, a counterfactual explanation of an ML prediction describes the smallest change to the input values that changes the prediction to an alternative output (Molnar, 2020).

The popularity of counterfactuals in XAI is often attributed to proposals in philosophy of science about their central role in causality (Lewis, 2013; Woodward, 2005), in psychology about their ability to elicit spontaneous causal thinking in people (Byrne, 2019; Legnado et al., 2013; Miller, 2019; Shin, 2021), and in the legal domain, where they are claimed to be GDPR compliant (Watcher et al., 2017). This popularity is reflected in the growing body of research on generating counterfactuals. Examples include generating counterfactual causal explanations for deep learning models (Del Ser et al., 2024), developing taxonomies for model-agnostic counterfactual approaches in XAI (Chou et al. 2024), computing counterfactual methods to help clinicians interpret automated results (Tanyel et al., 2023), and other efforts to generate *plausible* counterfactuals that aid human decision-making (Keane et al., 2023; Verma et al., 2021). However, research in psychology has also warned about potential burdens of counterfactual explanations, arguing that in complex situations the added cognitive work may impair rather than support human decisions (Kahneman and Miller, 1986; Byrne, 2007). Similarly, a recent review on theories and counterfactual approaches in XAI invited caution about psychological benefits of counterfactuals, stating that "*current model-agnostic counterfactual algorithms for explainable AI are not grounded on a causal theoretical formalism and, consequently, cannot promote causability to a human decision-maker.*" (Chou et al., 2024, p. 80). This initial evidence invites caution in adopting the notion that counterfactual explanations promote human causal thinking.

Human-subjects experiment can provide insights into the influence of counterfactual explanations by testing the effects of providing explanations on user performance in decision tasks, by comparison to either no-explanation baselines and/or other explanation methods. Other explanation methods typically included in user studies are example-based explanations presented as normative explanations and contrastive explanations. Example-based, normative explanations justify an automated result by showing examples describing "a norm" of what the input should be to fit a certain outcome (Cai et al., 2019; Gentile et al., 2023). Conversely, example-based, contrastive explanations explain why a prediction was made instead of another, showing counterexamples of the automated output, and answering questions such as "Why this advice instead of that advice?" (Lipton, 1990; Miller, 2019) or "why not that advice?" (Liao et al. 2020). Researchers in XAI have either separated contrastive explanations from counterfactual explanations (Stepin et al., 2021), or used the two terms intercheangeably (Shang et al., 2022). In this paper, we distinguish between contrastive and counterfactual thinking (McGill and Klein, 1993) and consider counterfactual explanations as a category of contrastive explanations (Stepin et al., 2021). The main motivation for this categorization is that counterfactuals can answer why-not questions as contrastive explanations, but they additionally highlight the minimum differences in the input that leads to changes from an output to another. According to this view, a more precise taxonomy would differentiate "generic contrastive" and "contrastive counterfactuals" (Stepin et al., 2021, p. 11,977). For simplicity, in the remainder of the paper, we will refer to contrastive explanations to mean "generic contrastive" explanations, and to counterfactual explanations to mean "contrastive counterfactual" explanations.

While the number of user experiments including these and other explanation methods has grown over the last few years (Keane et al., 2023; Leavitt, 2020), the available studies provide inconclusive evidence (Delaney et al., 2023). One of the first studies testing the effects of counterfactual on human performance is Lim et al. (2009). In that study, the participants were exposed to a series of input data related to an individual's physiological variables (e.g., body temperature, heart rate), the prediction of an ML model regarding whether the individual was exercising and had to select whether the individual was exercising or not. The explanations included a form of normative explanations (which they called "Why" explanations), a form of counterfactual explanations (which they called "Why-not" explanations), "How-to" explanations (showing the required inputs to derive a desired output), and "What if explanations" (showing a hypothetical output given a certain input), along with a baseline condition with no explanations (p. 2122, Table 1). They found that normative explanations resulted in better understanding, higher trust, and better task performance than counterfactuals. In contrast, counterfactuals resulted in better understanding and task performance than the "How-to" and "What-if" explanations. Further, Dodge et al. (2019) tested four explanation methods including normative (which they call "case-based"), counterfactual (which they call "sensitivity-based"), and two versions of global explanations to test how they impact people's judgement of an ML classifier predicting the risk of recidivism among a sample of criminal offenders (Grgic-Hlaca et al., 2018). They found that counterfactual explanations led to people judging the classifiers as less biased compared to the other explanations, suggesting that counterfactuals may have benefited participants' understanding of the model's results by directing their attention to the features relevant to a particular decision over another.

Other studies have found weaker evidence in support of the benefits of counterfactuals. For example, Warren et al. (2022; 2023a) used a simulated decision-making app that determined safe driving limits after drinking alcohol based on predicted blood alcohol content. They found that counterfactual explanations led to better task performance (i.e., whether the participant's prediction of the ML output aligned with the actual ML output) compared to no explanations, although not higher performance than causal explanations (describing the features determining whether an individual was driving within or above safe driving conditions). Additionally, counterfactual explanations resulted in higher satisfaction and trust than causal explanations.

Finally, some studies have also reported detrimental effects of counterfactuals on human performance. For example, Lage et al. (2019) had participants predict and verify the output of a ML model in a meal recommendation app and found that in tasks that involved explanations with counterfactual reasoning, participants not only showed lower performance, but also reported longer response times and perceived those tasks as being more difficult than the other tasks. Similarly, in Lucic et al. (2020) participants were given certain data inputs, along with ML generated outputs, and explanations related to those outputs. The researchers observed that participants' accuracy was lower when they were tasked with making changes to the input data (introducing a counterfactual change) compared to when they were simply asked to

Table 1
Confusion matrix with the four possible outcomes from Signal Detection Theory (SDT).

		State of the world	
		The ARDAS best estimation is incorrect (Signal; 18 % of cases)	The ARDAS best estimation is correct (Noise; 82 % of cases)
Did the user select the ARDAS second-best estimation?	**Yes**	Hit The user rejects a correct automated estimation.	False Alarm The user rejects an incorrect automated estimation.
	No	Miss The user accepts an incorrect automated estimation.	Correct Rejection The user accepts a correct automated estimation.

simulate the output based on the provided input data and explanations without altering the input.

A potential explanation for the inconsistent empirical evidence on counterfactuals could be that existing studies are paying limited attention to the link between explanation methods and the desired goal of the human-machine system (Delaney et al., 2023; Langer et al., 2021; Gentile et al., 2023). Providing users with the minimal change required in the input to elicit an alternative output may not be sufficient to support effective reliance on decision aids. Thus, counterfactuals may need to be put into context to understand when they may be beneficial, neutral, or detrimental to human performance. The literature presents initial efforts to contextualize the effects of counterfactual explanations. For example, a recent study presents a promising method to provide counterfactual explanations for groups of similar instances, instead of providing sperate counterfactuals for each local prediction (Warren et al., 2023b). One question currently not being addressed regards testing the human performance effects of providing different combinations of explanations for local predictions.

Our review of the literature leads to the research question of whether including counterfactual explanations in ML-based decision would support human performance. In a broader sense, we seek to determine whether post-hoc explanations including counterfactuals prove to be more effective in aiding decision-making processes than post-hoc explanations not including counterfactuals. The research question is the following: What is the impact of including counterfactual explanations on participants' performance metrics, such as task performance, workload, and decision time, in comparison to general contrastive explanations?

To answer this question, we present a human-subjects experiment where participants used an automated decision aid in the context of condition-based maintenance. Building on prior studies in counterfactuals (e.g., Lim et al., 2009; Dodge et al., 2019), we examine counterfactuals in comparison to different explanation methods, aiming to identify which one yields superior performance. Specifically, we explore the combined impact of counterfactuals with normative and contrastive explanations. We included normative explanations in all experimental conditions, aligning with the external validity approach adopted in Gentile et al. (2023) which used a similar experimental setup. In Gentile et al. (2023), we asserted that for most industrial use cases of automated decision aids, generic contrastive or counterfactual explanations should be provided not in isolation but as an addition to normative explanations. In these applications, the goal of the human-machine system is often to support people in making accurate decisions, within certain time limits, and within their cognitive workload capacity. Adopting the terminology from the recommender systems literature, these correspond to the explanation goals of *effectiveness* and *efficiency* (Tintarev and Mashtoff, 2012). In Gentile et al. (2023), we found that while normative explanations alone reduced participants' decision time and workload, the addition of contrastive explanations to normative ones further led to improvements in task performance, without impacting decision time and workload. The current experiment can be considered a follow-up to Gentile et al. (2023), where we simplified the experimental design to examine four experimental conditions (rather than three) and compare performance across different combinations of normative, contrastive, and counterfactual explanations. Given that in our previous study we found greater benefits in the condition with combined explanations, we excluded a normative-only condition from this study.

2. Methods

We conducted the experiment on a micro-world platform named Automated Reliability Decision Aid System (ARDAS), an open-source platform developed by Guanoluisa et al. (2020) to assist researchers in human-AI interaction to conduct user studies and develop design recommendations to support human performance in digital work environments (Fig. 1). ARDAS simulates decision processes within the context of condition-based maintenance (CBM), a preventive maintenance technique that utilizes real-time sensor data to determine the current (diagnosis) or future (prognosis) condition of a component in the equipment, and to recommend a maintenance action to ensure operational continuity (Jardine et al., 2006; Peng et al. 2010).

2.1. Participants

Twenty-four engineering students participated in the study (7 females, 17 males; mean age = 20.1 years, with standard deviation of 1.3 years). We targeted this user sample due to their accessibility and knowledge about engineering and data processes. We employed a rating scale ranging from 1 (indicating no experience or familiarity) to 7 (representing extensive experience or familiarity) to assess participants' proficiency in each of machine learning, experience in the process operation industry, and experience in maintenance processes related to hydraulic systems. Participants self-reported their experience scores as 2.75 for machine learning models, 1.7 for process operation, and 1.5 for hydraulic systems. The experiment was conducted online in the presence of an experimenter and took on average 2.2 h to complete. Participants

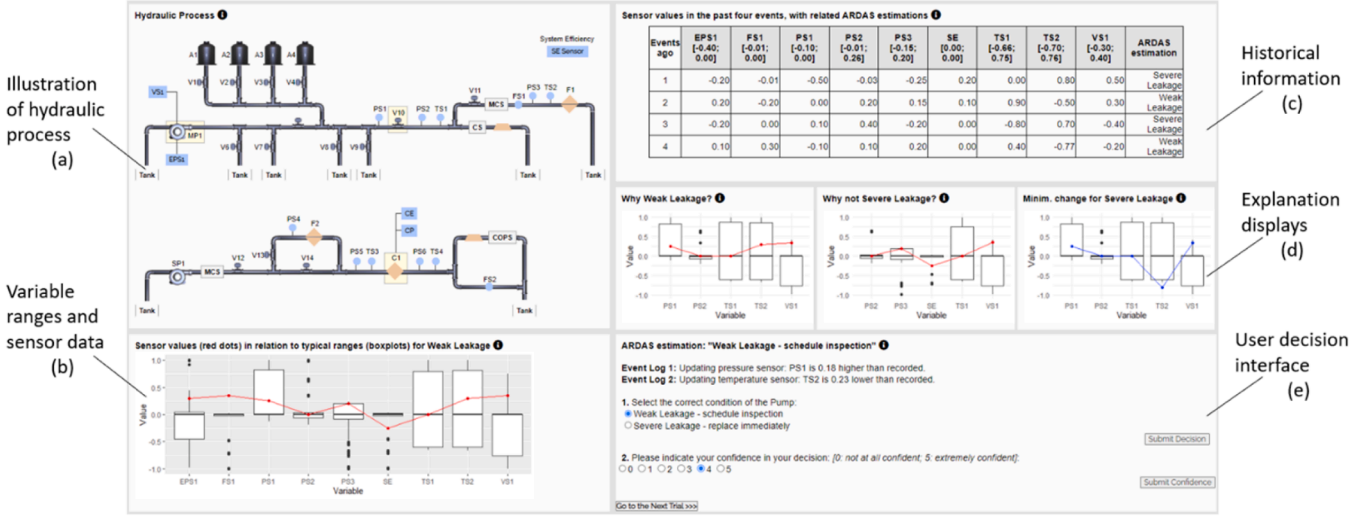

Fig. 1. ARDAS user interface. In this example trial, the component under investigation is the pump, and ARDAS is providing all three explanations (in 'Explanation displays', section d). A higher-resolution image is available in the Supplementary Materials; larger images for sections b, d, d, and e are presented in Figs. 2 to 5.

were compensated at a rate of CAD 15/hr, rounded to the closest half hour. To incentivize vigilance during the experiment, participants were informed that compensation would be reduced in case of poor performance (i.e., incorrect decisions and longer decision time relative to the other participants). However, we did not apply any reduction to the compensations, which were entirely based on study completion time. This research was approved by the Research Ethics Board at the University of Toronto (Protocol #: 00,044,164). Informed e-consent was obtained from all participants.

2.2. Experimental design

The experiment was a within-subject design with four experimental conditions, each presenting different explanatory information about the automated estimations. The four conditions were: baseline with no explanation, normative plus contrastive condition, normative plus counterfactual condition, and normative plus contrastive plus counterfactual condition. We excluded a "normative-only" condition from this study given the greater benefits we found in our previous study in the condition with combined explanations. Each experimental condition was presented in a separate block of eleven trials, for a total of forty-four trials per participant. The order of the conditions was counterbalanced across participants. The ML results presented to participants were simulated and manipulated as part of the experiment to assess how participants responded to algorithmic errors. ARDAS provided participants with two estimations: one best estimation, and one second-best estimation. The best estimation was the correct system state estimation in 82 % of trials. In the remaining 18 % of trials, the second-best estimation was the correct state estimation, which was attributed to malfunctions in the sensors that the ML system could not access. Participants, however, had access to information related to malfunctioning sensors in the form of event logs, and thus could identify when the ML did not provide the correct estimate.

2.3. Task

The experimental task involved participants consulting available information and making decisions to either agree with the ARDAS best estimation or reject it in favor of the second-best estimation. Participants were instructed to imagine themselves as consultants responsible for managing a hydraulic system with a prototype CBM system based on ML, following the company's best practices for estimating system conditions. The company's best practices dictated that operators should consider two factors in their decisions: i) the comparison of the input values from the sensor data (red dots in Fig. 1b) with the values in the training data typically associated with the ML best estimation (white area of boxplots in Fig. 1b), and ii) the most important variables used by the model to determine the ML best and second-best estimations. The company's best practices also provided that among the ARDAS best and second-best estimation, the one with a higher number of important variables looking similar to the input values was the correct estimation. Participants were instructed that if the input data looked typical for three or four of the five important variables, the ARDAS best estimation was likely to be correct; conversely, if the input data looked typical for only one or two of the five important variables, the ARDAS second-best estimation was likely to be correct. Participants were asked to submit their decision of the condition of a particular process component (pump, valve, cooler) after inspecting three elements of the interface: 1) the input values from the sensor data overlaid on the typical values in the training data typically associated with the ARDAS best estimation (Fig. 1b, replicated in Fig. 2), 2) the historical information that participants could use to assess which feature was important vs. non-important vs. of uncertain importance (Fig. 1c, replicated in Fig. 3), and 3) the event logs detailing updates in the sensor data, known to participants but not by ARDAS (in Fig. 1f, replicated in Fig. 4). The following subsections describe the task in three steps, followed by the final user decision.

2.3.1. Step 1: inspection of the sensor data

In Step 1, participants examined the input data overlaid on the typical values for the ARDAS best estimation; this allowed for visual comparison between the input data collected from the sensors and the values in the training data that are typical for the ARDAS best estimation. Fig. 2 displays the nine variables related to the pump on the x-axis, and their values on the y-axis. The white area in the boxplot indicates the typical interquartile range associated with the "Weak Leakage" estimation in the training data. The red dots represent the input data from the sensors for the nine variables, showing whether a certain input value was typical (i.e., within the white area) or atypical for a "Weak Leakage " estimation. In the example in Fig. 2, out of the nine variables related to the pump, six fall within the typical ranges of values for a "Weak Leakage" estimation, while four (EPS1, FS1, SE, and VS1) do not.

2.3.2. Step 2: inspection of the historical information

In Step 2, participants examined the historical information in tabular format to determine the importance of the variables observed in Step 1. The table in Fig. 3 displays the input values for the nine variables associated with the pump, and the related ARDAS estimations in the four events preceding the current trial. Participants followed specific rules to assess variable importance: for each variable (column), if two values fell outside the typical range (squared brackets in the header) and the ARDAS estimation changed from 'Weak Leakage" to "Severe Leakage", the variable was considered important; if two values fell outside the range but the ARDAS estimation stayed "Weak Leakage", the variable

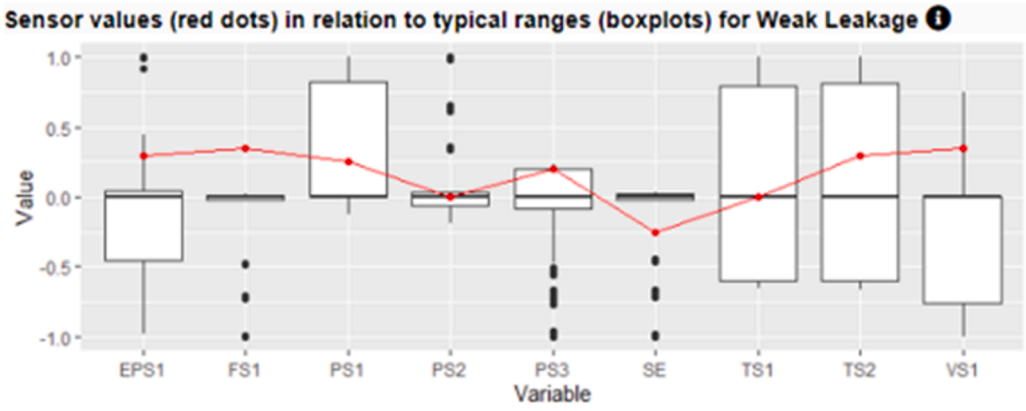

Fig. 2. Input values (red dots) of the nine variables related to the pump, overlaid on the typical ranges associated with a "Weak Leakage" estimation in the training data. This image corresponds to Fig. 1(b).

Sensor values in the past four events, with related ARDAS estimations

Events ago	EPS1 [-0.40; 0.00]	FS1 [-0.01; 0.00]	PS1 [-0.10; 0.00]	PS2 [-0.01; 0.26]	PS3 [-0.15; 0.20]	SE [0.00; 0.00]	TS1 [-0.66; 0.75]	TS2 [-0.70; 0.76]	VS1 [-0.30; 0.40]	ARDAS estimation
1	-0.20	-0.01	-0.50	-0.03	-0.25	0.20	0.00	0.80	0.50	Severe Leakage
2	0.20	-0.20	0.00	0.20	0.15	0.10	0.90	-0.50	0.30	Weak Leakage
3	-0.20	0.00	0.10	0.40	-0.20	0.00	-0.80	0.70	-0.40	Severe Leakage
4	0.10	0.30	-0.10	0.10	0.20	0.00	0.40	-0.77	-0.20	Weak Leakage

Fig. 3. Input data across the nine variables associated to the pump. Each row represent a past event, with its related ARDAS estimation. Participants could use this information to determine which variables were important, not important, or of uncertain importance in the ARDAS estimation. This table corresponds to Fig. 1(c).

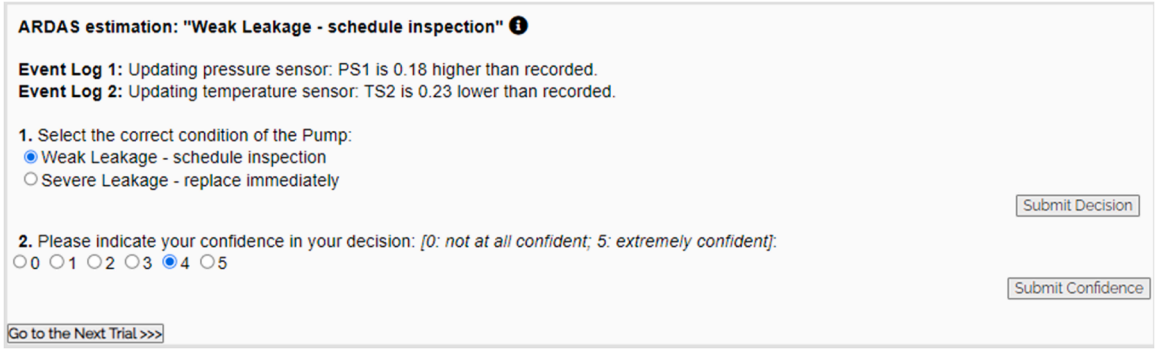

Fig. 4. User decision interface. It includes (from top to bottom): the ARDAS best estimation, the two event logs, a question on the final user decision, and a question about user confidence in their final decision. This image corresponds to Fig. 1(f).

was not important; if two variables fell outside the range, and the ARDAS estimation changed from "Weak Leakage" to "Severe Leakage" in one past event but not in the other, the importance of the variable was uncertain. For example, the variable PS1 in Fig. 3 demonstrates the application of the first rule. In the two cases when the value of PS1 exceeded the range [−0.10; 0.26], the ARDAS estimation was "Severe Leakage". This indicated that PS1 was one of the five most important variables for determining a "Weak Leakage" estimation in the valve. Other important variables in Fig. 3 are PS2, PS3 and VS1, while non-important variables included EPS1 and FS1. In this example trial, the importance of SE, TS1, and TS2 was uncertain. After inspecting the historical information table, participants knew which variables in Fig. 2 were important, unimportant, or of uncertain importance for a "Weak Leakage" estimation. The number of important, unimportant, and uncertain variables was consistent across the fourty-four trials.

2.3.3. Step 3: inspection of the event logs

In Step 3, participants inspected two event logs to assess their impact on the ARDAS best estimation (Fig. 4). If one of the event logs reported a critical change to one of the most important variables, that could have affected the ARDAS best estimation. However, in most trials, the reported update in the value of one of the most important variables (e.g., PS1 in Fig. 4) was not critical enough to exceed the typical range for the ARDAS best estimation. If the event logs reported changes in a non-important variable, the ARDAS best estimation would still be the correct estimation. However, if the reported update in the value of one of the most important variables was large enough for that variable to exceed its typical range associated with the ARDAS best estimation, then the correct estimation could change accordingly.

2.3.4. Final step: user decision

Participants considered the typical values associated with the ARDAS best estimation across all five important variables. If the input data aligned with these values, they could be 100 % confident in the ARDAS best estimation's correctness. Conversely, if the input data differed from the typical values across all five important variables, they could be 100 % confident in its incorrectness. However, the experimental design intentionally excludes these scenarios, requiring participants to make decisions under uncertainty. By analyzing the extent to which the input data resembled the typical values for the important variables, participants could infer the likelihood of the ARDAS best estimation being correct or incorrect. In the example of Fig. 2, participants knew that the input data aligned with the typical values associated with the ARDAS best estimation across three out of the five important variables. In most trials, they also had the explanation displays, which are described in the following section.

2.4. Explanation displays

Participants were tested under four experimental conditions: baseline display with no explanation, normative plus contrastive explanation display, normative plus counterfactual explanation display, and normative plus contrastive plus counterfactual explanation display. The experiment does not test a particular algorithm, but rather is tests explanations that are model agnostic. The graphical format of our explanation displays, as well as the process for deriving the variable importance, is based on the Failure Diagnosis Explainability method described in Zeldam (2018), where example-based explanations are represented in the form of boxplots, and the input values are transformed to range from −1 to 1, with the median set at 0.

2.4.1. Baseline display with no explanations

The baseline condition provided no explanation for the automated estimation. In this condition, participants saw the phrase "Currently not available" under the three questions/headers on top of the three graphical explanation sections in Fig. 1d (also replicated in Fig. 5).

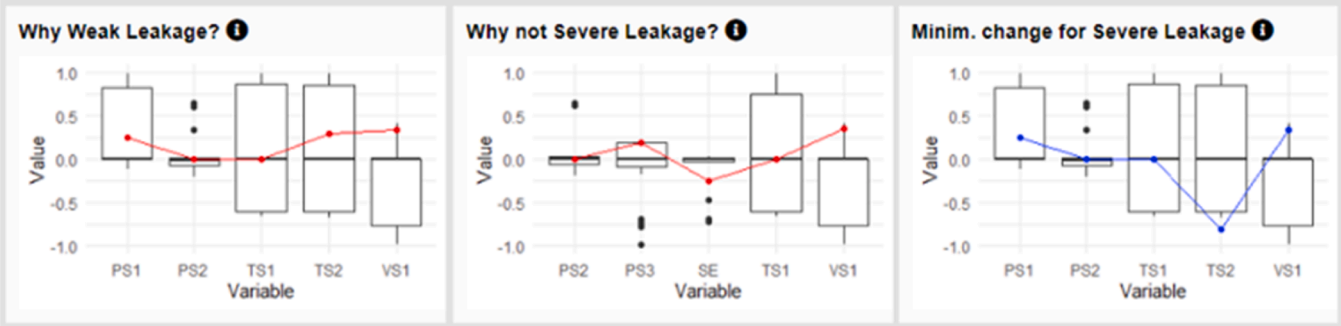

Fig. 5. Display with all three types of explanations. Current input values (the red or blue dots) are overlaid on top of the explanations to aid participants in the comparison between the current input values and the normative examples from the training data.

2.4.2. *Display with normative plus contrastive explanations*

The display with normative plus contrastive explanations provided participants with the five most important variables for the ARDAS best estimation (x-axis of the section with the header "Why Weak Leakage?"), the five most important variables for the second-best estimation (x-axis of the section with the header "Why not Severe Leakage?"), and the distribution of their values in the training data where ARDAS estimated a Weak Leakage or a Severe Leakage (white area of the boxplots in Fig. 5). In this condition, participants saw the phrase "Currently not available" under the header "Minim. change required for Severe Leakage" (Fig. 5, header of the third section). In the trial illustrated in Fig. 5, participants could conclude that ARDAS had estimated a Weak Leakage because the input data looked typical across four important variables for a Weak Leakage estimation, and only across three important variables for a Severe Leakage estimation.

2.4.3. *Display with normative plus counterfactual explanations*

This condition included a normative and a counterfactual explanation but excluded a contrastive explanation. Participants saw the phrase "Currently not available" under the header "Why not Severe Leakage?". A counterfactual explanation is operationalized in this study as a replication of the normative explanation, highlighting the smallest changes in the input data that would have turned an ARDAS best estimation into an ARDAS second-best estimation. In the counterfactual explanations, the dots are blue instead of red. In the trial illustrated in Fig. 5, the smallest change required to turn a "Weak Leakage" estimation into a "Severe Leakage" estimation was a decrease in the value of the variable TS2 from about 0.25 units to −0.75 units.

2.4.4. *Display with normative, contrastive, and counterfactual explanations*

In this display, all three types of explanations were provided.

2.5. *Procedure*

Participants signed the e-consent form, completed a brief demographic survey, and received 15-min audio-visual instructions. The instructions contained an explanation of the maintenance task as presented above, their role as the operator, and the trial structure. Participants were trained to understand all components of the interface including abbreviations (e.g., "T" for temperature, "P" for pressure, "F" for flow), and were reminded that the ML model could not access the event logs. In line with prior research in automation transparency (e.g., Bhaskara et al., 2021; Mercado et al., 2016), participants were informed that while ARDAS was highly reliable, it would not be 100 % accurate. Participants then completed three training trials, one for each explanation display. This was followed by 44 test trials without feedback. Each experimental condition was a block of 11 trials. At the end of each block, participants completed two surveys to assess their subjective workload and trust in automation. Participants were invited to take a 5 to 10-minute break between blocks. On average, the experiment duration was 2.2 h.

2.6. *Dependent measures*

We describe the dependent measures in the following order: task performance, decision time, subjective workload, and trust in automation.

2.6.1. *Task performance*

Task performance was evaluated with Signal Detection Theory (Green and Swets, 1966). Using SDT in this experiment offers the advantage of considering both the participants' ability to discriminate between correct and incorrect ML estimations, and their general bias toward accepting or rejecting the ML estimations. Since participants knew that the ML system was relatively reliable, and thus may have tended to follow the ML best estimation more often than to reject it, we defined a Hit as a correct selection of the ML second-best estimation; a False Alarm as an incorrect selection of the ML second-best estimation; a Miss as an incorrect selection of the ML best estimation; and a Correct Rejection as a correct selection of the ML best estimation. Table 1 defines the four possible SDT outcomes along a confusion matrix.

Following SDT, we operationalized task performance in terms of hit rate, false alarm rate, sensitivity (d') and response bias (c).

The hit rate is the proportion of trials where participants rejected an incorrect ARDAS best estimation, and the false alarm rate is the proportion of trials where participants rejected a correct ARDAS best estimation. Sensitivity, or parameter d′ (d prime), is a standardized score calculated in SDT as the difference between the (Gaussian) standard scores for the hit rate (H) and the false alarm rate (FA) [$d' = z(H)\text{-}z(FA)$]. In this experiment, sensitivity (d′) represents the participants' ability to discriminate between correct and incorrect ARDAS best estimations, not to be confused with "sensitivity" in information retrieval and pattern recognition, where it is also referred to as "recall". The response bias, or parameter c, assesses the degree to which participants are biased toward agreeing or disagreeing with the ARDAS best estimation; like sensitivity (d′), its calculation depends on the hit rate and the false alarm rate [$c = -0.5(z(H)+z(FA))$].

2.6.2. *Decision time*

Decision time was defined as the time taken (in seconds) by participants to select their final decision. It was measured automatically in the interface, from the trial onset to the participant clicking on the "Submit Decision" button (Fig. 4).

2.6.3. *Workload*

Workload was measured with a reduced version (i.e., without the last two items) of the NASA task load index questionnaire (NASA-TLX) also used in Gentile et al. (2023). The items in the questionnaire asked participants to rate their workload on a 7-point scale on six dimensions: mental demands, physical demands, temporal demands, own

performance, effort, and frustration. Examples of the six items include questions such as "*How hard did you have to work to accomplish your level of performance?*" or "*How insecure, discouraged, irritated, stressed, and annoyed were you?*". The workload score was averaged across the six items, resulting in a 1–7 range, where a range of 7 means that participants provided a rating of 7 to all the items in the questionnaire.

2.6.4. Trust in automation

Trust in automation was operationalized with the trust in automated systems scale from Jian et al. (2000). The scale included twelve items, such as "The system behaves in an underhanded manner" and "The system provides security." Items were measured on a 7-point scale, ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). The final score was obtained by averaging the sum scores of the twelve individual items. The final range was 1–7.

3. Results

The analysis was conducted in the R statistical package in version 4.0.4. On average, participants responded correctly to 70 % of the trials. Table 2 presents the mean and standard deviation of the dependent variables along with their correlations. All outcome measures were collected (trust and workload) or calculated (hit and false alarm rates, sensitivity (d'), and response bias) at the experimental condition level, except for decision time, which was recorded in each trial, and then averaged for each condition before analysis (i.e., for each variable, each participant had four values, one per condition). Pearson correlation coefficients were used, for which weak, moderate, and strong effect sizes are defined as 0.10, 0.30, and 0.50, respectively (Cohen 2016).

Hit rate showed a strong positive correlation with sensitivity (d'), and a strong negative correlation with response bias (c). Strong negative associations were also found between sensitivity (d') and response bias (c) ratings. Several other weak-to-moderate correlations were also found ($0.2 < r < 0.4$). For example, participants with a higher false alarm rate had also a lower response bias towards accepting the automated estimations. Additionally, participants who exhibited less response bias showed higher workload scores. Further, higher workload scores were associated with longer decision time and higher trust ratings.

We analyzed task performance (hit rate, false alarm rate, sensitivity (d'), and response bias (c)), decision time, workload, and trust in automation data using one-way repeated measures ANOVAs. We followed up significant F-tests with contrasts (*t*-tests) comparing the four conditions (Rosenthal and Rosnow, 1985). A Bonferroni correction was applied to correct for multiple comparisons. We tested for the assumptions of normality and of homogeneity of variance. Table 3 presents the descriptive statistics for each dependent variable across the four experimental conditions.

3.1. Task performance

We describe task performance measures in the following order: hit rate, false alarm rate, sensitivity (d') and response bias (c).

The hit rate resulted in three different values due to the presence of only two incorrect automated estimation per condition: 0 if participants did not identify any incorrect automated estimation, 0.5 if they identified only one, and 1 if they identified both. Due the non-normality of the data, we analyzed participants' hit rate with a Kruskall-Wallis test (McKight and Najab, 2010). Despite higher mean and median in the condition with normative plus counterfactual explanations, the results from the Kruskall-Wallis test indicated no significant differences in hit rate ($H(3)=0.79$, $p = .85$). In contrast, we found statistically significant differences in false alarm rate between conditions ($F(3, 69) = 5.20$, $p = .003$, $\eta_g^2 = 0.14$), with eta squared indicating a large effect size (Cohen, 2016; Miles and Shevlin, 2001). Post hoc testing indicated that the mean false alarm rate score was significantly lower in the condition with all three explanation displays than they were in the baseline condition ($F(1, 23) = 3.49$, p. adj. $= 0.01$). While there were no significant differences in false alarm rates between the other conditions, visual inspection of the data suggests that all explanation displays were associated with a general reduction in false alarm rate compared to the baseline (Fig. 6).

We found no significant effect in either mean sensitivity (d') ($F(3, 69) = 0.89$, $p = .45$) or mean response bias scores ($F(3, 69) = 0.94$, $p = .42$). However, the descriptive statistics suggested that the mean values of sensitivity were generally higher in the two conditions that included a counterfactual explanation, with the normative plus counterfactual

Table 3
Summary statistics, i.e., Mean (SD), of dependent measures across explanation conditions.

Variable	Explanation display			
	Baseline	*Normative plus Contrastive*	*Normative plus Counterfactual*	*Normative plus Contrastive plus Counterfactual*
1. Hit rate	0.62 (0.33)	0.55 (0.39)	0.65 (0.37)	0.58 (0.34)
2. False alarm rate	0.40 (0.21)	0.27 (0.17)	0.24 (0.18)	0.21 (0.15)
3. Sensitivity (d')	0.88 (1.61)	0.99 (1.59)	1.62 (1.84)	1.42 (1.94)
4. Response bias	−0.13 (0.84)	0.27 (1.10)	0.12 (1.01)	0.31 (0.80)
5. Decision time	95.7 (45.0)	86.1 (56.1)	70.9 (46.7)	66.9 (32.7)
6. Workload	3.54 (0.98)	2.83 (1.29)	3.10 (1.12)	3.01 (1.08)
7. Trust	3.39 (0.49)	3.07 (0.88)	3.46 (0.54)	3.30 (0.40)

Note. Larger values of d′ indicate better sensitivity. A value of $d' = 3$ represents good performance (the lower limit is −3); a value of $d' = 0$ represents chance performance (i.e., participants are guessing). Positive values of response bias signify a tendency toward agreeing with the ARDAS best estimation, whereas negative values signify a bias toward choosing the ARDAS second-best estimation.

Table 2
Overall means, standard deviations, and Pearson correlations for task performance (hit rate, false alarm rate, sensitivity and response bias), decision time, subjective workload, and trust in automation.

			Correlations						
Variable	*M*	*SD*	*1*	*2*	*3*	*4*	*5*	*6*	*7*
1. Hit rate	0.61	0.36	–						
2. False alarm rate	0.28	0.18	0.11	–					
3. Sensitivity (d')	1.28	1.75	0.90**	−0.30*	–				
4. Response bias (c)	0.11	0.94	−0.92**	−0.47**	−0.66**	–			
5. Decision time	78.83	45.70	−0.01	0.05	−0.05	−0.02	–		
6. Workload	3.13	1.13	0.18	0.32*	0.01	−0.30**	0.28**	–	
7. Trust in automation	3.31	0.60	0.19	0.01	0.16	−0.18	−0.18	0.27*	–

Note. Range for sensitivity (d′) = [−3; 3]. Range for response bias (c) = [−1; 1]. Decision time was measured in seconds. Range for workload = [1; 7]. Range for trust = [1; 7]. *$p < .05$. **$p < .01$.

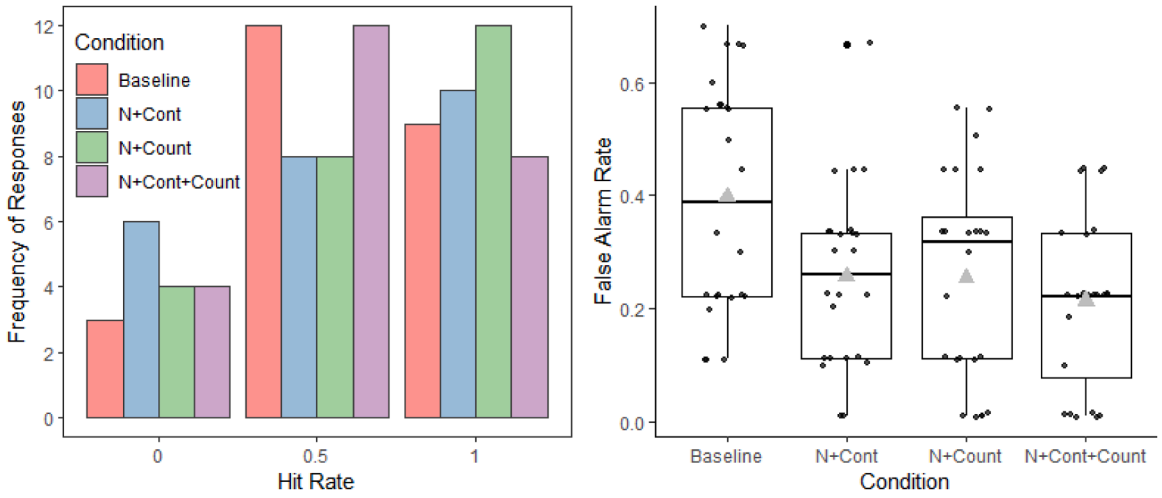

Fig. 6. Hit rate and false alarm rate across explanation condition. N+Cont = Normative plus contrastive explanations display; N+Count= Normative plus counterfactual explanations display; N+Cont+Count = Normative plus contrastive plus counterfactual explanations display.

display showing the highest mean across conditions (Table 3). Again, based on descriptive statistics, the mean values of the response bias measure suggest that participants could be more inclined to accept the ARDAS estimations in the experimental conditions that included normative and contrastive explanations, compared to the baseline and the condition with normative and counterfactual explanation (Table 3).

3.2. Decision time

Due to a potential violation of the assumption of normality, we log transformed decision time. There was a marginally significant difference between the conditions ($F(3, 54) = 2.89$, $p = .06$, $\eta_g^2 = 0.06$), indicating a medium effect size. Compared to the baseline, significantly faster decision time was reported in the condition with the normative plus counterfactual display ($t(23) = 2.96$, p. adj. = 0.04), while a marginal difference was also found in the condition with all three explanations ($t(23) = 2.73$, p. adj. = 0.07). These results suggest that participants may have experienced a reduction in decision time in the conditions that included a counterfactual explanation, as compared to the other conditions.

3.3. Workload

We found a significant difference in workload ratings between the explanation conditions ($F(3, 54) = 4.21$, $p = .009$, $\eta_g^2 = 0.05$), with a small to medium effect size. Compared to the baseline, participants reported significantly lower workload in the condition with all three explanation displays ($t(23) = 4.09$, p. adj. = 0.003), and marginally significantly lower workload with the normative plus counterfactual explanation display ($t(23) = 2.66$, p. adj. = 0.07).

3.4. Trust in automation

We found no differences in trust ratings between explanation conditions ($F(3, 54) = 1.28$, $p = .28$, $\eta_g^2 = 0.03$). The means and standard deviation values from Table 3 suggest no further trends across explanation conditions.

4. Discussion

We conducted a within-subject experiment to examine the effects of including counterfactual explanations in an automated decision aid leveraging a simulated machine learning model, and test whether counterfactual explanations led to improved decision performance compared to generic contrastive explanations. Although the number of user studies investigating the impact of explanations of automated results on human performance is growing, this is the first study to the best of our knowledge that examined counterfactual explanations in combination with normative and generic contrastive explanations. Despite the lack of significant performance differences between conditions, we observed a reduction in false alarm rates in the condition with all three explanations, and a reduction in decision time and subjective workload in the two conditions that included counterfactual explanations. These findings potentially caution against overestimating the theoretical promises of counterfactual explanations in supervised machine learning for decision support, however they can also inform the design and development of decision support systems for improved efficiency (i.e., reduced decision time and workload) in digital work environments characterized by high-frequency decisions. The rest of this section discusses some theoretical and practical implications.

4.1. Theoretical implications

The findings of this study contribute to the ongoing research on the human performance consequences of providing counterfactual explanations in decision tasks. Contrary to earlier studies like Lim et al. (2009) and Lage et al. (2019), we did not replicate clear improvements in task performance or find detrimental effects on workload and decision times. Instead, our results suggest that the inclusion of multiple types of explanations (counterfactual, normative, and generic contrastive) may be associated with certain performance benefits, such as reduced false alarm rates and lower decision time.

The results also contribute to the understanding of psychological mechanisms potentially associated with experiments in explainable AI. For example, our study addresses the issue of workload associated with explanations by showing that counterfactual explanations did not necessarily increase decision time or perceived workload compared to other conditions. This stands in contrast to Kahneman and Miller's (1986) assertion regarding the potential cognitive burden of counterfactual reasoning and implies that the specific context in which counterfactuals are presented may moderate their cognitive effects. On the other hand, the limited performance benefits associated with counterfactual explanations in this study (i.e., no difference in sensitivity d') support recent theoretical findings that model-agnostic counterfactual approaches in XAI do not support human causal thinking due to their lack of theoretical basis (Chou et al., 2024). This suggests that model-agnostic counterfactual explanations could be potentially leveraged to support users in making quick decisions with lower workload in

non-safety-critical work environments.

Having found greater benefits in the condition with combined explanations, in our earlier study, here we focused on implementing experimental conditions that presented combined explanations. To the best of our knowledge, this is one of the first studies exploring the effects of combined explanation methods on user performance. While comparison with existing studies is difficult due to differences in experimental conditions (i.e., we do not test counterfactual explanations on their own), some potential effects of counterfactual explanations emerge. First, in line with previous studies, the effects of counterfactual explanations on performance varied. We replicated neither the benefits to task performance found in Lim et al. (2009), nor the detrimental task performance effects found in Lage et al. (2019) and Lucic et al. (2020). In addition, with respect to decision time and workload, our results contrasted those reported by Lage et al. (2019), where participants experienced longer decision times and higher workload in tasks that involved counterfactual reasoning. This suggests a difference between tasks that require participants to apply counterfactual reasoning (as in Lage et al., 2019) and tasks that require participants to accept or reject automated estimations in the presence of counterfactual explanations (as in the present study). In the first case, empirical data might confirm the potential psychological burden associated with counterfactuals (Kahneman and Miller, 1986; Byrne, 2007). In the second case, the findings might align with studies reporting improved understanding with counterfactuals due to the isolation of the features that characterize one output versus its alternative (Dodge et al., 2019); this would challenge the notion that counterfactual explanations necessarily impose additional cognitive burden. The nuanced performance effects in our study are closer to the findings from Warren et al. (2022), who found that counterfactual explanations led to improvements in task performance compared to the baseline, but not compared to causal explanations. Although in our study we did not test causal explanation specifically, we observed a significant reduction in false alarm rate between the baseline and the condition with all three explanations; however, graphical inspection of the data suggest that all three conditions with explanations had a lower false alarm rate than the baseline. This points to general performance benefits of post-hoc explanations compared to baselines with no explanation. However, this says little about the individual contribution of counterfactuals on this effect. In the case of our study, participants rejected correct ARDAS estimations less in the presence of explanations, and particularly in the presence of all three types of explanations. Our study did not reveal significant differences in trust across different explanation conditions. This contrasts with previous research indicating that counterfactual explanations may positively influence trust (Warren et al., 2022). These results suggest a need to further explore the relationship between types of post-hoc explanations and users' trust in automation.

Differences in experimental design between our experiment and the studies cited above might help explain differences in our findings. One difference from the Lim et al. (2009) study is that they tested counterfactual explanations in isolation and were able to compare performance in the other explanation conditions more directly. Differences with the Lage et al. (2019) and Lucic et al. (2020) studies are more conceptual. These studies tested performance in relation to tasks that involved counterfactual reasoning, rather than under conditions that presented counterfactual explanations to justify an automated prediction, as in the present experiment. While there is conceptual overlap between these two types of tasks, they are different in nature, which may explain why we did not find any detrimental effects in task performance compared to the baseline or the other explanation conditions. One fundamental difference with our study and theirs is that our participants had the information available to complete the task without having to rely necessarily on the automated output and its estimation, although that would have saved them from completing the steps described in Section 2.3. Another difference, specific to the Dodge et al. (2019) study, concerns the nature of the task given to participants. In the Dodge et al. (2019) study, participants were not tested on a particular task, as they were in our study or in the Lim et al. (2009) study, but rather, their perception of the biased/unbiased classifier was reported. Nonetheless, it could be speculated that participants' positive perception of the classifier may be motivated by improved understanding offered by the counterfactual explanation in highlighting the feature differences that contributed to a certain automated prediction.

We also failed to replicate one finding in Gentile et al. (2023), where we had found that the combination of normative and contrastive explanations led to improvements in task performance, operationalized in terms of sensitivity (d'). This contrasting finding may be attributable to changes in experimental design, including the number of experimental conditions (from three to four) and in the number of trials that included a system failure. Since we wanted to keep the length of the experiment within three hours to avoid potential fatigue or learning effects, going from three to four conditions significantly reduced the opportunities for participants to accept an incorrect best ARDAS estimation (i.e., "miss"). This may have made a potential effect of explanations not detectable. We may conclude that two incorrect trials per experimental condition may not be sufficient to capture effects of explanations on performance in highly reliable decision support systems. Our suggestion for future user studies with highly reliable systems is to minimize the number of conditions tested to allow for reasonable trade-offs between experiment length and opportunities for participants to follow incorrect estimations.

4.2. Practical implications

These findings can provide insights for the design of decision support systems based on machine learning algorithms and similar feature-based decision aids. Specifically, the inclusion of counterfactual explanations along with normative and contrastive explanations appears to offer advantages in reducing users' workload, decision time and tendency to reject correct automated estimations (i.e., false alarms), which can be arguably acceptable in non-safety- critical digital work environments with high-frequency decisions.

4.3. Limitations

Beyond the experimental design offering limited opportunities for participants to detect an incorrect ARDAS estimation, a few other limitations can be identified. First, the results in this study may only be generalizable to environments involving binary decision tasks. While there are practical benefits of analyzing human behavior in binary decision tasks, it should also be recognized that participants could exclude one option and be sure that the other is correct. This would not hold in situations with more than two options; for example, in a scenario where ARDAS provided three plausible estimations. Another limitation is the limited sample size. While twenty-four participants represented a sufficiently large sample to detect main effects in our earlier experiment (Gentile et al., 2023), a larger sample in the current experiment would have offered more statistical power. Third, testing conditions with combinations of different post-hoc explanations does not allow us to isolate the effects of the individual types of explanations. The reductions in decision time and workload in the conditions with counterfactuals suggest that the explanations leading the effect may be the normative and counterfactual ones, but this needs to be tested specifically. An alternative interpretation would be that the performance benefits observed in this study could be due to the combined effects of explanations, and that we would not have been able to find the same benefits with conditions including individual explanations. Future research should extend these findings to diverse decision contexts and larger participant groups for a more robust understanding of the generalizability of our results.

5. Conclusions

We conducted a user experiment that tested different combinations of model-agnostic types of explanations on several measures of human performance. The experiment showed no major differences in task performance across explanation displays, however it suggests a reduction in decision time and workload in conditions that included normative and counterfactual explanations. This reduction may be due to counterfactual explanations offering the possibility of isolating the features that characterize a certain automated estimation over the other. On the other hand, this has not reflected in an improved general ability of participants to discriminate between correct and incorrect automation estimation. The reduction in false alarm rate under conditions with normative, contrastive, and counterfactual explanations warrants further exploration to investigate how different explanation methods may influence task performance, especially in domains where rejecting correct automated estimation may be particularly problematic.

CRediT authorship contribution statement

Davide Gentile: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Birsen Donmez:** Writing – review & editing, Supervision, Methodology, Investigation. **Greg A. Jamieson:** Writing – review & editing, Supervision, Methodology, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the Mitacs Accelerate program (IT16417) in collaboration with Ericsson GAIA (Global Artificial Intelligence Accelerator), and by the NSERC Collaborative Research and Training Experience (CREATE) Training and Research in Autonomous Vehicles for Reliable Services in the Air and on Land (TRAVERSAL) program.

Data availability

The authors do not have permission to share data.

References

Bhaskara, A., Duong, L., Brooks, J., Li, R., McInerney, R., Skinner, M., Loft, S., 2021. Effect of automation transparency in the management of multiple unmanned vehicles. Appl. Ergon 90, 103243.

Byrne, R.M., 2007. The Rational imagination: How people Create Alternatives to Reality. MIT press.

Byrne, R.M., 2019. Counterfactuals in explainable artificial intelligence (XAI): evidence from human reasoning. In: International Joint Conference on Artificial Intelligence, pp. 6276–6282.

Cai, C.J., Jongejan, J., Holbrook, J., 2019. The effects of example-based explanations in a machine learning interface. In: Proceedings of the 24th international conference on intelligent user interfaces, pp. 258–262.

Chou, Y.L., Moreira, C., Bruza, P., Ouyang, C., Jorge, J., 2024. Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications. Infor. Fus 81, 59–83.

Cohen, J. (2016). A power primer.

Del Ser, J., Barredo-Arrieta, A., Díaz-Rodríguez, N., Herrera, F., Saranti, A., Holzinger, A., 2024. On generating trustworthy counterfactual explanations. Inf. Sci. (Ny) 655, 119898.

Delaney, E., Pakrashi, A., Greene, D., Keane, M.T., 2023. Counterfactual explanations for misclassified images: How human and machine explanations differ. Artif. Intell. 324, 103995.

Dodge, Q.V.Liao, Zhang, Y., Bellamy, R.K., Dugan, C., 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 275–285.

Gentile, D., Donmez, B., Jamieson, G.A., 2023. Human performance consequences of normative and contrastive explanations: an experiment in machine learning for reliability maintenance. Artif. Intell 321, 103945.

Grgic-Hlaca, N., Redmiles, E.M., Gummadi, K.P., Weller, A., 2018. Human perceptions of fairness in algorithmic decision making: a case study of criminal risk prediction. In: *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Cant.

Hall, A., Murray, P., Boring, R.L., Agarwal, V., 2024. Human-centered and explainable artificial intelligence in nuclear operations. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications, Sage CA, 10711813241276463Los Angeles, CA.

Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J., 2018. Metrics For Explainable AI: Challenges and Prospects. *arXiv preprint arXiv:1812.04608*.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H., 2019. Causability and explainability of artificial intelligence in medicine. Wiley Interdisc. Rev: Data Min. Knowl. Discov 9 (4), e1312.

Jardine, A.K., Lin, D., Banjevic, D., 2006. A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mech. Syst. Sig Proc 20 (7), 1483–1510.

Kahneman, D., Miller, D.T., 1986. Norm theory: comparing reality to its alternatives. Psychol. Rev 93 (2), 136.

Keane, M.T., Kenny, E.M., Delaney, E., Smyth, B., 2023. If only we had better counterfactual explanations: five key deficits to rectify in the evaluation of counterfactual XAI techniques. In: *Proceedings of the 30th International Joint Conference on Artificial Intelligence* (IJCAI-21).

Kenny, E.M., Ford, C., Quinn, M., Keane, M.T., 2021. Explaining black-box classifiers using post-hoc explanations-by-example: the effect of explanations and error-rates in XAI user studies. Artif. Intell 294, 103459.

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S., & Doshi-Velez, F. (2019). An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv: 1902.00006.*.

Lagnado, D.A., Gerstenberg, T., Zultan, R.I., 2013. Causal responsibility and counterfactuals. Cogn. Sci 37 (6), 1036–1073.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., Baum, K., 2021. What do we want from Explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artif. Intell 296, 103473.

Leavitt, M.L., Morcos, A., 2020. Towards Falsifiable Interpretability Research. *arXiv preprint arXiv:2010.12016*.

Lewis, D., 2013. Counterfactuals. John Wiley & Sons.

Liao, Q.V., Gruen, D., Miller, S., 2020. Questioning the AI: informing design practices for explainable AI user experiences. In: Proceedings of the 2020 CHI conference on human factors in computing systems, pp. 1–15.

Lim, B.Y., Dey, A.K., Avrahami, D., 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp. 2119–2128.

Lipton, P., 1990. Contrastive explanation. Roy. Inst. Philos. Suppl 27, 247–266.

Lucic, A., Haned, H., de Rijke, M., 2020. Why does my model fail? contrastive local explanations for retail forecasting. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 90–98.

McGill, A.L., Klein, J.G., 1993. Contrastive and counterfactual reasoning in causal judgment. J. Pers. Soc. Psychol 64 (6), 897.

McKight, P.E., Najab, J., 2010. Kruskal-wallis test. The Corsini Encyclopedia of Psychology, p. 1. –1.

Mercado, J.E., Rupp, M.A., Chen, J.Y., Barnes, M.J., Barber, D., Procci, K., 2016. Intelligent agent transparency in human–agent teaming for Multi-UxV management. Hum. Fact 58 (3), 401–415.

Miles, J., Shevlin, M., 2001. Applying Regression and correlation: A guide For Students and Researchers. Sage.

Miller, T., 2019. Explanation in artificial intelligence: insights from the social sciences. Artif. Intell 267, 1–38.

Molnar, C., 2020. Interpretable Machine Learning. Lulu. com.

Rajabiyazdi, F., Jamieson, G.A., 2020. A review of transparency (seeing-into) models. In: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, pp. 302–308.

Riley, V., 2018. Operator reliance on automation: theory and data. Automation and Human Performance. CRC Press, pp. 19–35.

Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., Giannotti, F., 2021. Glocalx-from local to global explanations of black box AI models. Artif. Intell 294, 103457.

Shang, R., Feng, K.K., Shah, C., 2022. Understanding users' needs for counterfactual explanations in everyday recommendations. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1330–1340.

Shin, D., 2024. Artificial misinformation: Exploring human-Algorithm Interaction Online. Springer Nature.

Shin, D.D., 2023. Algorithms, humans, and interactions: How Do Algorithms Interact With people? Designing meaningful AI Experiences. Taylor & Francis.

Shin, D., 2021. The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. Int. J. Hum. Comput. Stud 146, 102551.

Skraaning Jr, G., Jamieson, G.A., 2023. The failure to grasp automation failure. J. Cogn. Eng. Decis. Mak, 15553434231189375.

Stepin, I., Alonso, J.M., Catala, A., Pereira-Fariña, M., 2021. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access 9, 11974–12001.

Tanyel, T., Ayvaz, S., & Keserci, B. (2023). Beyond known reality: exploiting counterfactual explanations for medical research. *arXiv preprint arXiv:2307.02131*.

Tintarev, N., Masthoff, J., 2012. Evaluating the effectiveness of explanations for recommender systems. User Mod. User-Adapt Inter 22 (4), 399–439.
van de Merwe, K., Mallam, S., Nazir, S., 2022. Agent transparency, situation awareness, mental workload, and operator performance: a systematic literature review. Hum. Factors, 00187208221077804.
Verma, S., Dickerson, J., Hines, K., 2021. Counterfactual Explanations For Machine learning: Challenges revisited arXiv preprint arXiv:2106.07756.
Wachter, S., Mittelstadt, B., Russell, C., 2017. Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv. JL Tech. 31, 841.
Warden, T., Carayon, P., Roth, E.M., Chen, J., Clancey, W.J., Hoffman, R., Steinberg, M. L., 2019. The national academies board on human system integration (BOHSI) panel: explainable AI, system transparency, and human machine teaming. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63. SAGE Publications, pp. 631–635. Sage CA: Los Angeles, CA.
Warren, G., Byrne, R.M., Keane, M.T., 2023a. Categorical and continuous features in counterfactual explanations of AI systems. In: Proceedings of the 28th International Conference on Intelligent User Interfaces, pp. 171–187.
Warren, G., Keane, M.T., Gueret, C., & Delaney, E. (2023b). Explaining groups of instances counterfactually for XAI: a use case, algorithm and user study for group-counterfactuals. *arXiv preprint arXiv:2303.09297*.
Warren, G., Keane, M.T., & Byrne, R.M. (2022). Features of explainability: how users understand counterfactual and causal explanations for categorical and continuous features in XAI. *arXiv preprint arXiv:2204.10152*.
Woodward, J., 2005. Making Things happen: A theory of Causal Explanation. Oxford university press.
Zeldam, S.G., 2018. Automated Failure Diagnosis in Aviation Maintenance Using Explainable Artificial Intelligence (XAI) (*Master's thesis, University of Twente*).