Exploring the Influence of Human System Interfaces: Introducing Support Tools and an Experimental Study

International Journal of Human–Computer Interaction

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hihc20

Exploring the Influence of Human System Interfaces: Introducing Support Tools and an Experimental Study

Chidera W. Amazu, Joseph Mietkiewicz, Ammar N. Abbas, Houda Briwa, Andres Alonso-Perez, Gabriele Baldissone, Davide Fissore, Micaela Demichela & Maria Chiara Leva

To cite this article: Chidera W. Amazu, Joseph Mietkiewicz, Ammar N. Abbas, Houda Briwa, Andres Alonso-Perez, Gabriele Baldissone, Davide Fissore, Micaela Demichela & Maria Chiara Leva (2025) Exploring the Influence of Human System Interfaces: Introducing Support Tools and an Experimental Study, International Journal of Human–Computer Interaction, 41:10, 6300-6317, DOI: 10.1080/10447318.2024.2376354

To link to this article: https://doi.org/10.1080/10447318.2024.2376354

Published online: 19 Jul 2024.

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
2025, VOL. 41, NO. 10, 6300–6317
https://doi.org/10.1080/10447318.2024.2376354

Exploring the Influence of Human System Interfaces: Introducing Support Tools and an Experimental Study

Chidera W. Amazu[a], Joseph Mietkiewicz[b], Ammar N. Abbas[c], Houda Briwa[d], Andres Alonso-Perez[d], Gabriele Baldissone[a], Davide Fissore[a], Micaela Demichela[a], and Maria Chiara Leva[d]

[a]Politecnico di Torino, Turin, Italy; [b]Hugin Expert A/S, Aalborg, Denmark; [c]Software Competence Centre Hagenberg (SCCH), Hagenberg im Mühlkreis, Austria; [d]Technological University Dublin, Dublin, Ireland

ABSTRACT

Situational awareness and decision support tools such as procedures and alarm systems are vital for effective interaction among control room operators, especially in safety-critical situations. In safety-critical environments such as process plants, there remains a gap in evaluating specific tools during actual operations, or "work-as-done." Additionally, the underlying factors that might impact operators' cognitive states and performance concerning safety have not been thoroughly explored. The need for such an evaluation is further bolstered by current interaction configurations where operators are more passive than active, thus reducing their cognitive performance. Therefore, this experimental study addresses the highlighted evaluation gap by introducing and comparing three human system interfaces/decision support tools in four human-in-the-loop configurations. The supports include two alarm design formats (prioritized vs. non-prioritized) and three procedure representation formats (paper, screen-based digitized, and an AI-based support system built with an integrated Bayesian network and reinforcement learning model). Ninety-two people (n = 92) participated voluntarily in the test. They were divided equally into four groups. Each group tested three safety-related events in a simulated formaldehyde production facility. Individuals belonging to the group with prioritized alarms and utilized paper procedures rated procedural support slightly higher on average than others in different groups. Unlike the other groups, their assessment of alarm prioritization support remained consistent across all scenarios. Further analysis of the impact of the setup on cognitive states and actual performance will be performed.

KEYWORDS

Human–machine interaction; process control rooms; decision support; situational awareness; workload; stress; psychophysiological measures

1. Introduction

Elements that comprise human system interfaces, such as mimics, alarms, and, in some cases, computerized or screen-based procedures, are of paramount importance for the effective performance of operators in both normal and abnormal plant states. The operator, responsible for stabilizing the state of the plant following feedback observable through alarms or different cues on the display, has to rely on experience, training, and intervention procedures to achieve this goal. Accidents such as the Buncefield oil storage depot in Hertfordshire (HSE, 2011) and the BP America Texas city refinery fires and explosions in 2005 (US Chemical Safety & Investigation Board, 2007) have highlighted the importance of these organizational elements. In the Buncefield case, it was noted that the display mimics overlapped each other, thereby obscuring the identification of some critical alarms. The operators had to make extra effort to navigate the different mimics. Furthermore, it was noted that the procedures for filling the storage tank were short of details (HSE, 2011). Similar reasons can be identified in the U.S. Chemical Safety and Hazards Investigation report for the BP Texas accident (US Chemical Safety & Investigation Board, 2007). The report noted issues such as the use of outdated and ineffective procedures, the prevalence of false alarms and control indications, and the failure of operators to follow standard operating procedures correctly. Similar recommendations regarding these accidents were given in the survey by Amazu, Abbas, et al. (2023), which investigates the state-of-the-art process industries on human-centered design and management of these elements. This shows that much is yet to be done regarding these issues, especially in process industries.

These elements support operators' cognitive state, particularly their situational awareness, essential for accurate decision-making in alarm handling and process control. Specifically, interaction configurations characterized by intermediate levels of automation help mitigate the risk of operators experiencing out-of-the-loop conditions. (Endsley & Kaber, 1999). Their importance has led to research attempting to perform human-in-the-loop studies, which involve a holistic evaluation of the impact of several possible factors on operators' cognitive state and performance. These experimental

CONTACT Chidera W. Amazu chidera.amazu@polito.it Politecnico di Torino, Turin, Italy

studies have investigated factors such as displays as a standalone performance shaping factor with varying levels and, in some cases, in combination with a few other factors, to understand their impact on operators' cognitive load and situational awareness, as summarized in a review by Amazu, Demichela, et al. (2023). Such studies have also motivated the exploration of novel monitoring techniques, such as eye tracking, electroencephalogram, heart rate monitoring, etc., for these evaluations. In a few cases, these novel techniques have been used simultaneously with more traditional methods, such as the situational awareness rating technique (SART) or its variations for situational awareness, the NASA task load index (NASA-TLX) for workload, etc. (Braarud, 2021; Shi & Rothrock, 2022). The limited available literature suggests that much remains to be explored in combining traditional techniques like SART with novel methods such as eye tracking for assessing human-in-the-loop situational awareness and the impact of decision support tools. Additionally, there are gaps in performing a deeper analysis of multiple vital performance-shaping factors combined. Emphatically, those that are key to decision-making and situational awareness.

This study introduces new situational awareness and decision support tools alongside an experiment design, which has the potential for adoption within the industry for human-in-the-loop evaluations during drills or training. The experimental study utilized the following support tools: alarm design, procedures, and a display interface, each with distinct levels of implementation. Notably, the study examined two levels of alarm designs: prioritized alarms versus non-prioritized alarms and three levels of procedures: conventional paper-based, digitized screen-based, and AI-based procedures. As a result, the combination of procedural variations and alarm design levels within different test groups determines the type of display interface to be utilized during the study. Using data from this study, the authors can further compare how these tools impact operators' cognitive states of attention, situational awareness, workload, performance, and safety. The study uses subjective and objective questionnaire-based methods and novel monitoring tools for data collection.

For this article, the reviewed literature on the critical situational awareness and decision support tools to be studied, that is, the alarm systems and procedures, are discussed in (cf. Section 2) with a clear elaboration of the research gaps, contributions, and hypotheses. This section is followed by a presentation of the methods, specifically the design of the experiment, the case study, and the evaluation tools and measures implemented for this study (cf. Sections 3, 4 and 5). After reporting the experiment's results (cf. Section 6), the article concludes with an overall evaluation of the approach, a discussion (cf. Section 7) and finally a conclusion section (cf. Section 8).

2. State of the art

2.1. Alarm systems

Alarm systems are essential in large industrial facilities. They are crucial to ensure operational efficiency, maintain safety standards, and avoid potential disasters. Historical incidents, such as the Piper Alpha accident and the BP Texas refinery explosions, underscore the critical link between effective alarm management and overall plant safety (Crompton, 2021). The emphasis has been placed on good alarm rationalization as a vital process in alarm management, recognizing that the design and functionality of alarm systems directly impact operator workload and response efficiency (Ghosh & Sivaprakasam, 2020). Despite advances and existing industry guidelines, challenges persist in achieving optimal alarm management practices within process industries. A few experimental investigations, such as in Simonson et al. (2022), have been carried out to analyze the impact of the alarm system design.

2.2. Paper vs. digitized screen-based intervention procedures

Operating or troubleshooting procedures are crucial to successful human-machine interaction in process control rooms. They provide operators with information on "what to do" during normal and abnormal situations. The mode of reading interaction and the format in which the documents are presented can affect comprehension (Leroy et al., 2023) and the error rate (Xu et al., 2008), respectively. For example, paper-based procedures have been written in formats that have been cumbersome to follow or difficult to update subsequently. Hence, there have been efforts towards computerized procedures and other procedural support systems (Kim et al., 2013). However, very little work has been done to investigate or validate these different representation formats and how they support operators. Xu et al. (2008), while comparing two different representation formats of emergency procedures, identified that representation formats can significantly influence high error rates during tasks. Gao et al. (2013) in their experimental study evaluated the impact of low- vs. high-complexity screen-based emergency operating procedures on operator workload. Their focus has been mainly on mental workload with no consideration of situational awareness yet. However, the few experimental studies on situational awareness in process plants have primarily focused on assessing interface displays using eye tracking and EEG metrics for predicting human errors, employing a one-task step procedure in their scenarios (Bhavsar et al., 2017; Kodappully et al., 2015). In comparison, some simulated operator interventions were based on training and judgment (Iqbal & Srinivasan, 2018). Given the state-of-the-art procedure designs in process plants where bulky papers are used, no work has been done to empirically analyze the impact of procedures on operators' performance, especially on operators' situational awareness. Also, the effect of procedures on behavioral metrics and situational awareness is yet to be explored.

2.3. AI-based recommendation systems

Given the increasing complexity of control rooms, operators often face complex information streams. In critical scenarios, operators may be inundated with hundreds of alarms and pieces of information simultaneously. Such information

overload can be counterproductive, as the data presented in the control room might add to their confusion instead of aiding the operator. In these circumstances, the implementation of a decision support system becomes imperative. Unlike human operators, mathematical models can efficiently process vast amounts of information and determine the optimal decision. The fusion of intelligent systems within manufacturing and operations management has traditionally been seen as a beneficial confluence of operational research (OR) and artificial intelligence (AI). This collaborative potential is underscored in studies by Proudlove et al. (1998) and Kobbacy and Vadera (2011). However, in safety-critical infrastructures like process plants, such systems are hardly used (Amazu, Demichela, et al., 2023).

Weidl et al. (2005) introduced a methodology for root cause analysis in industrial operations using object-oriented Bayesian networks (OOBNs), showcasing their ability to model industrial system uncertainties and dependencies for enhanced decision support. The study highlights OOBNs' adaptability and advantages in predictive maintenance and operational efficiency despite challenges with data quality and model complexity. Horvitz and Barry (2013) further explored Bayesian networks in time-sensitive decision-making, proposing interface designs that display probabilistic information for quicker, informed decisions. Both studies underscore the potential of Bayesian networks in improving decision-making processes. However, a notable limitation of both studies is the lack of participant-based testing to assess the impact on human performance, workload, and situational awareness empirically.

2.4. Contribution

- Unlike previous studies examining decision support tools in isolation, this work uniquely investigates the combined effects of alarms, procedures, and interface displays. Rather than focusing solely on alarm design or interface displays, we explore how these elements interact within process control room operations.
- A novel aspect of our research is the examination of actual procedures used in process control rooms, as opposed to some other studies where the tasks are presented as pop-ups on the display (Bhavsar et al., 2017; Iqbal & Srinivasan, 2018; Kodappully et al., 2015).
- Another distinctive feature of our study is using three different displays, which exceeds the norm in human-in-the-loop process control experiments. By incorporating multiple displays, the study offers a more comprehensive analysis of operator interactions within the control room environment.
- This study also introduces four human-in-the-loop configurations, which vary based on the combination of support tools employed. These configurations include different types of procedures (paper-based, digitized screen-based, AI-based) and alarm systems (with or without prioritization), each influencing the type of interface display utilized.
- Furthermore, our study introduces an AI-based decision support tool for comparison with traditional procedures.
- Methodologically, our evaluation employs a blend of qualitative and quantitative techniques. In addition to the use of questionnaires like NASA-TLX for workload assessment and SART, which are known for situational awareness evaluation, the authors also designed questions to further assess the situational awareness of the operators using the SPAM methodology. Additionally, eye tracking and health monitoring tools offer valuable insights into cognitive states such as mental workload, situational awareness, and stress levels.
- The data generated from this study enriches the fields of human factors and cognitive science as it can be explored to provide detailed insights into the factors that impact operators during safety-critical scenarios in different process control configurations. Beyond examining operator-system interactions and psycho-physiological states, we also gather training, experience, and demographic data. These can be explored further for human reliability analysis.
- Finally, given the recurrent role of procedures and alarm systems in accidents across various sectors, including nuclear operations (Gao et al., 2013), our experimental design and findings hold promise for enhancing safety practices in these domains.

3. Methods

3.1. Human system interfaces

Three key human system interfaces, considered decision-support tools, are designed for this study and varied to form different human-in-the-loop configurations. These are the Alarm systems prioritization, Support Procedures (Paper, Screen, and AI-Based Support), and the varying displays due to the alarm and procedure representation format (Table 1). They support the cognitive processes of the operators until the execution and evaluation of actions, as represented in Figure 1. The tools designed for this study are further described below.

3.1.1. Alarm systems

Eighty alarms were introduced in this plant for this study. Most importantly, these alarms were implemented to simulate nuisance and alarm floods. The alarm system design was divided into two parts, as seen in Figures 7–10. The first part is the alarm list, which refers to the layout of the alarm box and how it is displayed. This includes details on the acknowledgement box, sound, priority numbering, alarm state, activation time, tag, and plant section. Boxes are

Table 1. Key independent variables in this study.

Decision support	
Procedures	AI recommendation
	Paper-based procedures
	Digitized screen-based procedures
Alarm design	Prioritized alarms
	Non-prioritized alarms
Interface configurations	G1,G2,G3,G4

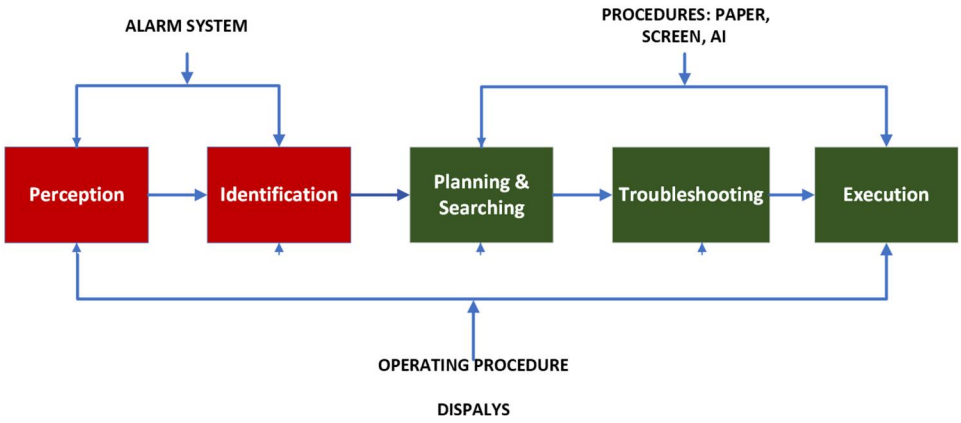

Figure 1. Support tools and the cognitive processes involved.

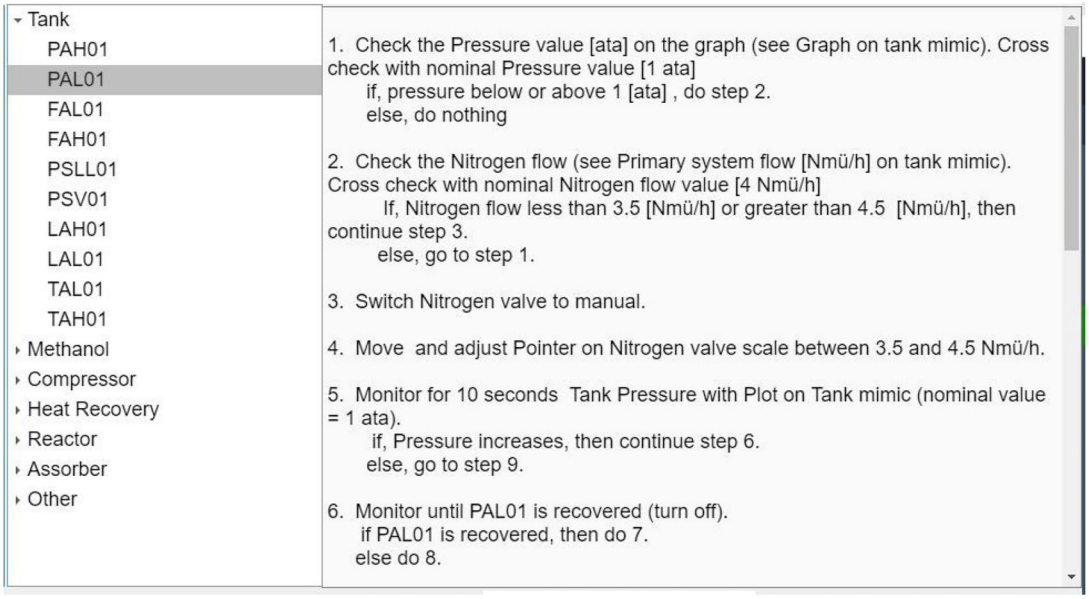

Figure 2. Procedure on screen: first steps of intervention procedure for low-pressure alarm (PAL01).

provided for the operators to silence and acknowledge each Alarm. In this study, the critical alarm is expected to be acknowledged first so that the supervisors can assess their perception and ensure the most important task is done first. The second part is alarm prioritization, which refers to the priority assigned to the alarms using three key color schemes (yellow, orange, and red for low, medium, and high priorities, respectively). The non-prioritized alarms remain white for all levels (see Figure 7).

3.1.2. Intervention procedures

The intervention procedures in this study are written in a hierarchical rule-based task representation format, as shown in Figure 2. The paper- and screen-based procedures follow a similar writing format. However, the presentation on paper rather than on screen makes the difference.

The screen-based procedures are organized by alarms and plant sections to facilitate the operator's search process, as seen in Figure 2. For example, in Figure 2, six Plant sections are visible: Tank, Methanol, Compressor, Heat Recovery, Reactor, and Absorber. When the operator is interested in an alarm within the Tank section and clicks on the word "Tank," the alarm tags as shown below "Tank" in Figure 2 are displayed. Upon clicking, for example, PAL01, the task steps to be followed to resolve that alarm are further displayed on the right side of the box.

The paper procedure instead has a table of contents to ease navigation when using the booklet. This table of contents is organized according to alarm numbers. For example, alarms ending with 01 come before those ending with 02. In addition, just like the screen-based procedures, the alarms are further grouped according to plant sections within the paper. For example, the alarms ending with 01 generally belong to the Tank section, which comes first in the paper, and so on. Yellow tags were also placed before the start of each plant section to ease the search process.

Each intervention procedure is written under three broad tasks: a) Troubleshooting, b) Control, and c) Evaluation. The first two task steps, as seen in Figure 2, represent the troubleshooting phase, 3 and 4 illustrate the control action phase, and 5 and 6 represent the evaluation phase. These were written following HPOG guidelines for procedure, job list, and checklist design (HPOG Steering Committee, 2021). In addition to the

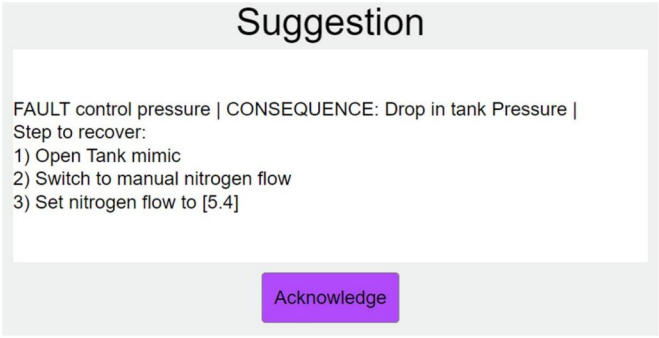

Figure 3. Simplified procedure recommended to the operator.

intervention procedure, an operating procedure was provided that contains a summary of the task expected of a control room operator and the nominal values and limits for each process variable for additional support. The operating procedure summarizes the expectations during Process Monitoring, Alarm Handling, and Intervention Planning.

3.1.3. AI decision support systems

This study also evaluates a decision support system (DSS) enhanced by artificial intelligence, specifically designed to assist control room operators by providing timely procedural guidance. The core of the DSS is an integrated system utilizing influence diagrams and reinforcement learning to deliver actionable insights during operational scenarios. These methodologies are chosen for their proven efficacy in improving decision accuracy and operational safety through adaptive learning and predictive analytics.

The architecture of the reinforcement learning component is adapted from the specialized reinforcement learning agent (SRLA) framework, tailored for use in safety-critical industries (Abbas et al., 2022), and specifically instantiated for the process control industry (Abbas et al., 2023). This adaptation ensures the DSS is robust and applicable to the specific challenges faced in process control environments.

While one of this manuscript's primary focus is assessing the DSS's impact on operator situational awareness and workload, a comprehensive outline of the AI system's design is detailed in referenced sources (Mietkiewicz et al., 2023). These references provide an in-depth look at the construction and theoretical underpinnings of the influence diagrams, as well as the reinforcement learning models used. We encourage readers to consult these works for a thorough understanding of the system's foundational elements.

Figures 2 and 3 of our paper depict the streamlined decision-support procedures implemented. During the experiments, operators had the option to use either traditional procedures, the AI-enhanced DSS, or a combination, allowing us to study the impact on trust and reliance on decision-support technologies.

3.2. Experimental design

The present study assesses the effect and impact of the above decision support tools on control room operators' cognitive state, performance and behavior in safety-critical scenarios. A formaldehyde production facility case study is considered for the experimental study as explained below under *Case Study*.

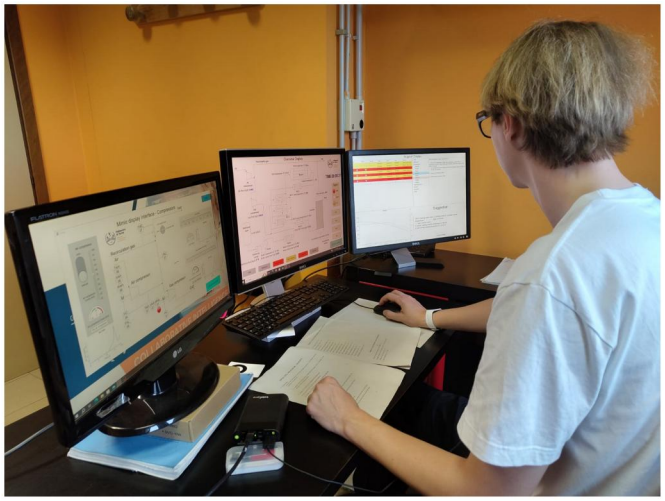

Figure 4. Example operator with the AI configuration (G4) (Mietkiewicz et al., 2024).

A further description of the use case and experiment design is presented below. Four groups are set up to study these tools at different levels. Hence forming unique process control human-in-the-loop configurations. The participants are equally clustered within these groups, 20 per group, each performing the scenarios in order of assumed complexity, that is, scenarios 1–2–3. To counterbalance the possible effect of the time of day of participation, the participants per group were randomized based on the time of day.

3.2.1. Case study

The case study involves the production of formaldehyde from the partial oxidation of methanol and air and a secondary reaction, which completes the oxidation to carbon monoxide, reducing the yield of formaldehyde in the reaction. This case study on formaldehyde production, initially utilized by Demichela et al. (2017) and later adapted by the authors, focuses on a plant comprising six main sections.

Illustrated in Figure 5, the plant's layout begins with a Tank Section housing a methanol storage tank equipped with alarms to signal deviations from expected parameters (as shown in Figure 6). Following this is the Methanol Section, where liquid methanol transforms into a gas via a pump and heater before being combined with compressed gas from the Compressor Section. The resulting mixture then passes through a heat exchanger (REC2), raising its temperature to around 200 °C before entering the reactor. The Reactor Section houses the reactor, while the Heat Exchanger Section accommodates heat exchangers REC1, REC2 and REC3. REC1 and 3 cool the reactor product. The heat recovered from REC1 and 3 are then exploited for water boiling and cooling, respectively.

The Reactor is housed within the Reactor Section of the plant, while the heat exchangers are housed within the Heat Exchanger Section. The other heat exchangers, REC1

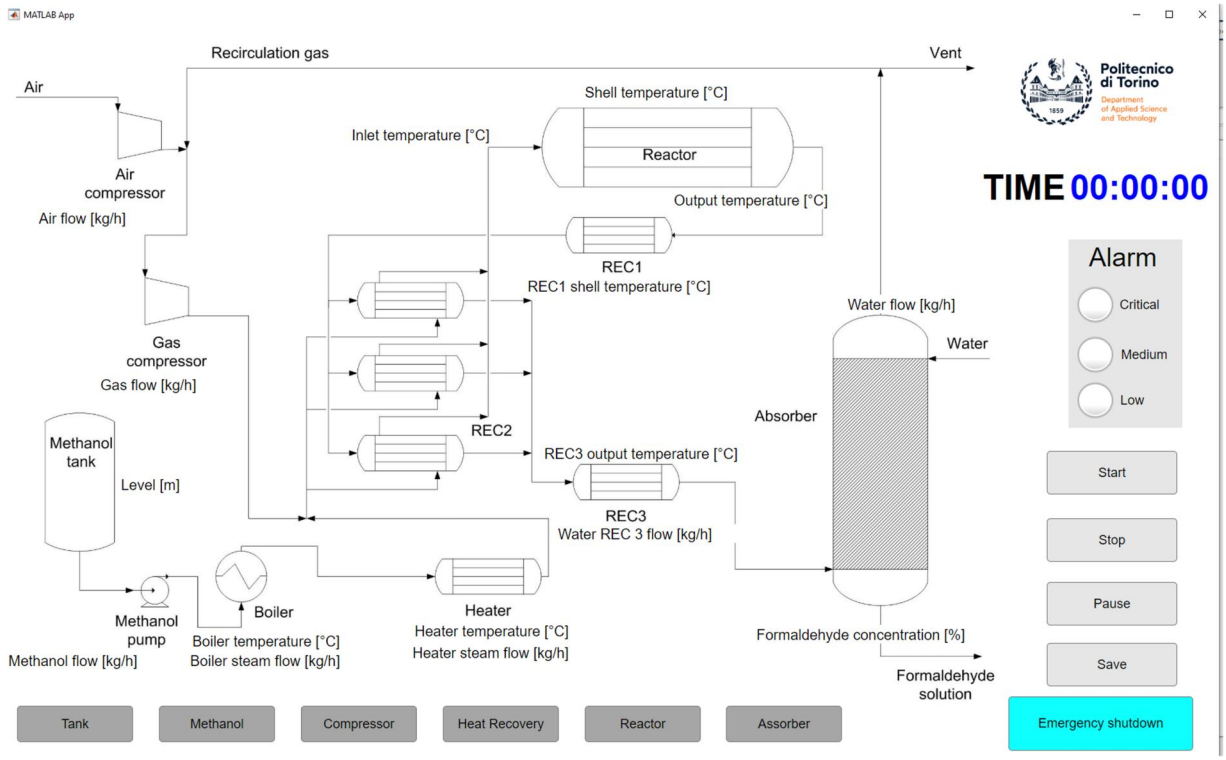

Figure 5. Main screen.

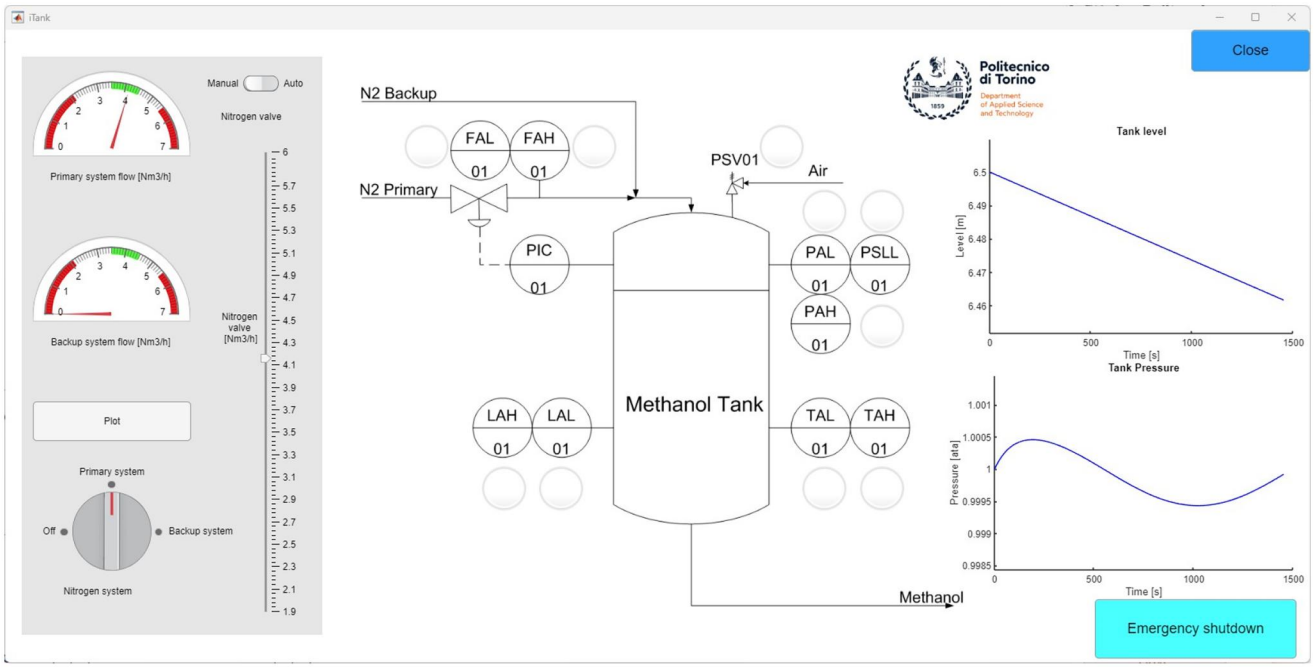

Figure 6. Tank section of the plant (Mietkiewicz et al., 2024).

and REC3, cool the product from the reactor, with the heat subsequently exploited for boiling and cooling water in the plant. REC3 aims to cool the product to an absorption temperature of around 67 °C before entering the absorber, where it is absorbed by water flowing in the opposite direction. This absorber resides in the Absorber Section. Navigation through these plant sections is facilitated by buttons at the interface display's bottom (see Figure 5).

Several accident scenarios are likely in this type of plant. The reactor or absorber could be overheated if the temperatures from the mixture or product are far above the expected. An interference to control this temperature by increasing the water input into the absorber can lead to losing the target product volume.

Comparing this present work to the original design by Demichela et al. (2017), changes to the interface display are discussed in Section 3.2.2. Additionally, the safety-critical

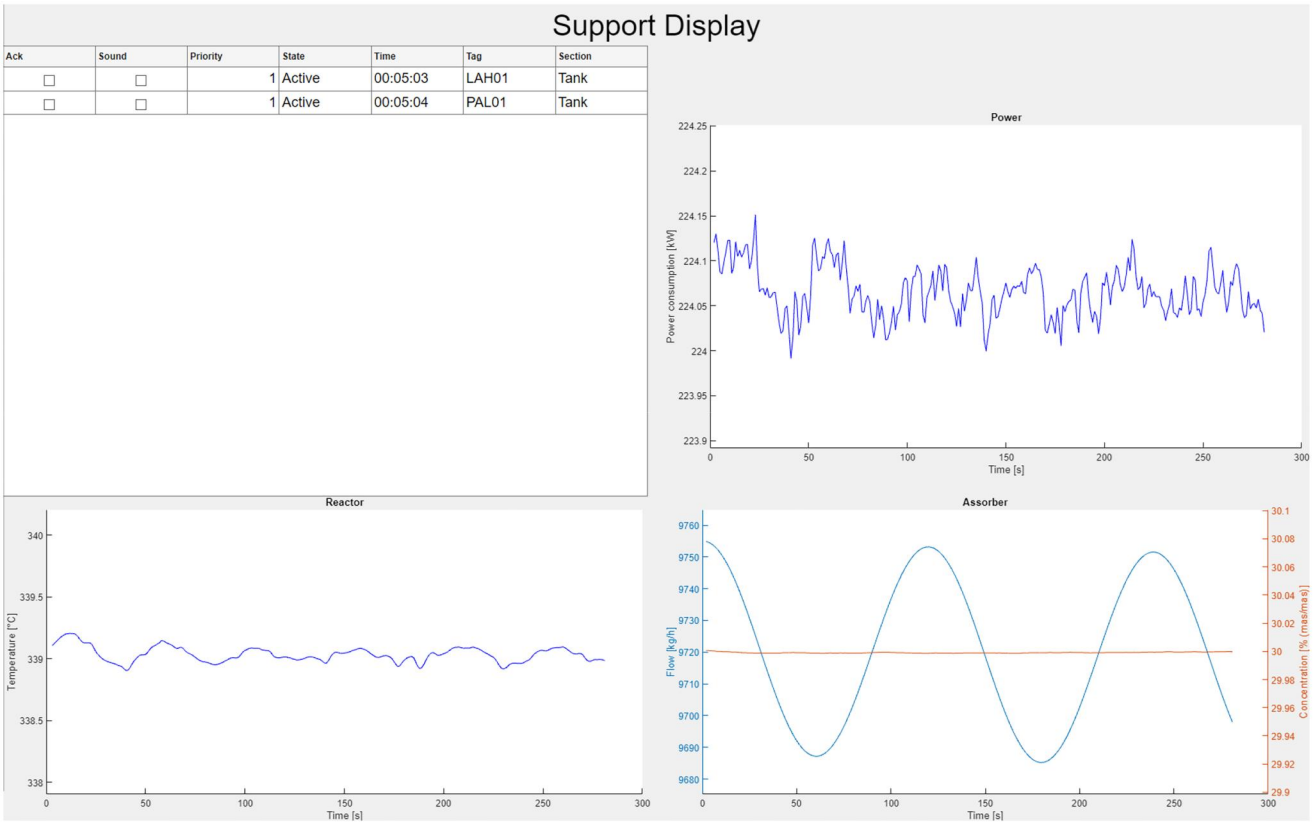

Figure 7. Support display G1. The alarms are all in white without prioritization.

scenarios identified to be presented as task challenges to participants in this study are later discussed in this paper.

3.2.2. The interface displays

The following improvements have been made to the initial design: An increase in the number of alarms in all six sections of the plant (tank, methanol, compressor, heat recovery, reactor and absorber), additional user interaction buttons on the overview display for the experiment (see Figure 5), redesign of buttons to switch from auto to manual after a few usability tests (see example in Figure 6), addition of an extra "close" button for easy visibility and design of a support interface display (see example in Figure 6). The simulator employs three displays. The central or overview display presents the process flow diagram of the plant with a central alarm notifier, buttons to navigate the different sections of the plant, and buttons to control scenario selection or start and end of scenarios (see Figure 5). The left interface shows the mimics of the plant sections when the operator clicks to open them (see the example tank mimic in Figure 6). The right interface is the support display (see Figures 7–10). A holistic view of the setup and interfaces can be seen in Figure 4.

Four distinct human-in-the-loop configurations, and effectively, the human-machine interfaces, were developed following a combination of the support tools:

- The first interface does not include alarm rationalization.
- The second interface incorporates alarm rationalization.
- The third interface displays procedures on-screen.
- The fourth interface, in addition to the features of the third, integrates an AI-based decision support system.

The difference between the groups appears more specifically in the support display, as seen in Figures 7–10.

3.3. Procedure

This work involved human subjects and has received approval for all ethical and experimental procedures and protocols from the internal ethics committee of the collaborative Intelligence for Safety Critical (CISC) project in Dublin, Ireland, with a supporting letter from the Technological University Dublin.

Participants were selected based on the following criteria: age limit of 18+, proficiency in English, and normal or corrected-to-normal vision. The flyers were distributed among the university and student groups. The flyer contained a link to register, which included information on the selection criteria. Participants were sent certificates of participation after the entire data collection phase was completed.

Upon arrival, participants were given a brief introduction to the experiment, after which they read a two-page document containing more information on what to expect during the test. This was followed by signing the consent form and questionnaire to collect demographic data on age, gender, and course or level of study. They then watched a

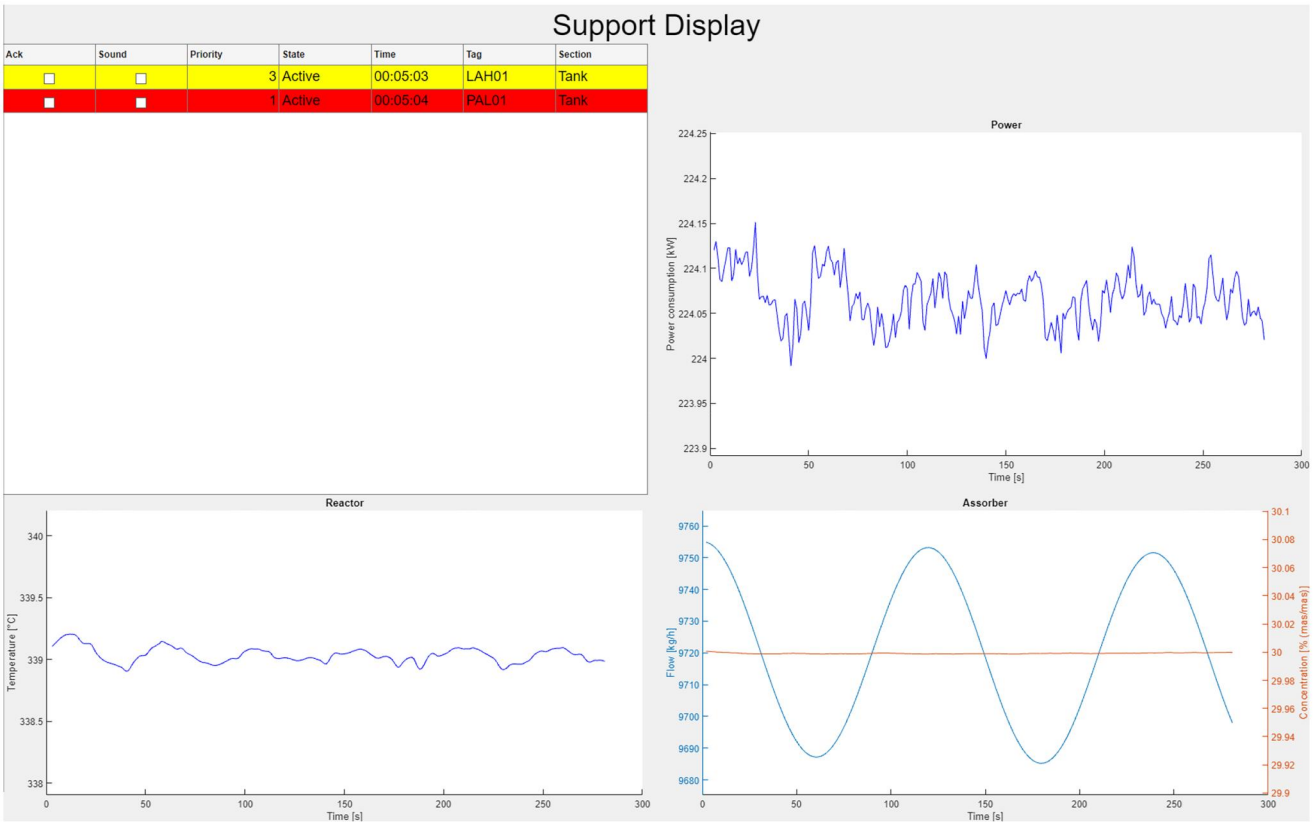

Figure 8. Support display G2. The alarms are color-coded based on priority.

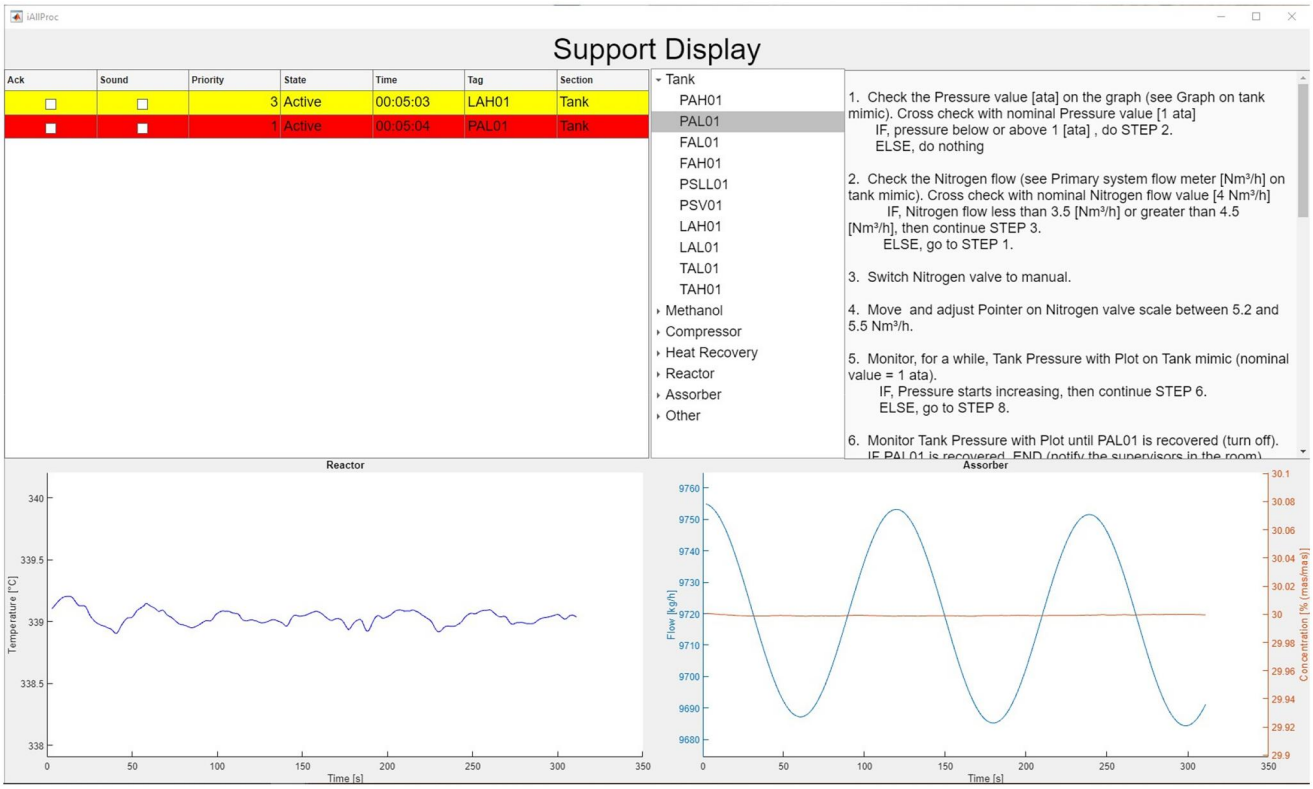

Figure 9. Support display G3. The procedure specific to an alarm can be found in the support display.

5-min video as the introductory phase of training, which contained a brief intro on the different interface displays and rules to follow during the experiment. This was followed by a 30- to 45-min training on the process dynamics, operating procedures, and the procedures that varied depending on the study group. After the training, the

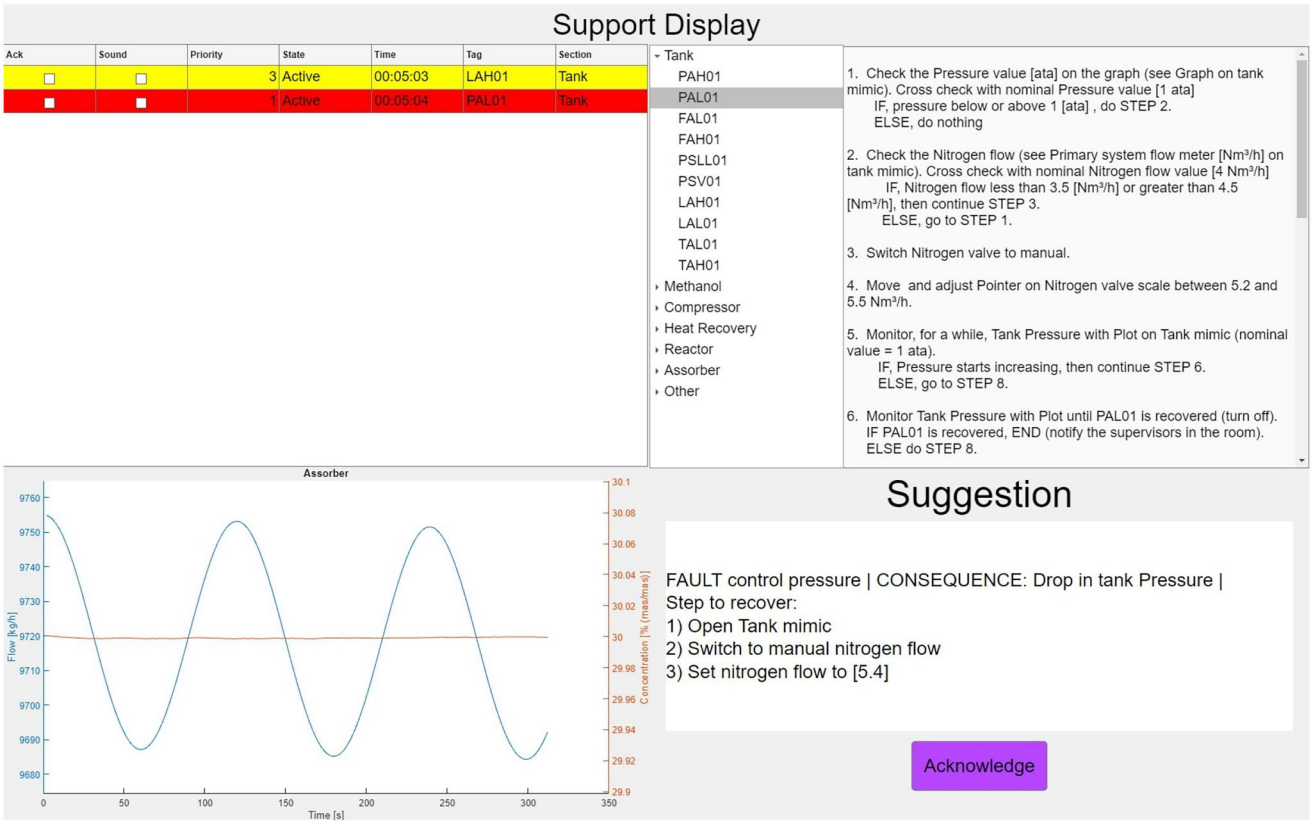

Figure 10. Support display G4. A suggestion box presenting optimal actions to the operator (Mietkiewicz et al., 2024).

participants were allowed to take a 5- to 10-min break before setting up devices. The eye-tracker and health monitoring watch were then set up after the break. Each test lasted 15–18 min, followed by completing the NASA-TLX, SART and support rating questionnaire. The SPAM technique was applied during the test without interrupting the process. These subjective techniques are further described in Section 5.

3.3.1. Participants

A total of 92 participants (36 female, 56 male), comprising students and staff of the Politecnico di Torino and a few externals, voluntarily participated in this study. The participants were between 21 and 61 years of age (M = 25, SD = 5.4) and had different experience levels. Most of the participants were junior process engineers selected voluntarily from among the students of the master courses in chemical engineering at Politecnico di Torino. In addition, there were more experienced engineers, such as PhD candidates and some professors.

3.3.2. Groups

In the structured experiment, we delineated four groups tailored to assess the incremental benefits of various support interfaces within a simulated control room environment. Comparative analysis was methodically planned to isolate the impact of each added feature by juxtaposing each group with its predecessor. Here is a more detailed exploration of the group design and comparative objectives.

- Alarm Rationalization Impact: The first group functioned as the baseline, operating without the benefit of alarm rationalization, digital procedure, or AI-based support system. The second group, in contrast, was equipped with an alarm rationalization system to compare its influence on the operator's workload and decision-making process. Alarm rationalization is expected to filter out noncritical alarms, thereby reducing the cognitive load on operators and enabling them to focus on the most pertinent issues.
- On-Screen vs. Paper Procedures: The third group was provided with on-screen procedures, a step up from the second group that relied on traditional paper-based methods. This comparison evaluated the operational efficiency and response time between the two mediums. Screen procedures could offer quicker access to necessary information, reduce the time spent searching through physical documents, and streamline decision-making by seamlessly integrating with other digital tools in the control room.
- AI Decision Support System Evaluation: The fourth group was provided the largest level of support by integrating an AI decision support system. This group was compared against the third group to gauge the incremental benefits of AI assistance. The AI decision support system is designed to synthesize information and provide recommendations. This comparison sought to quantify

the effectiveness of AI in improving operator performance, reducing errors, and improving overall system safety and efficiency.

Each group's performance was recorded and analyzed to determine the efficacy of alarm rationalization, digitalization of procedures, and AI-driven decision support in a control room setting. The outcomes of these comparisons are anticipated to offer insights into the design of future control room interfaces and decision support systems, ultimately contributing to the advancement of safe and efficient industrial operations. The characteristics of each group are summarized in Table 2.

3.3.3. Scenarios

We selected three scenarios to evaluate our study's different human-in-the-loop (HITL) configurations. Each scenario simulates a specific failure or challenge an operator might encounter in a plant environment. These scenarios test the robustness of the elements in the setups and the operator's ability to respond effectively under varying conditions. The details of each scenario are as follows:

1. Pressure indicator control failure. In this scenario, the automatic pressure management system in the tank ceases to function. Consequently, the operator must manually modulate the nitrogen inflow into the tank to preserve the pressure. During this scenario, the cessation of nitrogen flow into the tank results in a pressure drop as the pump continues to channel methanol into the plant.
2. Nitrogen valve primary source failure. This scenario is an alternative version of the first. In this case, the primary source of nitrogen in the tank fails. The operator has to switch to a backup system. While the backup system starts slowly, the operator has to regulate the pump power to maintain the pressure inside the tank.
3. Temperature indicator control failure in the Heat Recovery section. In this scenario, there's a risk of the reactor overheating, with subsequent pressure increase. The primary objective is to prevent the activation of the pressure switch (PSL01) within 18 min following the initial Alarm.

Table 2. Characteristics of the groups.

	Alarm rationalization	Procedure on screen	AI support
G1	No	No	No
G2	Yes	No	No
G3	Yes	Yes	No
G4	Yes	Yes	Yes

4. Data collection and plan for analysis

Data was collected using the devices shown in Figure 11. These tools are further explained in this section.

4.1. Surveys

To assess workload and situational awareness as perceived by the participant, including gaining insight into the perceived level of support of the different support systems, the demographics, training, and experiences of the participants, we administered questionnaires, which were completed at various points in the test as described in the protocol. For situational awareness assessment, the questionnaires used include the situational awareness rating technique (SART) and situation present assessment technique (SPAM). For workload, NASA-TLX was used. The other questions for demographics, etc., were designed into questionnaires by the researchers. A detailed description of the measures is presented in Section 5. These are to be analyzed and compared between groups to address the research hypothesis and understand the key organizational and individual factors that impact the operators the most in the different configurations and given certain safety-related initiating events. The comparison of the variables, especially those based on

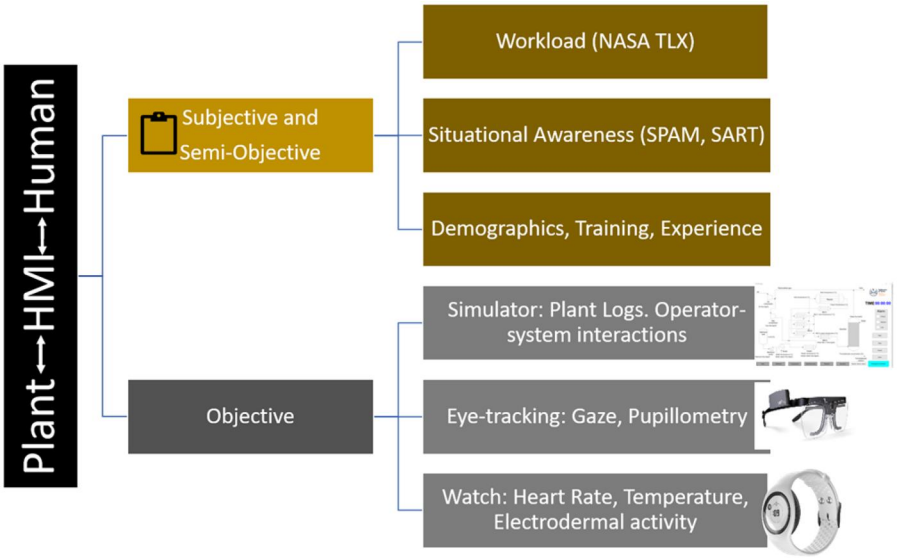

Figure 11. Data collection tools during the human interface interactions including subjective and objective measures obtained.

operators' perception, with more objective measures can benefit from a wholesome analysis of workload, situational awareness and stress.

4.2. Simulator

The simulator data from the plant for each scenario were recorded per participant. This included data from the plant systems and the operators' interactions with the plant through the displays, such as manipulating valves, acknowledging alarms, and more. By collecting such data, the authors plan to derive performance and behavioral metrics that can further be used to compare the different HITL configurations. Also, future studies in related industries can identify key metrics to predict performance and outcomes in safety-related scenarios.

4.3. Eye-tracker

The eye-tracking technique was deployed for data collection to gain in-depth insight into operators' cognitive states and understand their mental workload, situational awareness, and fatigue states. The glasses used were Tobii Glasses 3, 50 Hz. It records various eye-related gaze and pupillometry-related metrics, such as pupil diameter, average fixation duration, etc. The goal is to gain insight into the differences between group cognitive states given the varying configurations and during each scenario. The raw recordings are to be analyzed in-depth using Tobii Pro Lab software.

4.4. Watch

In this study, we utilized the EmbracePlus smartwatch to monitor vital physiological parameters: heart rate, temperature, and electrodermal activity. These metrics facilitate the comparison of physiological responses across different groups. Our analysis focuses primarily on discerning any significant differences in these measurements between the study groups.

5. Measures

Participants assessed the effectiveness with which decision support tools facilitated their performance at the end of each scenario. The outcomes of the alarm system and procedural aspects are presented in Section 6.

The questions asked for the alarm and procedure support assessment, including task load assessment, are as shown below, with participants scoring them on a scale of 1 to 5 (1: Low, 5: High):

- Task load: How complex did you perceive the task to be?
- Alarm list support: How clear was the information received on the interface for the list of alarms, and how do you acknowledge, silence, or follow up on them?
- Alarm prioritization support: How helpful was the support received in differentiating the alarm priorities?
- Procedure support: How helpful was the support of the consulting procedures during the scenario?
- AI support: More questions were collected to assess the operators' level of trust in the AI, quality, level of the help of suggestion, and explainability, complexity, and validity of the support. Some sample questions are shown below:
  1. How would you rate the suggestion of AI support?
  2. How would you rate the level of explainability of the AI support?
  3. How high is your trust in the decisions suggested by the recommendation system?

5.1. Subjective measures

For the subjective measures, NASA TLX was used to assess workload, while SPAM and SART were used to assess situational awareness.

1. NASA-TLX: the NASA Task Load Index used in this study is the same as described in Braarud (2021). It follows the standard TLX question format with little word changes for easy comprehension and alignment with the type of task. The NASA TLX index is calculated using the formula:

$$f_{(a_1,\ldots,a_6)} = \sqrt{\frac{1}{6}\sum_{i=1}^{6} a_i}$$

   The a_i represents the NASA TLX dimensions. That is, a) perceived mental, physical and temporal demands from tasks, b) effort and perceived performance on task, c) perceived effort utilized and frustration level.
2. SART: the standard questions used in the situational awareness rating technique were used for this study. The SART index is a composite measure of situational awareness comprising three dimensions: SART Understanding, Demand, and Supply Braarud (2021), and is calculated with the formula:

$$U - (D - S)$$

   Situation Understanding (U) comprises Information Quantity, Information Quality, and Familiarity. Situation demand (D) includes the situation's Instability, Complexity, and Variability. At the same time, the Supply of attentional resources (S) comprises Arousal, Concentration, Division of Attention, and Spare Capacity.
3. SPAM: Questions were developed and asked the participants at three points during each scenario: during alarm handling, while planning to intervene/first part of reading the procedures, and finally, after the intervention. In most cases, these questions were asked in the scenario's 6th, 8th, and 12th min. This is based on a non-freeze approach as applicable in SPAM, with each question assessing the perception, understanding and projection levels of situational awareness, respectively, as defined by Endsley. A concurrent think-aloud approach was

used for the participants' response assessment. The SPAM Index is calculated as:

$$f_{(a_1,\ldots,a_3)} = \sqrt{\frac{1}{3}\sum_{i=1}^{3} a_i}$$

The a_i represents the SPAM dimensions of Perception, Understanding and Projection as further explained by the questions asked to the participants. The responses were recorded by the supervisors/researchers and scored on a scale of 1–5. 1 means 'very low' situational awareness, and 5 means high situational awareness. The same questions used in scenario 1 to assess perception, understanding, and projection were used in scenario 2. The questions are detailed below:

Scenarios 1 and 2:

- Perception (Question 1): Which of these alarms, in your opinion, requires to be verified first [$FAL01, PAL01$]? and why? (AI system: What is the AI decision support system about?)
- Understanding (Question 2): Why do you think the PAL01 alarm is activated? And what do you intend to do? (AI system: What was the suggestion on? Was it clear what you were expected to do and why?) (1: SA level Low, 5: SA level High)
- Projection (Question 3): Now that you have done this, what do you think will change in the system? Why?

For scenario 3, the questions assessing "perception" and "understanding" differed from those in scenarios 1 and 2 due to the type of alarms annunciated. However, the question on "projection" is similar across all scenarios, as shown below:

Scenario 3:

- Perception (Question 1): Which of these alarms, in your opinion, must be verified first [FAL11, TAH17]? and why? (AI system: What is the AI decision support system about?)
- Understanding (Question 2): Why do you think the TAH17 alarm is activated? And what do you intend to do? (AI system: What was the suggestion on? Was it clear what you were expected to do and why?)
- Projection (Question 3): Now that you have done this, what do you think will change in the system? Why?

5.2. Objective measures

1. Performance:

Performance and behavior data were derived from the operational logs/simulator. Below are some of the selected metrics.

- Overall performance: this considers the time it takes to recover the low-pressure Alarm or, in some cases, those who fixed the fault even before an alarm. Those who fall below or equal to the 25th percentile are grouped as "optimal performance," those who fall below or equal to the 50th percentile are classified as "good," and the rest as "poor performance."
- Reaction Time: this is the time it takes to switch the Nitrogen valve button from Auto to Manual depending on the scenario and initial task as written in the procedures.
- Response Time: the time it takes to act. For example, scenario 1 means the time it takes to adjust the nitrogen valve scale to the correct value.

2. Eye-tracking Metrics

Our study used Tobii Pro Glasses 3 for eye tracking and Tobii Pro Lab for robust data analysis. Combining these technologies allowed us to investigate various eye-tracking metrics, providing valuable insights into participants' visual behaviors. The metrics considered in our analysis are enhanced by identifying areas of interest (AOI) for data mapping onto the snapshot and Time of Interest (TOI) delineations. Specifically, we categorized TOIs into baseline (pre-alarm occurrence), critical alarm regions, and the alarm flood, allowing for a more granular examination of visual attention dynamics. The following metrics were considered in our analysis:

- Heat Map: The heat map generated from the eye-tracking data visually represents the areas that captured the participants' visual attention. Brighter regions on the heat map indicate higher fixation density, offering insights into the focal points within the visual stimuli. When overlaid with AOIs, the heat map offers a detailed spatial representation of visual attention within specific regions of interest. This insight becomes particularly valuable when correlating the heat map with TOIs, revealing how visual attention evolves across different experiment phases.
- Number of visits: This metric quantifies the frequency with which participants revisit specific AOIs. It provides valuable information about the temporal patterns of attention, indicating whether certain areas are consistently revisited or if attention shifts over time.
- Visit Duration: Visit duration measures participants' time on each visit to specific AOIs. This metric complements the "Number of Visits" by offering insights into the temporal persistence and engagement within these areas.
- Number of Fixations: The number of fixations highlights the frequency with which participants shift their gaze between different points of interest. This metric aids in discerning patterns of visual exploration and identifying areas that consistently capture attention. By associating the number of fixations with AOIs and TOIs, we gain insights into the frequency of gaze shifts between different points of interest across different experimental phases. This spatial-temporal correlation allows us to identify consistent attention patterns throughout the study.

- Fixation Duration: Fixation duration represents when participants focus on a specific point of interest. This metric helps to identify elements that attract prolonged attention, contributing to our understanding of information processing and cognitive engagement. When aligned with AOIs and TOIs, fixation duration helps pinpoint elements that elicit long attention during baseline, critical alarm scenarios, and the subsequent alarm flood. This temporal alignment facilitates a nuanced exploration of cognitive engagement and information processing dynamics.
- Saccade Duration: Saccades are rapid eye movements between fixations. Analyzing the duration of these movements provides insights into the efficiency and fluidity of participants' visual scanning patterns. It can reveal information about the decision-making process, contributing to our understanding of cognitive processing speed. By analyzing saccade duration with Areas of Interest (AOIs) and Times of Interest (TOIs), we can study how efficiently and smoothly participants scan visually under different conditions: baseline settings, critical alarms, and alarm floods. This approach contributes to understanding how cognitive processing speed evolves in response to changing stimuli.
- Pupil Diameter: Pupil diameter is a metric that reflects changes in cognitive load and emotional arousal. By measuring pupil size variations, we understand the mental effort and emotional responses associated with specific visual stimuli. The variations in pupil diameter, observed with AOI and TOI, provide an understanding of cognitive load and emotional responses related to specific points of interest. This approach allows us to discern how visual stimuli affect participants' cognitive processes over time intervals.

By leveraging these eye-tracking metrics, we aim to unravel the intricacies of participants' visual attention and cognitive processes in response to the stimuli presented, shedding light on the underlying mechanisms that shape human perception. Incorporating AOIs and TOIs into our analysis framework enhances eye-tracking metrics' precision and contextual relevance, ensuring a comprehensive exploration of visual behavior. Leveraging Tobii Pro Glasses 3 and Tobii Pro Lab, we aim to unravel our study's complex interplay between visual attention, cognitive processes, and emotional responses.

3. Smart Watch

The following are some key measures collected during the test using the EmbracePlus smartwatch.

- Heart Rate: Heart rate is the number of heartbeats per unit of time, usually expressed as beats per minute (bpm). It is a vital physiological parameter that reflects the cardiovascular system's activity. It is often used to indicate arousal, stress, or emotional responses. In cognitive and health studies, monitoring changes in heart rate can provide insights into the autonomic nervous system's activity and overall cardiovascular health.
- Temperature: Temperature refers to a body or environment's degree of hotness or coldness, usually measured in degrees Celsius (°C) or Fahrenheit (°F). Body temperature is a fundamental physiological parameter. In cognitive science, changes in body temperature may be linked to stress responses, mental workload, or emotional states.
- 3. Electrodermal activity (EDA): Electrodermal activity, or galvanic skin response, measures the skin's electrical conductance. It is influenced by sweat gland activity. EDA often indicates sympathetic nervous system activity associated with emotional responses and arousal. In cognitive science, monitoring electrodermal activity can provide information on stress levels, emotional engagement, and the overall response of the autonomic nervous system during various cognitive tasks or stimuli.

5.3. Control variables

Data is also collected on the participants' demographics, including their age and gender, their course, year of study, familiarity with the process industry and control rooms, and their training assessment.

6. Result

Two participants per group were omitted to analyze the overall questionnaire data, given the number of missing data from them. The missing data left were further filled up using the mean of the data samples. Hence, 21 participants per group were used for the survey-data-related analysis below.

Table 3 shows the mean of the perceived support of the alarm system and procedure as rated by the participants on a scale of 1–7. 7 means excellent support, and 1 means poor support.

A Shapiro-Wilk L1 Test, Shapiro and Wilk (1965), was first used to assess the normality of the data for each group. The group data were not normally distributed for each scenario in all cases, as seen in Table 4. Therefore, the Mann-Whitney U Test, Mann and Whitney (1947), was used to evaluate further the statistical significance of our mean data in the group-scenario comparison, as shown in Table 5.

6.1. Perceived support

6.1.1. Alarm list support

- Scenario 1:
  Although Group 4, which used AI, had procedures displayed on the screen and employed alarm prioritization, it reported the highest average score for alarm list support (M = 4.38). This did not significantly differ from the scores of Group 2 (M = 4.33), which used procedures on paper with alarm prioritization, nor from Group 1 (M = 4.33), which used procedures on paper without alarm prioritization, or from Group 3 (M = 4.14), which used procedures on-screen with alarm prioritization.

Table 3. Comparison of operator perceived support while using the support tools of alarm list, alarm prioritization, and procedures (paper for Group (G) 1 and 2, screen for Group 3 and 4). Note that not all participants in Group 4 used the procedures on screen since they had the AI. *M = mean, SD = Standarddeviation, Med = Median*.

		Alarm list support			Alarm prioritization support			Procedure support (paper and screen)		
Scenario	Group	M	SD	Med	M	SD	Med	M	SD	Med
S1										
	G1	4.33	0.80	5	2.81	1.54	3	4.14	1.15	5
	G2	4.33	0.97	5	4.38	0.74	5	4.33	1.11	5
	G3	4.14	0.71	4	4.27	0.83	4.5	4.23	0.97	4.5
	G4	4.38	0.67	4	4.43	0.60	4	4.26	1.02	5
S2										
	G1	4.14	1.01	4	2.67	1.35	3	4.05	1.12	4
	G2	4.24	1.04	5	4.33	0.73	4	4.57	0.98	5
	G3	4.15	0.93	4	4.40	0.75	4.5	4.15	0.99	4
	G4	4.25	0.79	4	4.18	0.94	4.25	3.98	0.98	4
S3										
	G1	3.57	1.25	4	2.19	1.50	1	3.62	1.36	4
	G2	3.86	1.06	4	4.14	1.15	5	3.48	1.12	4
	G3	3.64	1.26	4	3.68	1.36	4	3.32	1.52	3.5
	G4	3.38	0.86	3	3.62	1.16	4	3.10	1.34	3

Table 4. Shapiro-Wilk Test results for procedure support, alarm list support, and alarm priority support across three scenarios (S1, S2, and S3). $p = 0.00$ shows non-parametric distribution.

	S1				S2				S3			
	G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4
Procedure support	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.06	0.01	0.06
Alarm list support	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.00
Alarm priority support	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.02

Table 5. Mann-Whitney U Test results for Group-wise comparison across the three scenarios. $p < 0.05$ shows statistical significance.

	S1			S2			S3		
	G1 vs. G2	G2 vs. G3	G3 vs. G4	G1 vs. G2	G2 vs. G3	G3 vs. G4	G1 vs. G2	G2 vs. G3	G3 vs. G4
Procedure support	0.51	0.51	0.89	0.04*	0.06	0.50	0.61	0.88	0.58
Alarm list support	0.84	0.19	0.25	0.71	0.61	0.86	0.50	0.64	0.25
Alarm priority support	0.00**	0.72	0.67	0.00**	0.72	0.51	0.00**	0.24	0.71

**also for G4, ** also for G3 & G4*

Notably, Group 4 had a median score (Med = 4) lower than Group 2 (Med = 5), indicating that despite having a higher mean, the central tendency of the ratings for Group 4 was lower. The absence of statistically significant differences, as shown in the Mann-Whitney U Test results, suggests a homogeneous perception of alarm list support across all groups in this scenario.

- Scenario 2:
  The perception patterns in Scenario 2 mirrored those of Scenario 1, with no group displaying a statistically significant difference from the others in their perception of alarm list support. Group 4 again reported a marginally higher mean score (M = 4.25). Still, this slight variation did not reflect a significant difference in the actual support experience, evidenced by the consistent median scores across groups and the non-significant *p*-values from the statistical tests.
- Scenario 3:
  Scenario 3 presented a shift in the pattern, with Group 4 reporting a notably lower average rating (M = 3.38) for their experience of alarm list support. Despite this lower mean score for Group 4, all groups maintained a similar perception of alarm list support. Group 4's mean score was affected by lower ratings, but the consistent median across the groups indicates a shared central experience. The differences were not statistically significant according to the Mann-Whitney U Test

6.1.2. Alarm priority support

- Scenario 1 (S1):
  In Scenario 1, the reported median scores for alarm prioritization support show Group 1, which lacked alarm prioritization, at a clear disadvantage with a median of 3, compared to Groups 2, 3, and 4, which had higher medians of 5, 4.5, and 4, respectively. This is reinforced by the mean scores, where Group 1's mean of 2.81 is significantly lower than those of Groups 2 (4.38), 3 (4.27), and 4 (4.43). The Mann-Whitney U Test results indicate a statistically significant impact of alarm prioritization on perceived support, placing Group 1 at a marked shortfall.

- Scenario 2 (S2):
  Maintaining the pattern observed in Scenario 1, Group 1 showed the lowest median score for alarm prioritization support (Med = 3) in Scenario 2, underlining a persistent perception of inadequate support, with this finding being

statistically significant as indicated by the Mann-Whitney U test. Conversely, Groups 2, 3, and 4 reported higher median scores (Med = 4, 4.5, and 4.25, respectively), further substantiating the hypothesis that the provision of alarm prioritization is linked to an enhanced perception of support. The absence of this feature in Group 1's workflow leads to a continuous perceived support gap compared to the other groups.

- Scenario 3 (S3):
  In Scenario 3, Group 1's median support for alarm prioritization remained the lowest (Med = 1), indicating a persistently lower perceived support. The other groups—Group 2 (Med = 5), Group 3 (Med = 4), and Group 4 (Med = 4)—again reported higher median scores. The Mann-Whitney U Test results illustrate the significant difference in perceived support between Group 1 and Group 2, underlining the critical role of alarm prioritization in shaping operators' support perceptions.

6.1.3. Procedure support

- Scenario 1 (S1):
  Groups 1,2, 3, and 4 had mean scores of 4.14, 4.33, 4.23, and 4.26, respectively. No statistically significant differences were found between groups based on the Mann-Whitney U test in Scenario 1.

  In Scenario 1, the median scores for procedure support were consistent across Groups 1, 2, and 4 (all with Med = 5), while Group 3 had a slightly lower median of 4.5. These median scores, alongside mean scores of 4.14 for Group 1, 4.33 for Group 2, 4.23 for Group 3, and 4.26 for Group 4, suggest an overall similar perception of support across all groups, a finding supported by non-significant results from the Mann-Whitney U test, indicating no substantial differences in procedure support experience.

- Scenario 2 (S2):
  In Scenario 2, Group 2 stood out with the highest mean (4.57) and median (5) scores for procedure support, indicating a positive perception of paper-based procedures. Group 1 followed closely with a median of 4 despite a slightly lower mean (4.05). Group 3 also held a median of 4, with a mean of 4.15, and Group 4, even with AI support, had the lowest mean (3.98) and a median that matched Group 3 (4). The Mann-Whitney U test results suggest significant differences between Groups 1 and 2, which could indicate a more difficult use of the procedure without alarm rationalization.

- Scenario 3 (S3):
  Scenario 3 showed a general decline in mean scores for procedure support across all groups, with Group 1 at 3.62, Group 2 at 3.48, Group 3 at 3.32, and Group 4 at the lowest with 3.10. However, the median scores were closer, with Group 1 at 4, Group 2 at 4, Group 3 at 3.5, and Group 4 at 3. Despite the observable drop in mean scores, the lack of statistically significant differences per the Mann-Whitney U test underscores a shared perception of procedure support across all conditions.

7. Discussion

7.1. Support systems: Groups per scenario

Alarm list support: The lack of statistically significant differences between the groups implies that, despite numerical discrepancies in mean, all groups in the first scenario have a consistent baseline experience of alarm list support. This is expected because they all have the same alarm list support format.

Alarm priority support: The pattern observed as we move from *S*1 to *S*2 indicates a trend where participants in Group 1 consistently perceived lower levels of alarm priority support compared to Groups 2, 3, and 4 across all scenarios. While Groups 2, 3, and 4 had varying mean scores across scenarios, the absence of statistically significant differences between these groups suggests a relatively stable perception of alarm priority support within each group across the scenarios. The consistently lower perception of alarm priority support in Group 1 across all scenarios suggests alternative support is needed. Insight on the importance of this factor when predicting performance would be required to weigh the actual impact given such configurations. These results from group 1 with non-prioritized alarms align with the literature indicating the downsides of non-prioritization or rationalization of alarms (EnergyInstitute, 2010; Meshkati, 2006).

Procedure Support: The significant differences observed between Groups 1 and 2 in *S*2 and between Groups 2 and 4 highlight variations in participants' perceptions of procedure support. Group 2 consistently rated higher, indicating a more robust perception of support than Group 3 or 4 with AI assistance. Interestingly, no statistical difference was found between Group 3 (using procedures on screen) and the other groups, suggesting similarities between these representation formats. Further investigations may provide additional insights, including questions assessing the different formats' effectiveness.

Surprisingly, Group 4, despite AI support, initially exhibited higher ratings than Group 3 in *S*1. This result, including the close nature of the mean values with that of other groups, shows a reliance on the procedures despite the AI suggestion. This behavior contrasts with literature suggesting more concise forms of representation (Park & Jung, 2003). Seeing that the concise format was not sufficient enough, the digitized procedures had to be consulted. Therefore, this might be ascribed to cases with new technology designs for procedures, such as using AI, as trust in AI can be an initial issue.

However, as observed in other groups, Group 4's ratings also declined as scenarios progressed, indicating an increased reliance on AI over procedural support. The decline in ratings across all groups in S3 may be linked to the escalating difficulty of the scenarios, especially scenario three. These inexplicable observations underscore the

importance of exploring the contextual factors influencing participants' perceptions during each scenario.

Significantly, Group 4's ratings were found to be similar to those of Group 3, underscoring that incorporating an AI decision support system does not negatively affect other support components within the interface. This observation reinforces that introducing AI-based decision support systems in control rooms can be advantageous without disrupting the established support frameworks.

7.2. Limitations

It is essential to recognize the study's limitations, including its reliance on self-reported measures and lack of more detailed questions for evaluating the procedure format. The study can also be improved with more complex scenarios and alarm conditions and the use of actual control room operators.

Furthermore, the controlled nature of the study might not adequately reflect the dynamic and unpredictable aspects of real-world safety-critical situations. The results of the self-perceived ratings by the operators might also be influenced by the controlled nature of the study, given that they are more tensed to perform accurately and within the available time.

The focus on specific human system interfaces, and decision support tools may overlook other contextual factors and individual differences influencing operators in control room environments. This limitation can impact the outcomes in adequately capturing some underlying factors needed to predict error, situational awareness, or workload.

Additionally, the study's reliance on a single simulated formaldehyde production facility may limit the generalizability of findings to different industrial processes and control room settings. Moreover, the study does not consider the potential long-term effects of various interfaces and tools on operators' cognitive states and performance, as it only assessed immediate impacts without exploring fatigue, burnout, or other lasting effects over time. Thus, it is probable that the long test duration, especially for the last scenario, might influence the participants' subjective ratings on their perception of the support of the systems, workload and situational awareness. Addressing these limitations is suggested for future research.

Although we tried to include experienced operators, most participants were engineering master's students, with a few being more experienced engineers, such as PhD candidates and professors. This disparity in practical experience may compromise the transferability of our findings, as real-world operators could face distinct challenges and respond differently to decision support systems.

Also, implementing AI decision support systems presents challenges. First, modelling an entire industrial process in a DSS is highly complex, and the decision to focus it on a specific segment of the system or to try to broaden it could affect its efficacy. Further, deploying a DSS in an existing control room might present compatibility and safety problems. The trust of the operators in DSS could be achieved by specific training and instructions, and the allocation of responsibilities should be clearly defined to ensure that the human role in the control room is supported rather than replaced.

8. Conclusion

This research discussed state-of-the-art alarm systems, intervention procedures, and AI-based recommendation systems, highlighting gaps in holistic human-in-the-loop configurations for process control rooms. We also present a set of decision support tools evaluated through an experimental study of a simulated formaldehyde production facility. Through this study, different data collection techniques were used to gain insight into the level of support of supporting tools and the impact of the support tools on the situational awareness, workload and performance of process control room operators.

This study has shown that the Comparison between Group 1 (with paper procedure + no-alarm prioritization) and other groups on alarm list support suggests a notable consistency in the perception of alarm list support across scenarios, which was quite the opposite for alarm priority support. Instead, as expected, there was a statistically significant difference between Group 1 and the other groups in alarm priority. Based on the subjective procedure support ratings, there were no significant differences between Group 3 (the group with the procedure on screen) and Group 4 (the group with AI support). Furthermore, there was no significant difference based on this subjective rating between Groups 3 and 2 (with the Procedure on paper + alarm rationalization). However, the differences between Groups 1 and 2 and Groups 2 with 4 were statistically significant. The almost statistically significant difference between groups 2 and 3 on the support of the procedure suggests the benefit of a more in-depth analysis. Therefore, incorporating the information obtained from eye tracking or EEG analysis of the procedures as an area of interest during the study could provide a more nuanced understanding of the configuration differences.

As the first experimental investigation on this topic and recognizing the constraints associated with assessing a larger control room environment and distributed control panels, our study focused on an academic setting. The scenarios devised for this study were tailored to match task levels that could be tackled by the profiles included in our research—namely, master's and PhD level students and certain professors. The prospect of incorporating a larger panel, control room operators, and more complex scenarios remains a consideration for future research efforts by the authors.

Furthermore, future research could investigate quantitative and qualitative aspects to explore the factors influencing participants' perceptions of alarm lists, alarm priorities, and procedure support in greater detail. More work is needed to analyze the data collected for AI support and the combined impact of the different HITL configurations on situational awareness, workload, and performance assessment. This includes analyzing the data to understand vital performance-

shaping factors that stood out in each configuration, which decision-makers can potentially pay closer attention to.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Consent and ethics statement

Participants were briefed before the study, read a detailed description of what the study entailed via an information sheet, and signed the necessary consent form before participating. These documents and the ethics application were approved by the Internal Ethical Committee of the Collaborative Intelligence for Critical Safety Systems after first approval by the Ethics Review Committee of the Technological University of Dublin, Ireland, with approval number REC-20-52.

Funding

This work has been carried out within the Collaborative Intelligence for Critical Safety Systems (CISC) project. The CISC project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Marie Skodowska-Curie Grant Agreement No. 955901. We thank Rob Turner and Adrian Kelly at Yokogawa and EPRI Europe for contributing to developing the human-system interfaces. Finally, thanks to the participants for the time they dedicated to this study and their feedback.

ORCID

Chidera W. Amazu http://orcid.org/0000-0001-8788-3173
Joseph Mietkiewicz http://orcid.org/0009-0007-3109-5865
Ammar N. Abbas http://orcid.org/0000-0002-2578-5137
Gabriele Baldissone http://orcid.org/0000-0001-7015-8995
Davide Fissore http://orcid.org/0000-0002-0914-7901
Micaela Demichela http://orcid.org/0000-0001-5247-7634
Maria Chiara Leva http://orcid.org/0000-0002-6770-8332

Data availability statement

The data supporting this study's findings are available on GitHub at https://github.com/CISC-LIVE-LAB-3/dataset.

References

Abbas, A. N., Chasparis, G. C., & Kelleher, J. D. (2022). Interpretable input-output hidden Markov model-based deep reinforcement learning for the predictive maintenance of turbofan engines. *International Conference on Big Data Analytics and Knowledge Discovery* (pp. 133–148). Springer. https://doi.org/10.1007/978-3-031-12670-3_12

Abbas, A. N., Chasparis, G. C., & Kelleher, J. D. (2023). Specialized deep residual policy safe reinforcement learning-based controller for complex and continuous state-action spaces. arXiv preprint arXiv: 2310.14788. https://doi.org/10.21203/rs.3.rs-3918353/v1

Amazu, C. W., Abbas, A., Demichela, M., & Fissore, D. (2023). Decision making for process control management in control rooms: A survey methodology and initial findings. *Chemical Engineering Transactions*, *99*(May), 271–276.

Amazu, C. W., Demichela, M., & Fissore, D. (2023). Human-in-the-loop configurations in process and energy industries: A systematic review. *Proceedings of the 32nd European Safety and Reliability Conference (ESREL 2022)* (pp. 3234–3241). Research Publishing. https://doi.org/10.3850/978-981-18-5183-4_S33-04-572-cd

Bhavsar, P., Srinivasan, B., & Srinivasan, R. (2017). Quantifying situation awareness of control room operators using eye-gaze behavior. *Computers & Chemical Engineering*, *106*, 191–201.

Braarud, P. Ø. (2021). Investigating the validity of subjective workload rating (NASA TLX) and subjective situation awareness rating (SART) for cognitively complex human–machine work. *International Journal of Industrial Ergonomics*, *86*(October), 103233. https://doi.org/10.1016/j.ergon.2021.103233

Crompton, J. (2021). Chapter 5—Data management from the DCS to the Historian. In Bangert, P., editor, *Machine learning and Data Science in the oil and gas industry* (pp. 83–110). Gulf Professional Publishing.

Demichela, M., Baldissone, G., & Camuncoli, G. (2017). Risk-based decision making for the management of change in process plants: Benefits of integrating probabilistic and phenomenological analysis. *Industrial & Engineering Chemistry Research*, *56*(50), 14873–14887. https://doi.org/10.1021/acs.iecr.7b03059

Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, *42*(3), 462–492. https://doi.org/10.1080/001401399185595

Energy Institute (2010). *Research report: Human factors performance indicators for the energy and related process industries* (1st ed., pp. 1–79).

Gao, Q., Wang, Y., Song, F., Li, Z., & Dong, X. (2013). Mental workload measurement for emergency operating procedures in digital nuclear power plants. *Ergonomics*, *56*(7), 1070–1085. https://doi.org/10.1080/00140139.2013.790483

Ghosh, K., & Sivaprakasam, G. (2020). Aiding alarm rationalization by automatic identification of various sequential patterns in large volume of alarm and event log data. *IOP Conference Series: Materials Science and Engineering*, *778*(1), 012092. https://doi.org/10.1088/1757-899X/778/1/012092

Horvitz, E. J., & Barry, M. (2013). Display of information for time-critical decision making. arXiv preprint arXiv:1302.495. UAI-P-1995-PG-296-305. https://doi.org/10.48550/arXiv.1302.4959

HPOG Steering Committee. (2021). Best practice in procedure formatting, revision 1. *Human Performance Oil and Gas (HPOG)* (pp. 1–47). https://www.hpog.org/assets/documents/HPOG-Procedure-Best-Practice-Rev1.pdf

HSE (2011). Buncefield: Why did it happen? *Control of Major Accident Hazards* (pp. 1–36). https://webarchive.nationalarchives.gov.uk/ukgwa/20220701173308mp_/; https://www.hse.gov.uk/comah/buncefield/buncefield-report.pdf

Iqbal, M. U., & Srinivasan, R. (2018). Simulator based performance metrics to estimate reliability of control room operators. *Journal of Loss Prevention in the Process Industries*, *56*, 524–530.

Kim, J., Lee, S. J., Jang, S. C., Shin, Y. C., & Ahn, K. I. (2013). Design-related influencing factors of the computerized procedure system for inclusion into human reliability analysis of the advanced control room. *Journal of Nuclear Science and Technology*, *50*(11), 1110–1126. https://doi.org/10.1080/00223131.2013.836065

Kobbacy, K. A., & Vadera, S. (2011). A survey of AI in operations management from 2005 to 2009. *Journal of Manufacturing Technology Management*, *22*(6), 706–733. https://doi.org/10.1108/17410381111149602

Kodappully, M., Srinivasan, B., & Srinivasan, R. (2015). Towards predicting human error: Eye gaze analysis for identification of cognitive steps performed by control room operators. *Journal of Loss Prevention in the Process Industries*, *42*, 35–46. https://doi.org/10.1016/j.jlp.2015.07.001

Leroy, C., Gerjets, P., Oestermeier, U., & Kammerer, Y. (2023). Investigating the roles of document presentation and reading interactions on different aspects of multiple document comprehension. *International Journal of Human–Computer Interaction*, *39*(6), 1327–1340. https://doi.org/10.1080/10447318.2022.2062854

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. https://doi.org/10.1214/aoms/1177730491

Meshkati, N. (2006). Safety and human factors considerations in control rooms of oil and gas pipeline systems: Conceptual issues and practical observations. *International Journal of Occupational Safety and Ergonomics*, *12*(1), 79–93. https://doi.org/10.1080/10803548.2006.11076669

Mietkiewicz, J., Abbas, A. N., Amazu, C. W., Baldissone, G., Madsen, A. L., Demichela, M., & Leva, M. C. (2024). Enhancing control room operator decision making. *Processes*, *12*(2), 328. https://doi.org/10.3390/pr12020328

Mietkiewicz, J., Abbas, A. N., Amazu, C. W., Madsen, A. L., & Baldissone, G. (2023). Dynamic influence diagram-based deep reinforcement learning framework and application for decision support for operators in control rooms. *Proceedings of the 33rd European Safety and Reliability Conference*. Research Publishing. https://doi.org/10.3850/978-981-18-8071-1_P531-cd

Park, J., & Jung, W. (2003). The operators' non-compliance behavior to conduct emergency operating procedures—Comparing with the work experience and the complexity of procedural steps. *Reliability Engineering & System Safety*, *82*(2), 115–131. https://doi.org/10.1016/S0951-8320(03)00123-6

Proudlove, N. C., Vaderá, S., & Kobbacy, K. A. (1998). Intelligent management systems in operations: A review. *Journal of the Operational Research Society*, *49*(7), 682–699. https://doi.org/10.1057/palgrave.jors.2600519

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Shi, C., & Rothrock, L. (2022). Using eye movements to evaluate the effectiveness of the situation awareness rating technique scale in measuring situation awareness for smart manufacturing. *Ergonomics*, *66*(8), 1090–1098.

Simonson, R. J., Keebler, J. R., Blickensderfer, E. L., & Besuijen, R. (2022). Impact of alarm management and automation on abnormal operations: A human-in-the-loop simulation study. *Applied Ergonomics*, *100*(May 2021), 103670. https://doi.org/10.1016/j.apergo.2021.103670

US Chemical Safety and Investigation Board. (2007). Investigation report refinery explosion and fire BP Texas City. *Investigation Report* (pp. 1–341). https://www.csb.gov/bp-america-texas-city-refinery-explosion/

Weidl, G., Madsen, A. L., & Israelson, S. (2005). Applications of object-oriented Bayesian networks for condition monitoring, root cause analysis and decision support on operation of complex continuous processes. *Computers & Chemical Engineering*, *29*(9), 1996–2009. https://doi.org/10.1016/j.compchemeng.2005.05.005

Xu, S., Song, F., Li, Z., Zhao, Q., Luo, W., He, X., & Salvendy, G. (2008). An ergonomics study of computerized emergency operating procedures: Presentation style, task complexity, and training level. *Reliability Engineering & System Safety*, *93*(10), 1500–1511. https://doi.org/10.1016/j.ress.2007.09.006

About the authors

Chidera W. Amazu, a PhD student and researcher assistant at Politecnico di Torino, is part of the Collaborative Intelligence for Safety-Critical Systems network (CISC). Her research centres on the HMI, situational awareness, decision support of control room operators, and safety analysis. She studies operator behaviour during human–machine interaction.

Joseph Mietkiewicz, who has a Master's in Mathematics and ongoing PhD studies at Hugin Expert Denmark and TUDublin Ireland, contributes to the CISC project. Specialising in AI, particularly interpretable machine learning and Bayesian networks, he focuses on developing decision-support systems for control room operators.

Ammar N. Abbas, with a Master's in Mechatronics and pursuing a PhD in Deep Reinforcement Learning, focuses on optimal decision-making in safety-critical systems (detecting anomalies and prescriptive maintenance, optimising product quality, process scheduling.) He utilises Reinforcement Learning as an online method, integrating human expertise with explainable and interpretable AI.

Houda Briwa, a data and knowledge state engineer and recent Data Science Master's graduate from the Polytechnic University of Madrid, currently applies deep learning and Bayesian inference in social science and ergonomics. She aims to build a model for real-time assessment of human performance during human–machine interaction.

Andres Alonso-Perez, pursuing a PhD in Neuroergonomics, collaborates with mBrainTrain to develop deep data representation techniques for electroencephalograms. These methods, integrated with AI models, aim to assess mental workload and address task allocation challenges. Additionally, he explores AI implementation in manufacturing to enhance human collaboration, potentially improving labour conditions.

Gabriele Baldissone, is a Research Technician and Lecturer at Politecnico di Torino, Italy. He educates undergraduate and master's students in advanced control and environmental safety techniques and advanced technologies for risk-based decision-making. His expertise spans risk analysis, assessment, modelling, quantitative analysis, reliability, data mining, and process safety.

Davide Fissore, Full Professor of Process Control and Food Processing Technologies at PoliTo, specialises in process modelling and developing advanced model-based tools for process monitoring and control. His recent research extends to pharmaceutical engineering, focusing on applying the "Quality by Design" concept to optimise, monitor and control pharmaceutical freeze-drying processes.

Micaela Demichela, Full Professor at Politecnico di Torino, is actively involved in teaching and research and holds positions on the Board of Directors for the ESReDa Association and the Scientific Board for the R3C Inter-Department Centre. Her expertise lies in risk management, occupational safety, probabilistic risk assessment, etc.

Maria Chiara Leva, co-chair of the Human Factors Technical Committee for the European Safety and Reliability Association, previously chaired the Irish Ergonomics Society and co-chaired the Symposium on Human Mental Workload. She co-founded Tosca Solutions and currently lectures at TU Dublin. Her expertise spans Human Factors and Safety Management Systems.