Opsummering

I dette projekt udforsker vi anvendelsen af AI-detektering med forklaringer i et multi-drone søge- og redningssystem (SAR). Droner anvendes i en bred variation af applikationer i dag, og den danske beredskabsstyrelse (DEMA) benytter også droner i forskellige SAR-situationer. Vores mål var at teste, hvordan forklaringer fra en AI, der assisterer med at lokalisere savnede personer, kan hjælpe med at nedsætte den kognitive belastning og forbedre præstationen hos SAR personale. Når flere droner arbejder sammen om at lokalisere en savnet person, er der et stort område at søge i, og mange ting at holde styr på, som kræver et godt overblik. Derfor afgrænsede vi rollen som drone operatør og observatør til kun at være observatør. Observatøren skal ikke bekymre sig om hvor dronerne flyver hen eller hvis de skulle styrte ned. De skal kun fokusere på hvad dronerne ser. Vi undersøgte derfor, om de ekstra forklaringer fra en AI ville hjælpe observatøren med at træffe den korrekte beslutning, og om det havde en indvirkning på responstiden. For at teste dette, udviklede vi en online platform, som simulerede fem droner, der flyver i et område og søger efter en specifik person. Hvis en drone opdager et interessant objekt, rapporterer den en alert, i form af en alarmmarkør, der placeres på opdagelsesstedet. Alerts fra alle droner ville poppe op, mens dronerne fløj langs deres ruter, og brugerne blev bedt om at forsøge at besvare så mange alarmer som muligt og besvare dem korrekt. I alt blev der afholdt fire tests, der bestod af både højt og lavt kognitivt belastningsniveau, hver med eller uden AI-forklaringer. Der blev rekrutteret 8 deltagere til undersøgelsen. Deltagerne var primært personer fra DEMA, men vi rekrutterede også personer fra dronevirksomheden Robotto samt fra en social netværksgruppe for professionelle dronepiloter. Vores hypoteser omhandlede primært om der vil være stor variation i svarene, hvilket viste sig ikke at være sandt. Dernæst undersøgte vi om de ekstra forklaringer vil sænke responstiden for hver alert, men gøre deres svar mere præcise. Ved brug af ANOVA kunne vi også afvise denne hypotese. Den tredje hypotese omhandlede at den oplevede arbejdsbyrde ville være lavere med ekstra forklaringer, denne hypotese kunne vi dog også afvise, da der igen ved brug af ANOVA ikke kunne vises nogen signifikant sammenhæng mellem disse. Til sidst forventede vi, at når de ekstra forklaringer blev vist, så ville deltagerne følge et bestemt mønster når de skulle vælge den næste alert at svare på. Ingen synlige mønstre gjorde sig til kende. Langt de fleste alerts blev der svaret *'Ignore'* til, hvilket var overraskende, dog kunne grunden til dette være fordi der blev lagt meget vægt på at det vil koste ressourcer hver gang et objekt skulle inspiceres. Kommentarer fra deltagerne var meget positive omkring at have de ekstra oplysninger, dog kunne det mistænkes at de misforstod dem. Selvom ekstra oplysninger ikke havde en direkte indflydelse på deltagernes beslutningstagen, så har vi præsenteret vores undersøgelser i dette projekt, som understreger kompleksiteten ved at kombinere AI i SAR kontekst, og udfordringen ved at undgå at automatikken sker på bekostning af den individuelle erfaring.

Explainable Alerts for Drone Swarm-based Search and Rescue: How Information Detail Impacts Performance

ANDREAS SKJOLDGAARD ANDERSEN, Aalborg University, Denmark
PHILIP MICHAELSEN, Aalborg University, Denmark

The use of drone swarms for search and rescue is being increasingly explored. Because of the time-sensitive and life-critical nature of these missions they are very mentally demanding of the search and rescue operators. It is therefore apparent that incorporation of artificial intelligence at the right levels will be a deciding factor in the effectiveness of these systems. We set out to investigate how explanations about object detections made by the drone swarm can be used to improve the performance of operators engaged in search and rescue missions. We conducted an online study with 8 participants involved with the Danish Emergency Services, in which they were tasked with responding to AI-generated alerts under varying workload while being provided distinct levels of explanation detail. A combination of performance measures and subjective measures showed that under high cognitive load, participants become significantly faster at responding to alerts, and that this increase in speed is not necessarily at the cost of accuracy. Our findings also provide insight into the challenge of mitigating a drop-off in user expertise due to over-trust in the AI. We discuss the findings and provide implications for the design of alerts for search and rescue.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; *Graphical user interfaces.*

Additional Key Words and Phrases: Drone swarm, Automation, Alerts, Search and Rescue, Explanations

ACM Reference Format:
Andreas Skjoldgaard Andersen and Philip Michaelsen. 2023. Explainable Alerts for Drone Swarm-based Search and Rescue: How Information Detail Impacts Performance. In *AAU'23 - Master Thesis 2023 - Aalborg, DK.* ACM, New York, NY, USA, 27 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The use of unmanned aerial vehicles (UAVs) such as drones for search and rescue (SAR) is being increasingly targeted by researchers because of the drones' ability to speed up the search process, thereby reducing the time to find people [3, 25, 39]. A recent research focus has been to use swarms of drones, with the intended benefit of being able to cover a large area even faster [9, 11, 16, 34]. In order to realise the use of a drone swarm, increasing levels of automation are needed because each drone can't be controlled directly at all times. Interfaces for how to control the drone swarm and view live video generated by each drone have recently been suggested [11, 16]. Considering the benefits of coordinated swarms of drones, SAR operations could be improved with faster coverage of large areas, however, there is little research on interfaces to support higher levels of automation such as automatic identification of objects, and it is unclear how the operator can maintain a sense of control and at the same time benefit from the automated swarm behavior. These interfaces need to carefully balance control and automation in order to serve as a tool that operators with expertise can utilize, rather than be the complete solution.

Manuscript submitted to ACM

This work is a part of the HERD research project[1] which aims to allow an operator to deploy and control a swarm of drones that will fly semi-autonomously, and may perform automatic detection of objects of interest. Large steps have already been taken in this direction. In close collaboration with the Danish Emergency Management Agency (DEMA)[2] and local drone developer Robotto[3], suggestions have been made for a SAR oriented system for control of a drone swarm and an interface for viewing live video feed [11, 16]. System autonomy in regards to the identification of people in distress, and more specifically, how operators will interface and interact with the system once those detections happen is still highly uncertain.

In order to investigate how AI-generated alerts can be used to improve the performance of drone swarm operators during searches we conducted an online study to examine performance when responding to AI-generated alerts. The study looked at situations where users would be shown minimal details about the alert versus detailed information about what triggered the alert, while also experiencing low and high levels of cognitive load. The study was conducted with domain professionals from DEMA and Robotto. The main findings in this paper indicate that providing explanations for AI detections did not improve the speed or the accuracy of participants' responses. However, the findings do indicate that participants still found explanations useful for helping to make the decision of inspecting or ignoring an alert, even over-relying on them in some cases.

The main contributions of this work are: (1) We provide implications for the design of AI-generated alerts in SAR; (2) We provide insights into how explanation detail in alerts can impact the performance of SAR operators; (3) We develop a research platform that allows for simulation and interaction with any number of drones.

2 RELATED WORK

In this section, we examine related research on drone swarms for SAR purposes and discuss how it ties together with literature on how to design alerts that help to improve user performance and understanding without overwhelming them. We also illuminate existing research in the field of human-AI collaboration and investigate what practices can be used to guide alert design toward overcoming the challenges in a drone swarm SAR system.

2.1 Search and Rescue Multi-UAV systems

Using drones for SAR operations is becoming increasingly widespread in the domains of maritime SAR [22, 28] as well as ground based SAR [9, 16, 34]. One immediate benefit is that large areas can be searched more quickly, especially with a swarm of drones. Arnold et al. examine the challenge of optimizing the coverage of a search area [6]. They do this by extensively simulating swarms of up to 50 drones that are given different simple behaviors. They find that, with the optimal role distribution, the swarm is able to locate all missing people in a four square kilometer area in 40 minutes.

As of the writing of this paper, there is no universally agreed upon best way to interface with and control the swarm of drones. A tablet-based approach for mission planning and real-time swarm control is presented by Hoang et al. Through a co-design session with domain experts from DEMA they find that some level of automation is desired for supporting e.g. path planning and adjusting the camera angles according to the terrain. However, it is also evident that automation in such a system should serve as a tool to support the human operator rather than replace them [16]. We recognize the role of automation as a supportive tool rather than the complete solution. With respect to situational awareness, challenges associated with a single UAV differ significantly from a multi-UAV emergency response environment [3].

[1] https://direc.dk/herd-human-ai-collaboration-engaging-and-controlling-swarms-of-robots-and-drones/
[2] https://www.brs.dk/en/
[3] https://www.robotto.ai/

Christensen et al. study how to present video feeds from drone swarms to SAR operators while maintaining their situation awareness among other things. By performing a field study in collaboration with domain experts they find that video information overload is one of the main challenges to overcome, and that an operator at most can keep track of three video feeds at a time. They also find that participants wish to always be able to see what the drones are doing on the map because of a lack of complete trust in the new technology [11]. For these reasons, it will be beneficial to consider the use of a designated alert handler role separate from the one that controls the drone swarm.

While a lot of work has focused on the control mechanisms of multi-robot systems, an important part of these systems will entail how information, such as status updates, should be conveyed to whoever is responsible for it. Specifically in terms of drone swarms for SAR, we want to investigate how a user should be alerted to real time detections made by the drones.

2.2 Alert Design

Alerts and notifications are common in many contemporary information systems. However, work on the design of alerts in drone SAR systems is sparse. Challenges that are faced in air traffic control (ATC) notification systems such as high false alarm rates [38], attentional tunneling [19], and the inability to distinguish notification cause [21], are all expected to also be obstacles in alert systems for drone swarm SAR. To overcome these issues, researchers have suggested the use of likelihood alarms, that self-report their own confidence [38], utilization of designs that draw visual attention to import areas [19], as well as encoding notifications with categorical information about what caused them [21]. Although the tasks that are performed in ATC and SAR are different, there is a distinct similarity in the mental demands that those tasks place on the person performing them. Therefore, it would be reasonable to consider practices that are known from ATC as the first step toward developing system for drone swarm SAR.

In [3], Agrawal et al. focus on improving situational awareness in multi-UAV emergency response applications, as well as testing different ways to communicate with firefighters through their interface, which include notifications of when and where a drone has observed a drowning person. Domain experts express a need for imagery to be full screen and related buttons to be closely adjacent, which is also supported by the findings of Christensen et al. [11]. In [36], Van Berkel et al. study seven unique visual marker designs for AI detection notification during colonoscopy. One of the designs only displays the AI confidence score without a visual border around the detection, but participants did not find this design useful by itself, rather it was more confusing especially when managing false positives. The authors did not have time to test the different visual marker designs in combination with each other, but it is speculated that this would improve usefulness. Following these practices will be essential to ensure that the interfaces accommodate the challenges faced by real SAR operators.

2.2.1 Alert Placement. An important part of a multi-drone SAR system is to design how the alerts should appear to a user. In [11], domain experts suggest having a static bar at the top of the screen that flashes red whenever a drone has detected something. However, it is not immediately obvious how such an implementation would scale as drone numbers and alert pop-up frequency increase. Müller et al. study the relation between desktop visual importance and notification noticeability and discuss how notification placement with respect to the visual importance of the background can allow for more freedom of notification design without sacrificing noticeability. They also suggest that by taking user attention into account the quality of notification placement can be increased, which has a significant impact on noticeability [26]. Different kinds of notification types are studied in [40]. They find that notifications generally have a positive effect on player performance and that having icons in notifications is the most effective in getting the player's attention. Based

on these findings, drone swarm SAR interfaces should ideally place notifications near their corresponding drone as this will be the most visually important place on the map.

2.2.2 Alert Fatigue Mitigation. Depending on the sensitivity of the object detection algorithm, there is a risk of having a substantial amount of alerts appearing within a short time frame. Overly frequent or excessive use of alerts can lead to alert fatigue, which in turn can cause users to miss critical information [23]. Approaches for mitigating the occurrence of alert fatigue in medical scenarios are presented in [23]. These include clustering of alerts, checking for false positive alerts, adjusting alert design based on their severity, and delaying non-critical alerts. The authors of [24] explore the effects of notification intensity scaling on highly physical task performance. They find that, in some cases, more intense and obtrusive notifications do succeed in improving task performance. In the case of object detections from a drone swarm, it might be unwise to suppress or delay alerts, however, adjusting their design is one of the techniques that will be explored.

2.3 Human-AI Collaboration

Making use of AI disciplines like computer vision to enable fast detection of people and objects, path planning algorithms to cover an area quickly, or explainable AI to allow for better comprehension of AI decisions could markedly improve the effectiveness of SAR teams. Domain experts have expressed that a SAR drone swarm system would need to incorporate AI detection software if it is to really be useful for them [11, 17]. Uncertainties regarding the intrusiveness of the software among many other things are however still prevalent.

2.3.1 Computer Vision. Given the success that computer vision applications have had in areas such as video surveillance [8] and aerial imagery [4], it is no surprise that the potential benefits of its inclusion in drone-based SAR applications are being explored. Sambolek and Kos investigate the reliability of state-of-the-art detectors in SAR situations [30]. Initial tests show that the YOLOv4 model performs best in terms of detection speed and accuracy. After performing further tests of YOLOv4 on a self-made image dataset of people in SAR situations they find that on average 6% of detections are false positives. They argue that this is acceptable since the most important thing when searching for a missing person is that the detector locates that person, and it is less important how accurate it is. Because of the time-sensitive nature of many SAR missions, there exists a trade-off between speed (possible at high altitudes), and detection capabilities (better at low altitudes). In a maritime SAR setting, this relationship is studied by Qingqing et al. in [27]. Results indicate that the proportion of false negatives (instances where the algorithm fails to detect a person when it should) remains low even at high altitudes, but there is a significant drop-off in accuracy at around 100 meters. It is however not guaranteed that there will be the same low proportion of false negatives for land-based detections, and studies such as ours that target land-based SAR should not expect these results to be reflected. We also recognize that computer vision algorithms for SAR must never miss a person even if that causes a high number of false positive detections to occur. Lastly, we point out that, while great strides have been made to merge computer vision with SAR drones, there is a distinct lack of knowledge on how to present the output of that system to a human operator.

2.3.2 Explainable AI. Deep learning models, such as image classifiers, are generally considered black boxes, as their complexity makes it inherently difficult to understand what makes them arrive at their predictions, and a demand for more transparency is increasing [2, 7, 14]. At its core, eXplainable Artificial Intelligence (XAI) is about enhancing transparency and understandability by uncovering and presenting these otherwise hidden learned predictors [7, 31]. The following formal definition of XAI is given by Barredo et al. in [7]: *"Given an audience, an explainable Artificial*

Intelligence is one that produces details or reasons to make its functioning clear or easy to understand". This definition highlights understandability and clarity as the central purpose of XAI while also implying that different audiences can have different XAI needs, without making claims about what exact shape the details or reasons take.

Robot explanations and their impact on the effectiveness of human-robot teams are researched by Ezenyilimba et al. [13]. They seek to identify useful levels of robot transparency and robot explanations. Context-driven and readily available robot explanations are found to be a driving factor in effective human-robot teams. Additionally, both transparency and explanations were found to improve trust in the robot [13]. In [29], Rader et al. study the relationship between explanations and algorithmic transparency, viewing transparency as a way of preventing negative effects of complex algorithmic decision-making systems. They differentiate between *How* and *Why* explanations. *How* explanations provide information about how a system produces a certain output. *Why* explanations provide justifications for the output of a system, without providing any visibility into how the system works. Results show that some positive effects of transparency are often due to brief, easy-to-read explanations.

AI explanations have recently been used to reduce cognitive load [1, 20], and to improve usability in terms of the principles of transparency [29] and control [18], studied by including or excluding explanations. Ambiguities do however exist around the effects of the technique. A meta-analysis by Schemmer et al. finds no significant difference between AI-assisted and XAI-assisted performance throughout the literature [33], which indicates that it is not a "one-size fits all" method, but that it is more context-dependent. The authors of [37] study the behavior and experience of clinical professionals responding to annotated colonoscopy videos. Results show that AI recommendations significantly slow down decision making. Zhang et al. study the effect on the overall performance of showing confidence and local explanation in AI predictions. Results show that confidence information can improve trust calibration in AI-assisted decision making, however, it had little effect on decision outcomes [41]. For SAR purposes, explanations should therefore aim to be brief, context driven, and easy-to-read, while also providing a level of detail that improves SAR personnel's understanding of why the AI made an alert appear.

3 RESEARCH PROBLEM

Much recent work has investigated control methods for operation of SAR drone swarms [9, 11, 16, 34]. There has also been much focus on how to better support collaboration between humans and AI in decision-making systems [10, 13, 37, 41]. However, research on the next step of fusing AI with SAR drone swarms and their operators, and the associated challenges regarding cognitive overload and trust issues [11], is lacking. That fusion can be divided into three parts that are all accomplished with the help of computer vision: object detection, object recognition, and object tracking. Figure 1 shows an example of a typical object detection and recognition. While object tracking is very important in contexts such as autonomous driving [5], it is not well understood what role it will play in SAR. Object detection and object recognition, however, are already known to be essential and a way for an operator to interface with them is through alerts.

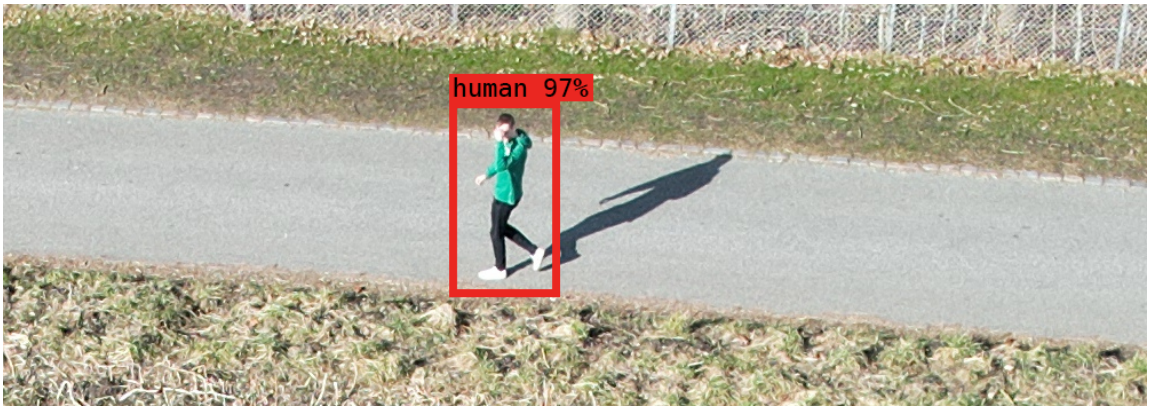

Fig. 1. Detection and recognition of a human by an AI. The red bounding box represents the detection, and the text "human 97%" represents the recognition. The percentage conveys how confident the AI was in its recognition.

To mitigate cognitive overload, and ultimately improve performance, we postulate that an AI-based detection alert system for SAR drone swarms must employ the use of explanations and specialized alert design. Due to the time-sensitive nature of many SAR missions and the potentially life-critical consequences of failing to notice a missing person, performance in this paper consists of speed and accuracy. However, the two are not equally important, and missing critical information is considered much more severe than taking a long time to respond. We, therefore, ask the question: ***How can explanations be designed to mitigate cognitive overload and improve the performance of SAR personnel responding to AI-generated alerts?***

To answer this question, we will design alerts with and without explanations, for SAR drone swarms, and evaluate their effectiveness with domain professionals. The intent of the study will primarily be to understand the impact of explanations on performance and user preference under high and low cognitive load. We also wish to gain insight into any habits, strategies, or patterns that domain professionals might consciously or subconsciously follow when responding to alerts. At a meeting with one of the officers in charge at DEMA, we were told that SAR drone operators have no defined guidelines for when to inspect an object further, and they are often encouraged to follow their intuition.

We put forth the following hypotheses:

(H1) Because SAR drone operators don't follow common guidelines, we hypothesise that for many alerts, both with and without explanations, there will be high variation in the response choices.

(H2) Based on the findings of [13] and [37], we hypothesise that explanations will slow down decision-making time, but will increase response accuracy.

(H3) Based on the findings of [1] and [20], we hypothesise that perceived workload will be lower when explanations are included.

(H4) We hypothesise that explanations will make participants follow common strategies based on the AI confidence, when deciding which alert to respond to next.

4 STUDY DESIGN

In order to study the effectiveness of providing explanations we recruited domain professionals to take part in an online study, in which they were asked to respond to AI-generated image detection alerts from drones under different

conditions with 2 independent variables: cognitive load and detail of explanations. The study is conducted as a 2x2 within subjects design, an overview of which can be seen in Table 1.

	Without explanations	**With explanations**
High cognitive load	HiLoadNoXAI	HiLoadWithXAI
Low cognitive load	LoLoadNoXAI	LoLoadWithXAI

Table 1. The independent variables of the study include the detail of explanations and the level of cognitive load.

To regulate high and low cognitive load we varied the number of alerts that are shown per minute. These numbers are 6 for low cognitive load and 14 for high CL. The values were found through testing with two colleagues. Low cognitive load was determined to be present when all alerts could comfortably be responded to within the two and a half minute time span. High cognitive load was determined to be induced when there were still a considerable amount of unopened alerts left after two and a half minutes.

4.1 Participants

Data collection took place from May 11–26, 2023. 13 people, all living in Denmark, registered for the study, and of those, 8 participants (8 male, 0 female) finished all tasks. Participant age spanned between the ages of 26 and 64 years old (*mean*=39.4, *SD*=11.1). 5 participants were recruited from DEMA, 1 was recruited from the SAR drone specialist company Robotto, 2 were recruited from a social network group for professional drone pilots. Participant experience with SAR ranged from 0 to 26 years (*mean*=7, *SD*=7.76), and experience with drones ranged from 2 to 10 years (*mean*=5.6, *SD*=2.78). None of the participants reported any issues with understanding the system or tasks they were given.

4.2 Alert Design Rationale

When developing the design of an alert it was important for us to specialize it towards the SAR setting and make an attempt to address the challenges that we know exist there. We used the following principles to guide design of alerts:

- *Task dedication.* Related research has shown that currently, the SAR drone swarm operators who keep track of the drones and the ones who look at video feeds experience a high level of mental workload from those tasks alone [11]. For that reason, the alert design would revolve around having a designated alert handler role separate from the one that controls the drone swarm. Alerts were then free to take up as much space and attention as necessary, and therefore the alerts we used take up the entire screen, as also suggested in [3] and [11].
- *A time-sensitive task.* Due to the time-sensitive nature of the task, it was important to design alerts with that in mind. Thus the explanations that we used were meant to be brief and easy to read, which has been shown to also have a positive effect on transparency [29].
- *Context captured by placement.* To provide more awareness and a greater sense of realism we wanted alerts to be placed by the drone that made the detection and we wanted the detection image to match its location on the map. Placing alerts near visually important areas has also been shown to improve noticeability [26].

- *Information at each level.* We wanted to give alert handlers the opportunity to make informed decisions about the response choice as well as the process of selecting alerts. Inspired by [23] and the idea of adjusting alert design based on their severity, we provide a preview on the map of the information that is contained at the full-screen level of an alert.
- *Easy access to mission objective.* While using mental effort on deciding how to respond to alerts, alert handlers may experience the situation awareness demon of memory overload [3], and forget the details of the person they are looking for. For this reason, we included the missing person description on every alert.
- *Inspiration from ATC practices.* By using design practices from fields that share some of the operator demands of drone swarm SAR it could help to mitigate the challenges that they share. Therefore we developed a design that drew visual attention to important areas [19], encoded alerts with information about what caused them [21], and self-reported its own confidence [38].

4.3 The Alert System

In order to make it easier for domain professionals from all over Denmark to participate, we implemented an alert system as part of an online website from where the entire study would be conducted. The language of the website is Danish, but we will translate any content into English in this paper. All participants used the system on their own laptop or desktop.

Each condition contains five simulated drones searching an area. All five drones start from roughly the same spot and continue to move throughout the session, covering approximately 800m each. Each drone path is a manually drawn linestring that is made to look like it follows areas of interest (e.g. treelines, streams, and lakeshores) to provide more psychological realism. To simulate a drone flying, the linestring is drawn at roughly five meters per second, which we know from our contacts at DEMA is reasonable (during real SAR missions, drone speed depends on altitude, but altitude is not simulated in this system). The head of the linestring is therefore meant to represent the drone. The tail of the linestring remains drawn and serves to represent the trail that the drone has already flown. Figure 2 shows what a single drone and its path look like.

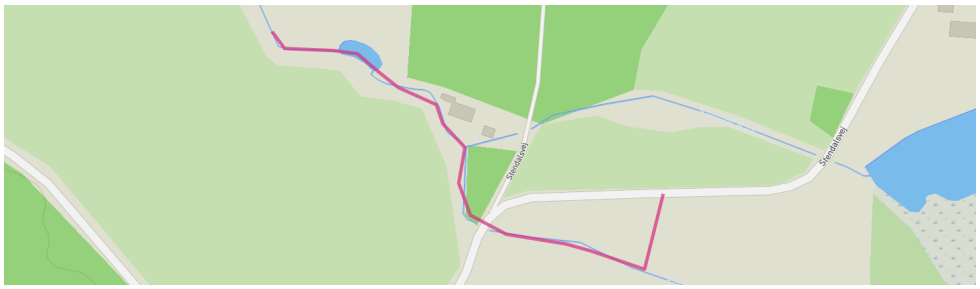

Fig. 2. A single trace from a drone. The path is being drawn from right to left.

At any given moment an alert pin may appear at the head of the path. Alerts appear randomly rather than at set intervals, but they are distributed along the full length of each path. The alert pin can be clicked to reveal an alert page, which contains an image of the detection that caused the alert, a red bounding box around the object that was detected, as well as a sidebar on the left. This sidebar is where the two buttons for responding to the alert are located, as well as an itemized description of the missing person, based on the principle of *Easy access to mission objective*. For the two

conditions that include explanations, the sidebar is where they will be presented. Those explanations are: the type of object that has been detected (Human, Clothing, or Trash), a percentage projecting how confident the AI is that the object is of the displayed type and a list of the colors that make up the object. These three explanations were chosen and designed based on the design principles of *Inspiration from ATC practices* and *A time-sensitive task*. The confidence scores range from 50% to 99%. Additionally, based on the design principle of *Information at each level*, if explanations are included, the alert pins that appear on the map will be shaded based on the AI confidence of each alert. An example of alert pins and an alert page for a condition that includes explanations can be seen in Figure 3. All explanations as well as the red bounding boxes were set manually by one of the researchers and then subsequently checked by the other, where any disagreements were discussed and adapted.

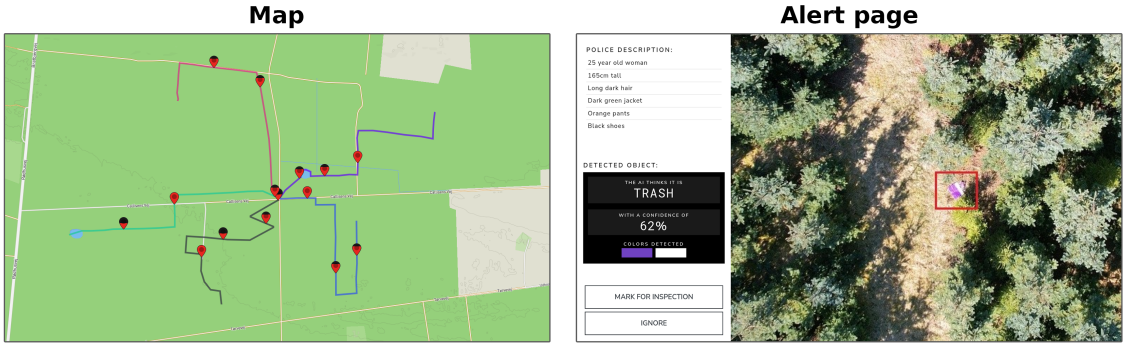

Fig. 3. The map (left) displays the paths of the drones along with the corresponding alert pins. The confidence levels of the alert pins are indicated by a varying degree of red. The alert page (right) shows the information panel and the detection image. In the information panel the missing person information is at the top, the AI explanations are in the middle, and the two buttons that are used to respond, are at the bottom.

One challenge we faced when designing the alerts was what to call the two response options. Initially, they were called **Acknowledge** and **Dismiss**, but internal pilot studies revealed that those labels made users have a tendency to respond based on how closely the explanations matched the detection picture. Consequently, they were renamed to **Mark for Inspection** and **Ignore**, and more space was dedicated to explaining their function and the impact that choosing either would have on other SAR resources. Each study condition is in a different location on the map, but they all include roughly the same proportion of the natural features that were used to categorize the images. Those categories are field, stream, lake, and forest. Figure 4 shows an example of each image category that was used in the study. In total, 142 unique images were used; 100 images for the conditions and 42 images for familiarization with the system.

All images were taken at a resolution of at least 1920x1080, but each of them was later cropped to make the red bounding box sit closer to the center and be noticeable. The average resolution of all used images is 570x380, the condition LoLoadWithXAI has the lowest average image resolution at 501x334, and the condition LoLoadWithoutXAI has the highest average at 612x408. All images were taken in typical Danish nature settings in the north of Jutland, but not in the exact locations of the study conditions. However, they were manually assigned to each alert and made to roughly match the alert's location on the map. All images are unique. Some images are similar and of the same object, but those are taken from different angles and often mirrored or rotated to make them less recognizable. Mirroring and rotating could easily be done without making images look strange. This is because most of them were taken with a

Fig. 4. Images from the four categories of natural features used as backdrop for detections.

near vertical top down view, a common practice in drone SAR literature [25, 34], and also how the operators at DEMA angle the drone camera during sweeping searches. To avoid one condition's detection images being significantly more difficult to respond to than others, we selected each set of images from the same superset. We used clothing, trash, and humans as the objects of detection. Some detection images were staged to be of obvious non-missing people and obvious false detections. These alerts were intended to be easy to respond *Ignore* to, and they were made to simulate the inaccuracies of real AI detection algorithms. Three examples of detections that were considered easy to respond to can be seen in Figure 5. Of the 100 images used for the four conditions, 24 were made to be easy, and of those, 10 were deliberate false detections, which is three more than the false positive ratio of state-of-the-art person detector YOLOv4 for images of nearly the same resolution [30].

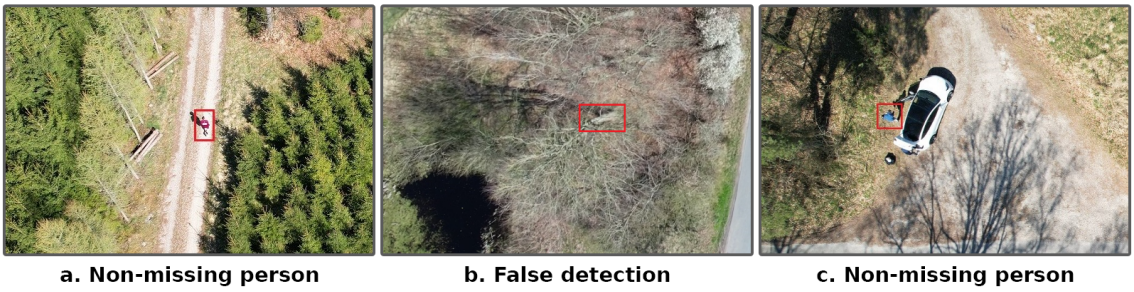

Fig. 5. Three alerts that are considered easy to respond to, given that the missing person description involves a man wearing all black clothes. The detection in image a. is of a jogger, b. is of a tree trunk, and c. is of a person standing by their car.

In order to measure the response error rate, each alert is associated with a ground truth. Because of the fact that for many alerts it can be difficult to see what exactly an object is from the image alone, and because operators are encouraged to follow their intuition, there are many alerts where strong cases could be made for both response options. Therefore ground truths are labeled either *Mark for Inspection*, *Ignore*, or *Ambiguous*. All ground truths were manually set by the researchers. This was done by first having both researchers independently note down their favoured ground truth for each alert, based on the guidelines that participants were shown. Then, for the alerts where the two researchers agreed, the agreed ground truth was assigned. Where there were disagreements the ground truth was labeled as *Ambiguous*. Figure 6 shows a missing person description and an image for each of the three ground truths and gives an explanation for why each image was given the particular ground truth. Of the 100 images used for the four conditions, 28 were assigned the ground truth *Mark for Inspection*, 36 were assigned *Ignore*, and 36 were assigned *Ambiguous*. Having a relatively lower proportion of "correct" images was chosen because the most important thing when searching

for a missing person is that the algorithm locates that person, and it is less important how accurate it is. [30]. Therefore we expect such algorithms to make more detections that can be ignored than must be inspected.

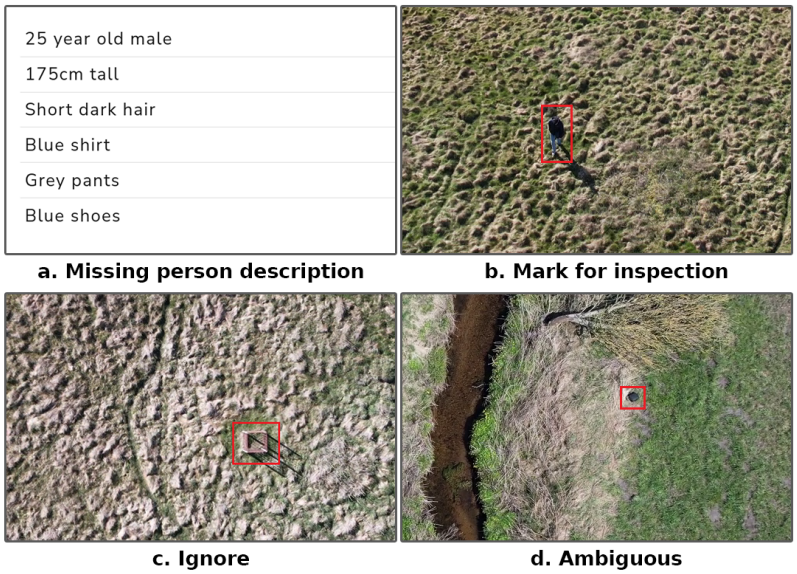

Fig. 6. The three types of ground truth, given only the missing person description in sub-figure a. Image b. is labeled *Mark for Inspection* because there is a close color match between the object and the missing person description. Image c. is labeled *Ignore* because there is no close match in the colors, and because it is clearly a random object. Image c. is labeled *ambiguous* because even though there is no direct match in the colors, it is not certain that the object is not related to the missing person.

4.4 Procedure

Each participant conducted the study online on their own laptop or desktop, without any of the researchers present. The total completion time was estimated at around 30 minutes, but participants were given as much time as they needed to read descriptions and instructions. They were expected to take this during their normal working hours, but it was not presented as a formal requirement.

Participants received a study invitation via email and clicked on a link that directed them to the study website. The website contained five core parts in sequence: registration, study introduction, familiarization, tasks, and post-study questions. Participants were first informed about what type of equipment was recommended in order to take part in the study. They were instructed to use either Google Chrome or Mozilla Firefox, the use of a mouse was requested, and a screen scaling of more than 125% was prohibited. Participants were also instructed to connect their power supply if they were using a laptop and finally, they were instructed not to use a phone or a tablet, only a laptop or desktop. Following this, they were presented with a consent form, and after accepting to take part in the study they could register. Participants had to enter their age, gender, drone experience, SAR experience, and participant category to move on. For the participant category, they could choose one of the following options: SAR professional (DEMA), Professional drone developer (Robotto), Volunteer investigator, and Professional drone pilot (Facebook group).

Next, participants read a short introduction to the purpose of the study and the steps that they would go through from start to finish. It is worth noting that high and low cognitive load was never explicitly mentioned. Rather, participants

were only told that they would be experiencing varying levels of difficulty. On the next page, they were given a brief explanation of the alert system including how drones are represented and what alerts look like. The text was accompanied by a GIF that showed a couple of seconds of drone flight until an alert pin appears, at which point a cursor presses the alert to make an alert page containing a detection image appear. After that, they saw a figure of the general layout of an alert page, as well as some descriptive text. Next, participants were given a thorough explanation of the two response options, where their differences and the consequences of pressing either one were emphasized. For example, pressing *Mark for Inspection* meant that SAR colleagues are asked to go take a look at the location of the alert, and that can mean that resources may be taken away from other tasks. On this page, they were also provided a few examples of when to definitely choose either response option. In the final step of the introduction, they were provided instructions on the task they would be asked to perform. They were told that they would take on the role of the alert handler and that it was their responsibility to observe the map and respond to any alerts that appeared. They were also asked to try to respond as fast and accurately as possible due to the time-sensitive nature of SAR missions.

After the introduction and task description, participants needed to familiarize themselves with the system. They were given two demos, one with explanations and one without. The order of the demos were switched for each participant. Before they got started with each, they read a short description of the content of alerts in the specific demo, as well as a reiteration of what to focus on. After having responded to five alerts they could proceed, or they could stay and respond to more alerts if they did not feel ready to move on.

Next, participants would perform the task for each of the four conditions. The ordering of the four conditions followed a balanced latin square. Before each condition, participants again read a short description of what the response options entailed, as well as what they should base their response on. They were also presented with a list describing the missing person they were looking for. The list contained: age and gender, height, color and length of hair, and color of shirt, pants, and shoes. Once participants pressed "Start test", they had two minutes and thirty seconds to respond to alerts. When an alert had been responded to it remained on the map and was given an icon that indicated what response was chosen. At any time, participants could re-open an alert and change their response. When the time ran out, or when they had responded to every alert, a prompt appeared that took them to a post-condition evaluation. First, they used NASA-TLX to assess their own experience, and then they were asked to evaluate perceived usefulness using a five-point Likert scale. Then, participants moved on to the next condition, where they read the same short task description to remind them what their task was. The missing person description, as well as the map location, changed for each condition. Finally, after participants had been through all four conditions, they were asked to choose which alert design they preferred, as well as encouraged to provide comments on why they had that preference and if they had suggestions for improvements.

4.5 Measures

During the study, we collected four objective measures: response variation, alert response time, response error rate, and alert selection order. To measure response variation and response error, each time a participant made a response we logged the choice that was made together with the alert that it belonged to. On that same data entry we logged the time when the alert was selected and the time when a response button was pressed, this time interval gave us the response time for each alert that a participant answered. Logging the selection time also gave us the data we needed to analyse the order in which a participant selected each alert.

We also collected data on three self-reported subjective measures: perceived workload, perceived usefulness and design preference. After each condition, participants assessed their perceived workload using NASA-TLX. Two modifications

were made: we excluded the subscale regarding *Physical demand*, and we eliminated the weighting process in order to limit the overall study length. Both of these are common modifications made to NASA-TLX [15]. After each NASA-TLX assessment, we used a five-point Likert scale to evaluate the perceived usefulness of the alert design from a set of three to five condition-specific statements, inspired by the Technology Acceptance Model (TAM) [12]. These statements can be seen in Table 2. The study ended with a forced-choice question regarding which alert design was preferred, and two comment boxes for explaining the preference and suggesting improvements.

Explanations	Statement
With	The supporting information helped me understand why an alert was triggered.
With	I found the supporting information useful when responding to an alert.
With	Knowledge of the algorithm's confidence helped me decide what to respond.
With	Knowledge of the type of object the algorithm had detected helped me decide what to respond.
With	Knowledge of the colors of the object the algorithm had detected helped me decide what to respond.
Without	The red box was helpful for me when responding to an alert.
Without	It was helpful to see the person description on each alert.
Without	The image and the red box were sufficient to let me respond to an alert with confidence.

Table 2. Statements used to evaluate perceived usefulness of the alert design. The "With" statements were presented after the conditions LoLoadWithXAI and HiLoadWithXAI. The "Without" statements were presented after the conditions LoLoadWithoutXAI and HiLoadWithoutXAI.

5 RESULTS

To determine the effects of including alert explanations, we set out to measure alert response variation, alert response time, alert response error rate, perceived workload, perceived usefulness, alert design preference, and alert selection patterns. We collected 677 data points in total. Aggregating the data for each participant gave us 32 evaluations considering that each of the 8 participants interacted with all four combinations of cognitive load levels and explanation details

Only one participant managed to respond to all 100 alerts. The participant with fewest alert responses had 37 and the average for all participants was 84.38 (SD = 18.8). Table 3 shows a more detailed overview of each condition, which further indicates that participants managed to respond to a high number of the available alerts. One participant who responded to less than half of the average stood out from the rest. In fact, the values of the *Minimum* column all equal that participant's response counts. The overall high number of responses was surprising and contrasted internal pilot testings.

Condition	Total alerts	Avg. response count	Minimum	Maximum
LoLoadWithXAI	15	13.62	6	15
LoLoadNoXAI	15	13.62	5	15
HiLoadWithXAI	35	30.25	16	35
HiLoadNoXAI	35	26.88	10	35

Table 3. Alert response count for each condition.

Participants were generally very critical of the alerts that they saw. Table 4 shows, for all data points, that on average 22.6% of alerts were marked for inspection, and 77.4% (SD = 11.25) were ignored. The average percentage of alerts that were marked for inspection is slightly lower than the percentage of alerts with ground truth *Mark for Inspection* (28%). This could mean that participants were naturally very critical in their assessments, or it could have stemmed from the described consequences of choosing *Mark for Inspection*. This is also evident when looking at the *Ambiguous* column, which shows that 88% of alerts labeled as ambiguous were ignored.

Response \ **Ground truth**	Ambiguous	Mark for Inspection	Ignore	Total average
Mark for Inspection	12%	52.24%	7.96%	22.6%
Ignore	88%	47.76%	92.04%	77.4%

Table 4. Alert response distribution for each ground truth and in total.

5.1 Response Variation

Knowing when to investigate something further is a learned practice that takes experience and tacit knowledge. Therefore we wanted to measure variations in alert responses to understand whether explanations made participants agree more or less. We define an alert with high variation in its responses as one which has an agreement percentage of less than 70%. In order to avoid making generalizations from alerts that only have one or two responses, this section only considers alerts with three or more responses. Out of the 100 total alerts, 98 had three or more responses.

Agreement (%)	**LoLoadWithXAI**	**LoLoadNoXAI**	**HiLoadWithXAI**	**HiLoadNoXAI**	**Total**
[90,100)	5	7	17	18	47
[80,90)	5	4	6	7	22
[70,80)	2	1	6	3	12
[60,70)	0	2	3	3	8
[50,60)	3	1	3	2	9

Table 5. Number of alerts with a given response agreement percentage for alerts with three or more responses. An alert with e.g. 90% agreement is one where 90% of the participants that responded to it made the same choice.

In Table 5 it can be seen that participants generally agreed on a large set of alerts. 69 of the alerts had an agreement rate of 80% or more and only 17 alerts fall under the definition of high variation. The general tendency to ignore alerts, as shown in Table 4, was surprising and may also have cascaded down through the variation in responses, causing much agreement in the way that participants respond to alerts regardless of explanation detail, thus rejecting **H1**.

5.2 Performance

In order to understand the effect that cognitive load and explanation details have on participant performance we collected the time it took for participants to respond to each alert and the response that they gave each alert. For each condition, we measured participants' performance by two indicators:

1) *Alert Response Time*, the average time it took to respond to an alert. In cases where a participant opened the same alert multiple times, the response time was an aggregation of the time that the alert was open.

2) *Response Error Rate*, the percentage of alerts with ground truth *Mark for Inspection* or *Ignore* where the response does not match the ground truth. Making errors on alerts where the ground truth is *Mark for Inspection* (a false negative) can be severely more consequential than the opposite case. We therefore present the error rate of *Mark for Inspection* and *Ignore* separately. The main difference between the two indicators is that, generally, the consequences of taking longer to respond are less severe than the consequences of making a false negative response. Therefore, we consider the response error rate to be more impactful on performance than alert response time.

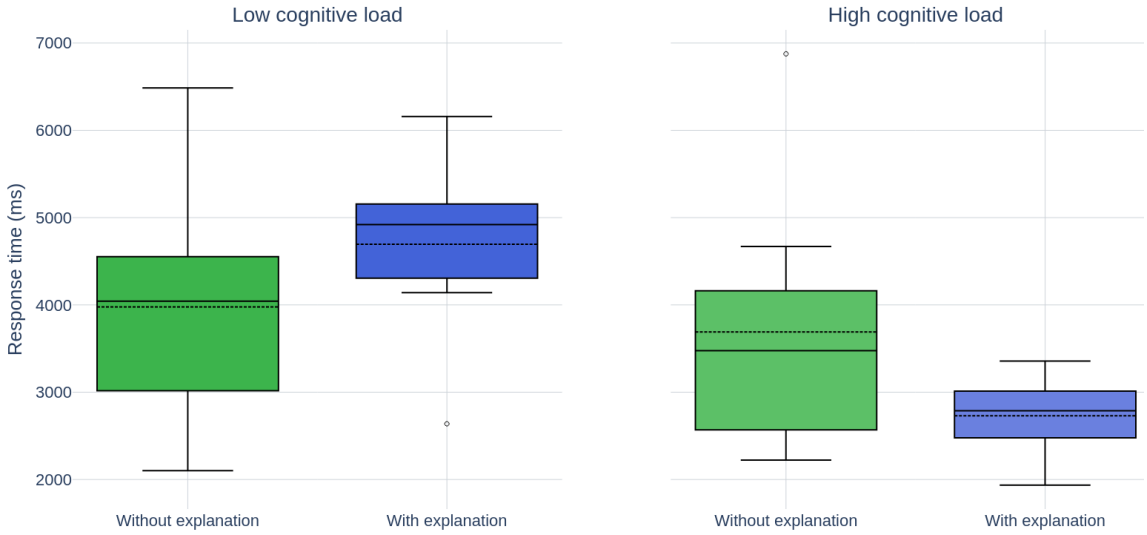

Fig. 7. Alert response times, measured as the average for each participant, across cognitive load and explanation detail. The dots indicate potential outliers.

Independent variable level	Mean	SD	Diff.	p
Alert response time under high cognitive load	3210.099	271.330	-1125.492	**.002**
Alert response time under low cognitive load	4335.591	366.206		
Alert response time with explanations	3712.867	222.089	-119.955	.738
Alert response time without explanations	3832.822	434.572		

Table 6. Table of means of alert response times showing the differences between the two levels of the two independent variables.

5.2.1 Alert Response Time. Figure 7 shows average alert response time across the cognitive load and explanation detail factors. A two-way repeated measures ANOVA was conducted to compare the average alert response time of the four combinations of cognitive load level and explanation detail. Analysis of the studentized residuals showed that there was normality, as assessed by the Shapiro-Wilk test of normality and no outliers, as assessed by no studentized residuals greater than ± 3 standard deviations. Sphericity always holds for factors with only two levels. The results revealed a significant main effect of cognitive load on average alert response time ($F(1,7) = 22.048$, $\mathbf{p = .002}$), but there was no significant main effect of explanation detail on average alert response time ($F(1,7) = .122$, $p = .738$). There was no statistically significant two-way interaction between cognitive load and explanation detail ($F(1,7) = 4,774$, $p = .065$).

Pairwise comparisons using Bonferroni adjustment revealed that the mean alert response time for the high cognitive load level (M = 3210.09ms, SD = 271.330) was lower compared to the low cognitive load level (M = 4335.591ms, SD = 366.206), **p = .002**. Table 6 shows that the difference between including or excluding explanations had a non-significant effect on alert response time.

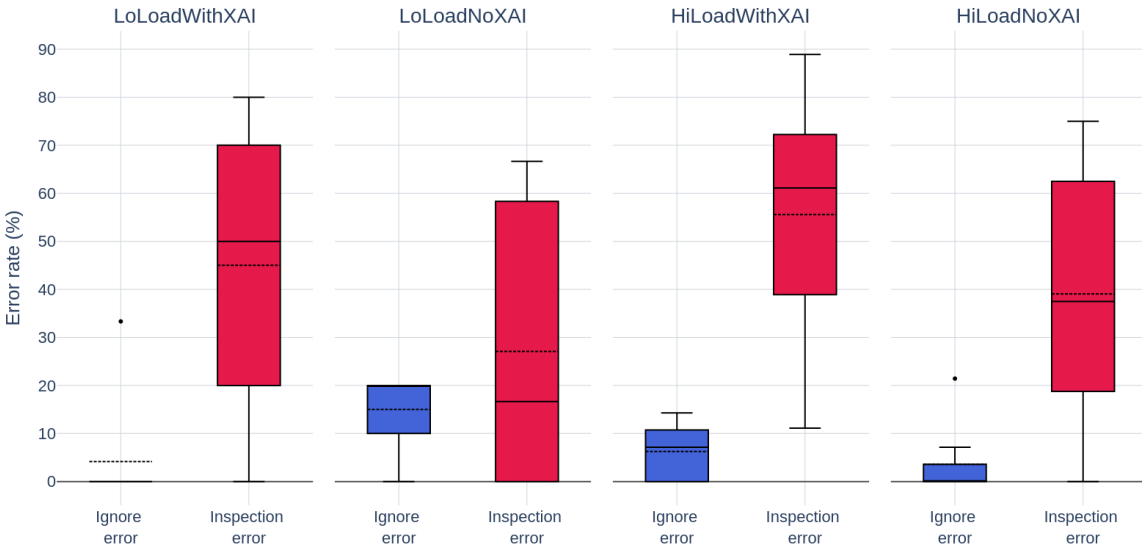

Fig. 8. Response error rate, measured for all participants as the percentage of responses that don't align with the ground truth, for each condition. Dots indicate potential outliers.

Independent variable level	Mean	SD	Diff.	p
Response error rate under high cognitive load	6.424	.859	1.410	.103
Response error rate under low cognitive load	5.014	.963		
Response error rate with explanations	6.636	.836	1.835	.52
Response error rate without explanations	4.801	.995		

Table 7. Table of means of the square root of *Mark for Inspection* error rate showing the differences between the two levels of the two independent variables.

5.2.2 Response Error Rate. Figure 8 shows response error rate for both *Mark for Inspection* and *Ignore* across the cognitive load and explanation detail factors. An *Ignore* error means that a participant responded *Mark for Inspection* on an alert with ground truth *Ignore*, and vice versa.

A two-way repeated measures ANOVA with square root transformation to correct for non-normality was conducted to compare *Mark for Inspection* error rate for the four combinations of cognitive load level and explanation detail. Analysis of the studentized residuals showed that there was normality, as assessed by the Shapiro-Wilk test of normality, and no outliers, as assessed by no studentized residuals greater than ± 3 standard deviations. The results showed no significant main effect of cognitive load on *Mark for Inspection* error rate ($F(1,7) = 3.510$, $p = .103$). There was also no significant interaction effect between cognitive load and explanation detail ($F(1,7) = .225$, $p = .650$). While the main

effect of explanation detail on *Mark for Inspection* error rate (F(1,7) = 5.483, p = .052) didn't reach significance, at p = 0.52 it is extremely close, so we performed the pairwise comparisons with Bonferroni adjustments. These revealed a higher mean square root of *Mark for Inspection* error rate when explanations were included (M = 6.636, SD = .836) compared to when explanations were excluded (M = 4.801, SD = .995), p = .052. Table 7 shows that the difference between high and low cognitive load had a non-significant effect on *Mark for Inspection* error rate. The fact that including explanations had no significant effect on alert response time, and possibly a negative effect on accuracy is puzzling, and it rejects **H2**.

Table 8 shows the *Ignore* error rate for each participant between the four conditions. Looking at participant 7 we see that they are the only one to make at least one *Ignore* error on each of the four conditions. This could indicate that, compared to many other participants, they were more cautious of not ignoring an alert that could have led to the missing person.

Participant	LoLoadWithXAI	LoLoadNoXAI	HiLoadWithXAI	HiLoadNoXAI
1	.00	20.0	7.14	.00
2	.00	20.0	.00	.00
3	.00	20.0	.00	.00
4	.00	.0	7.14	.00
5	.00	20.0	14.29	21.43
6	.00	.0	.00	.00
7	33.33	20.0	7.14	7.14
8	.00	20.0	14.29	.00

Table 8. *Ignore* error rate of each participant for each of the four conditions.

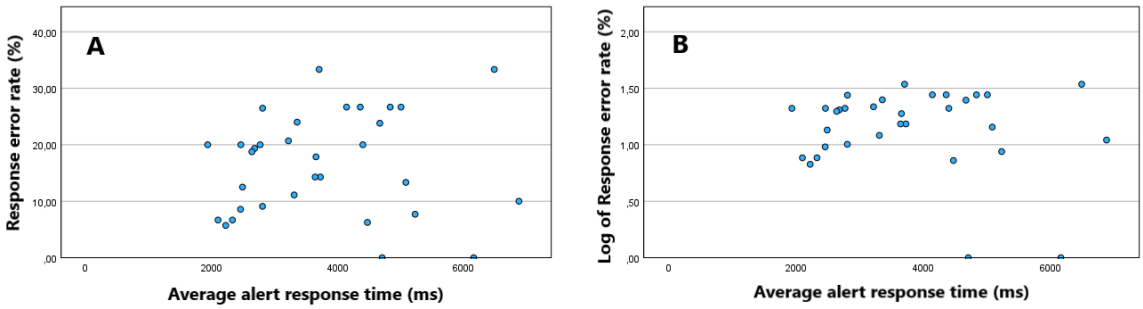

Fig. 9. Relationship between the average alert response time and the combined response error rate. (A) shows the non-transformed relationship. (B) shows the relationship when the response error rate has been log transformed.

5.2.3 Correlation between Performance Indicators. We wanted to evaluate whether there existed a correlation between the average alert response time and the overall response error rate, as we expected participants who took more time to respond to also make fewer errors. Two common methods for doing this are Pearson's correlation and Spearman's correlation. One of the assumptions of Pearson's correlation is that there exists a linear relationship between the two variables that are being evaluated. Figure 9 (A) shows the relationship between average alert response time and response error rate. It shows the relationship to be more clustered than linear. Instead of a linear relationship, Spearman's correlation assumes a monotonic relationship to exist, but as Figure 9 (A) shows, this also does not appear to have been

the case. In these cases it is recommended to try transforming either one, or both, of the variables. Figure 9 (B) shows the relationship between the average alert response time and the log transformed response error rate. This shows more characteristics of a linear relationship. However, another assumption of Pearson's correlation is the non-existence of significant outliers, and Figure 9 (B) indicates that there are a couple, for example the two with a response error rate of zero. It is possible to remove these and carry on with the analysis, but in this case an error rate of zero is more indicative of good performance than anomalous data, so we do not proceed. Therefore we conclude, from the data that was collected during our study, that we cannot with confidence say whether or not there is a correlation between the two performance indicators.

5.3 Subjective Reports

In order to shine a light on some conscious aspects of the participants' experience, we collected two types of subjective data: Perceived Workload using NASA-TLX and Perceived Usefulness using the TAM inspired statements. At the very end we also prompted participants to choose their preferred design between the one with explanations and the one without.

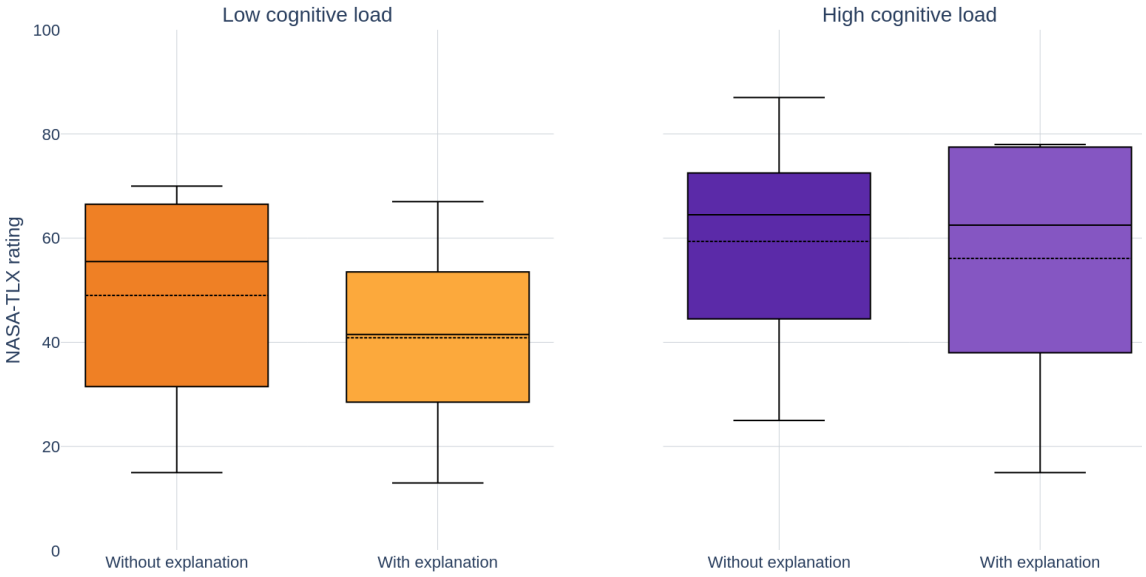

Fig. 10. Perceived workload from NASA-TLX assessment, across cognitive load and explanation detail.

5.3.1 Perceived Workload. Figure 10 shows the perceived workload across the cognitive load and explanation detail factors. A two-way repeated measures ANOVA was conducted to compare perceived workload for the four combinations of cognitive load level and explanation detail. Analysis of the studentized residuals showed that there was normality, as assessed by the Shapiro-Wilk test of normality and no outliers, as assessed by no studentized residuals greater than ± 3 standard deviations. The results revealed a significant main effect of cognitive load on perceived workload ($F(1,7) = 26.082$, $\mathbf{p = .001}$), but no significant main effect of explanation detail on perceived workload ($F(1,7) = 4.857$, $p = .063$) was found, rejecting **H3**. There was also no statistically significant two-way interaction between cognitive

load and explanation detail (F(1,7) = .625, p = .455). Pairwise comparisons using Bonferroni adjustment revealed that the mean perceived workload for the high cognitive load level (M = 57.750, SD = 8.076) was higher compared to the low cognitive load level (M = 44.938, SD = 6.601), **p = .001**. Table 9 shows that the difference between including and excluding explanations had a non-significant effect on perceived workload. This result is not surprising, as cognitive load and perceived workload are practically the same. Nonetheless, it confirms that the definition of high and low cognitive load that we used did in fact succeed at inducing significantly different cognitive load levels.

Independent variable level	Mean	SD	Diff.	p
Perceived workload under high cognitive load	57.750	8.076	12.812	**.001**
Perceived workload under low cognitive load	44.938	6.601		
Perceived workload with explanations	48.500	7.279	-5.688	.063
Perceived workload without explanations	54.188	7.482		

Table 9. Table of means of perceived workload showing the differences between the two levels of the two independent variables.

5.3.2 Perceived Usefulness. Where the NASA-TLX workload assessment was well suited for comparing designs across explanation detail, the perceived usefulness statements let us compare explanation detail across cognitive load levels. It also provided a look at the general sentiment of the usefulness of the two designs.

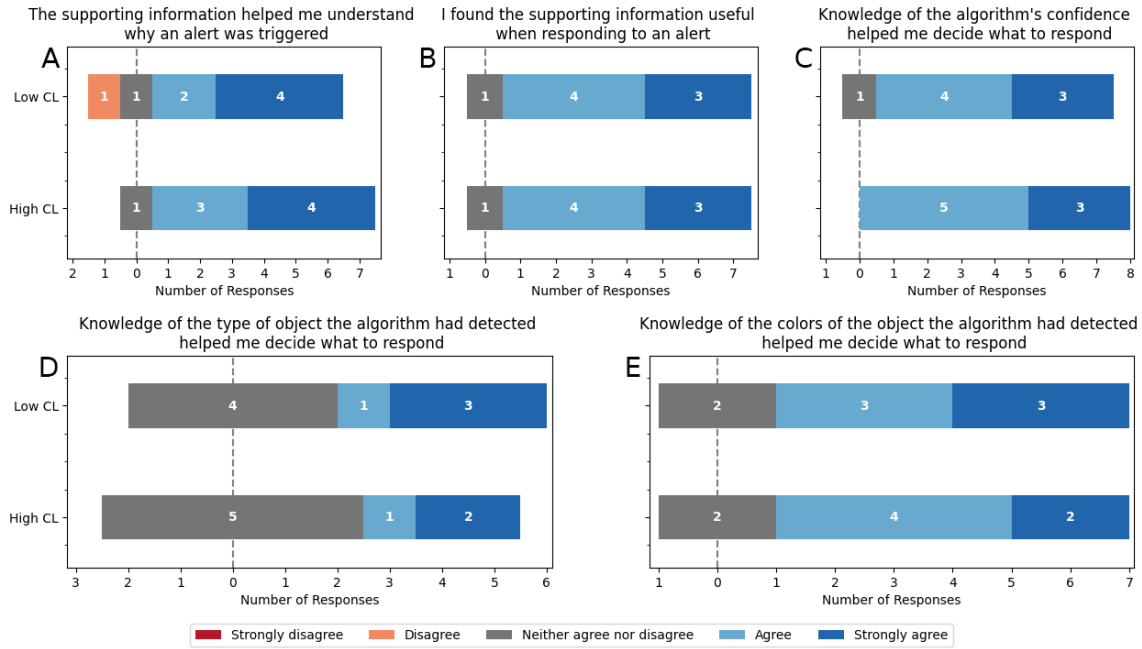

Fig. 11. Likert ratings of the perceived usefulness statements regarding the alert design with explanations. CL = cognitive load.

Figure 11 shows the resulting Likert rating of all five statements that were presented after the two conditions with explanations. It can be seen that the general sentiment of having supporting information available was quite positive (B).

More specifically, the results indicated that algorithm confidence level was the most useful (C), followed by the object colors (E), ending with the type of object that was detected (D), although no participant disagreed with the usefulness of any of the explanations. The ability of the explanations to help participants understand why an alert appeared was also quite good (A), with at least six participants agreeing under high and low cognitive load. Overall the results did not indicate that cognitive load levels have a meaningful impact on the usefulness of having explanations available.

Figure 12 shows the resulting Likert rating of two of the statements that were presented after the two conditions without explanations. The results indicated that having the missing person description presented on each alert was helpful to all participants (A). Interestingly, participants felt that they were generally able to confidently respond to alerts using just the alert image and the bounding box (B). As was the case for the conditions with explanations, cognitive load did not seem to have a meaningful impact on the perceived usefulness of the basic alert features.

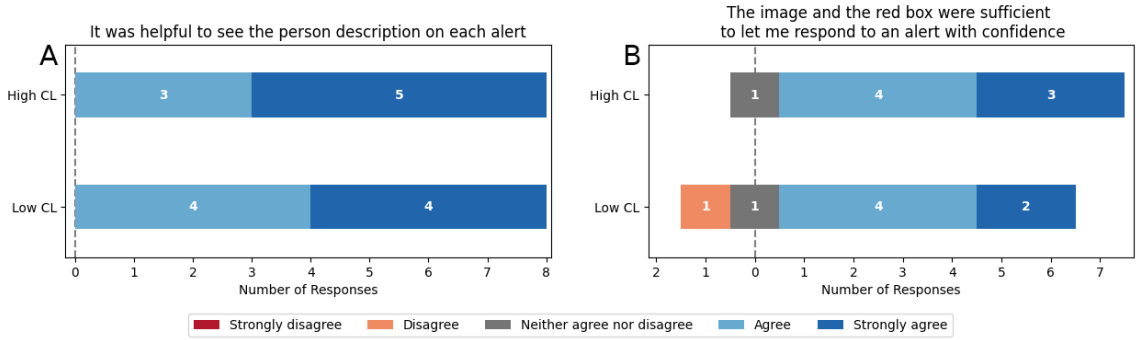

Fig. 12. Likert ratings of the perceived usefulness statements regarding the alert design without explanations. CL = cognitive load.

5.3.3 Design preference. At the end of the study, we collected each participant's preference regarding the alert design with or without explanations. Results showed that all participants preferred the design with explanations, even though the results showed no indication of it significantly improving their performance. This preference could be attributed to participants wanting the explanations there as a "nice to have" tool.

5.4 Alert Selection Patterns

To evaluate **H4**, we collected data on the order that alerts were selected for each participant, which we compared to the order that alerts appeared. As confidence was represented on the alert pins based on confidence intervals, the participants only had three different icons to help them decide which alert to select next, rather than the actual percentages. Therefore, we describe alert selection patterns based on the following three discrete values: low, mid, and high confidence, corresponding to the alert icons on the map.

Figure 13 shows the individual participants answers for the HiLoadWithXAI condition in the order they were opened. The y-axis represents each alert by their id, adjusted to the order they appeared, and the x-axis shows the sequence of chosen alerts for each participant. The black line going straight from (1,1) to (35,35) shows the order of alert appearances. Focusing on the points above this line, they seem to show that some participants have chosen an alert before it appeared in the system. However alerts may be appearing at a higher rate than they are answered, which makes it possible for a user to choose the 20th appearing alert as the 15th answer. The graph shows participant 1 and 6 having an almost straight line, indicating that the they may have chosen alerts based on the order they appeared. Generally for the rest

Fig. 13. Alert selection order for all participants. The red line passing through (3,4) represents that the third alert opened by participant 7, was the fourth alert to appear on the map.

of the participants, it seems they each have their own strategy when choosing alerts, a common pattern is visible in the start, but as the task goes on they get more separated. AI confidence may have been a factor for the choices the participants make when choosing an alert.

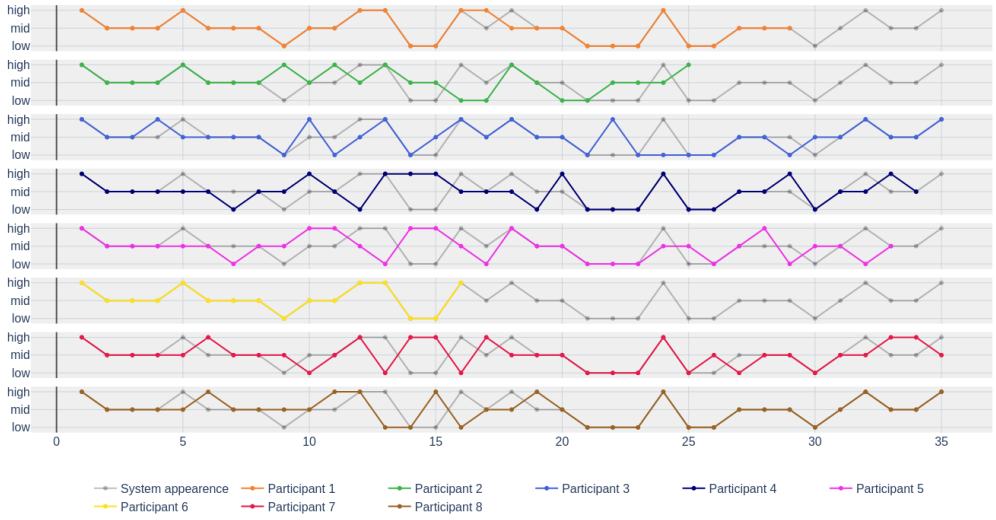

Fig. 14. Alert selection order for each individual participant. The common grey line shows the confidence level of each alert in the order they appeared on the map, e.g. the fifth alert to appear had a high confidence level. The colored lines represent the order in which participants selected alerts and the selected alert's confidence level. For example, the tenth alert participant 7 (red line) selected had a low confidence level.

Figure 14 shows each individual participants' alert choice and its corresponding confidence level. The data does not show a tendency to prioritize the selection of alerts based on the alert confidence level. There is also no indication that alerts of similar confidence levels are selected together. However, since it was not possible to see which alerts were available to participants at each selection, we can't say with certainty that they intentionally didn't follow a prioritization. Therefore we only partially reject **H4**.

6 DISCUSSION

Promoting understandability and clarity is critical for supporting effective cooperation between a SAR operator and an intelligent agent. This study investigated how explanations about the AI's decision making affected participants' performance and perception of usefulness under different cognitive load levels. Many of the study results were surprising and we now further discuss these, their implications for future research, and the limitations of our work.

6.1 A Tendency for Critical Assessment of Alerts

The participants that took part in our online study showed a vastly different behavior to the one that we had expected. They were fast enough to respond to the majority of alerts, they agreed on how to respond to many alerts, and they chose to ignore on average more than 75% of the alerts they saw. This may not necessarily be bad behavior, but it is surprising, and it might mean that there is a culture ingrained in the way that they are thinking. It could potentially also be due to our emphasis on the consequences of choosing to *Mark for Inspection*, which may have trumped the fear of failing to find a missing person. We also expected to see that if people took more time to respond they would be more accurate. In some instances this was the case, and in many others, being fast did not negatively impact accuracy. This might mean that there is a lot of diversity in participants' skill levels and the way that they approach the task. If researchers are to study more nuanced aspects of similar SAR systems, it would be beneficial to have consistency across participants, so that everyone treats the alert in a systematic way that a study can be built around.

6.2 Explanations and Performance

The fact that none of our hypotheses were supported, further highlights the surprising nature of the findings. Specifically, we were surprised by the inability of the explanations to show any positive impact on either response time, accuracy, or perceived workload. These findings also contrast the findings in [1] and [20], but they support the work of Schemmer et al. [33]. The fact that having explanations might even have lowered their accuracy is puzzling, since we effectively provide the answer by giving information that can be matched to the missing person description. That could also have an overall negative effect if users become so accustomed to just comparing words that they stop looking at the image, practically nullifying their experience and judgement which is otherwise a big contribution to the human-AI team. It is fully possible that the main potential of XAI in SAR systems is not as a direct improvement to performance, but an indirect improvement to effectiveness of the team through trust calibration [41]. However, it is also entirely possible that more carefully constructed explanations would show a desirable direct effect on performance.

6.3 Aligning AI Support and Human Expertise

The results from the statements that were presented after each condition and the forced choice feedback at the end indicated that participants probably had different interpretations of how to use explanations. When prompted to choose their preference between the alert design with explanations and the one without, all participants chose the one with explanations. One participant commented: "*It supports me better in my decision-making*", and another commented: "*Nice*

to see clearly what it thought it had found". One participants also commented: "*Supplementary information is never without importance*", which suggests that they might not have utilized the explanations often, but they were nice to have available in some cases.

The findings from the perceived usefulness statements showed that at least 6 participants found the image and the bounding box sufficient to confidently respond to an alert, which could further indicate that explanations, while useful, were not always necessary. However, this could also be attributed to the participants getting familiar with the explanations and starting to rely heavily on the confidence percentage, a problem similar to attentional tunneling [3, 19]. One participant commented: "*I trust the AI probability score quite a lot*", and another commented: "*It helps to make a quick decision if the text confirms what you see yourself*". This was not the intended value of the AI confidence, and it emphasises the importance of striving for trust calibration, rather than pure trust [13, 41]. What we actually wanted to portray was how confident the AI was that it had detected a certain object, not how confident it was that the object it had detected was related to the missing person. This concern is reinforced by the finding that confidence percentage was the most useful, ahead of both object type and colors detected.

Instead of the raw confidence score that we used, a score that corresponds to how well the detection matches the missing person description might be better suited. However, the risk of users neglecting their own experience and judgement in favor of over-trusting the AI recommendation still arises. We see this balance as one of the major challenges that similar SAR systems face when trying to partner an experienced operator with an intelligent agent.

Design implication	**Detailed description**
Explanations for alerts need to be unambiguous	Mental models are built on past experiences, which can cause them to be faulty in new environments [3]. Some participants used the AI confidence to a greater extent than it was designed for, which is supported by the finding that the most useful explanation was the confidence, and the comment: "*I rely quite a bit on the AI probability score*". When using explanations that are intended to support decision making, their meaning must be unmistakable.
The impact of each response action must be carefully defined	The results of our study showed that 77.4% of all alerts were ignored, indicating that participants didn't want to inspect objects unless it seemed absolutely critical. Using replicated real world scenarios, the consequence of each user decision should be explained thoroughly to ensure users know the implications of each action.
Value the expertise of the operator	For AI and humans to form partnerships they need jointly learn to utilize their capabilities [35]. In our study that was not always the case. However, one participant commented: "*I thought it was a good help because I first look at the picture and think about what it could be, if I'm in doubt I look over and see if it can help describe the object. You should just not trust it 100%.*", which indicates that they prioritized their own judgement.
The missing person description should be present when assessing an alert	Results from the perceived usefulness statements showed that all participants liked having the missing person description on each alert page. This helps to mitigate the situational awareness demon of mental overload described by Agrawal et al. [3].

Table 10. Design implications for alert design in a drone swarm-based SAR system.

6.4 Implications for Design

Insight provided by the study illuminates the challenges associated with using AI-generated explanations to enhance collaboration with SAR operators. We produce implications for design, as a method for facilitating insights to other designers or researchers [32]. Table 10 presents a set of suggestions to consider when designing alerts for a drone-swarm SAR system.

6.5 Limitations

There are multiple aspects of this work that present possible limitations. These relate to the ground truths being manually annotated, a small and non-diverse participant sample, and cognitive load levels not being tested on actual SAR personnel.

Explanations and ground truths were all manually annotated by the researchers. Using an actual AI algorithm in the case of the explanations, and using domain experts in the case of the ground truths would have made them more valid. When doing the annotations, we used the guidelines that were provided to participants regarding how to respond. We also openly discussed and challenged the explanation details and ground truth labels, thereby refining them before being finalized. When more tailored algorithms become available, they should be used to provide the explanations, as also suggested in [37].

The number of participants that were recruited for this study was quite low, and we have been hesitant to make generalizing claims about the findings. The participants we did recruit were all male and living in Denmark. We accept that since the gender distribution of the profession is largely male, and because nationality is not perceived to affect any of the objective measures in this study. However, participants from more conservative cultures may show greater hesitation to adopt AI technology. Additionally, expertise in SAR missions and experience with use of drones varied a lot between the participants. Since we were limited to a small population of potential subjects, including people with less experience could at least give some insights into how people would respond to such a system. Furthermore the system we developed is completely new to all participants and therefore shouldn't put less experienced participants too much at a disadvantage, apart from the missing SAR experience that could help with decision making.

Lastly, the criteria we use for inducing high and low cognitive load had only been tested among the researchers and two colleagues. To mitigate this problem, the alert per minute limits that were found during those tests were pushed further outwards to increase the likelihood that all participants would experience the same cognitive load levels as ourselves and our colleagues.

7 FUTURE WORK

A central aspect of presenting explanations relates to their long term effect on the operators skills and performance. Studying their behavior in more detail, by e.g. using an eye-tracking device, could illuminate some of the nuanced aspects of sustained exposure to AI explanations. This could provide valuable insight into what information and features of an alert people notice and prioritize, and such insights would greatly help to inform alert design that gets the best out of both the human and the AI.

An idea that unfortunately came up too late for us to make it a part of the study was to include an indication on the alert page of how well the detection matched the missing person description. Simplifying the explanations to just include this match indicator could make it more likely that blindly choosing *Mark for Inspection* when the match is

high, is actually a good decision more often than not. However, the primary goal should of course be to make sure that users make a decision based on a combination of their experience and the AI's information.

In this work we employed the use of a separate alert handler role. This gave us more freedom to design alerts that align with recommendations without having to worry about accommodating control of the swarm on top of it. If the practice of having a designated alert handler is widely adopted, it would be relevant to investigate the necessity of the map as the hub from where alerts are selected. We imagine that a more detailed preview of each alert could be provided on a page that was dedicated to the purpose of showing alerts, thus further helping users prioritize which alert to look at next.

One of the major challenges of this work was to determine how participants should behave when faced with the different situations in the study. Future work should seek to define such guidelines and best practices in SAR drone swarm systems. Specifically, defining what characterizes a good drone swarm operator or alert handler, and defining a systematic way to handle alerts, should be a priority and will be key to maturing a shared understanding of these systems.

8 CONCLUSION

With the rise of drone swarms for SAR on the horizon, it is important to establish how automation can be used to facilitate smooth and effective human-swarm collaboration. This study moved beyond drone swarm control, to contribute with insights into how the human operator can interact with the live detections made by such a drone swarm. Our main research question was: ***How can explanations be designed to mitigate cognitive overload and improve the performance of SAR personnel responding to AI-generated alerts?*** Through extensive engagement with emergency services we gained the knowledge to develop a system that presented detections from five simulated drones searching for a person. This system was used to facilitate an online study for 8 participants involved with the Danish Emergency Services. The results suggested that explanations in AI-generated alerts didn't have a direct positive impact on operator performance or cognitive overload, however they did help support understandability and decision-making. The knowledge gained from the user study allowed us to generate a set of design implications for alerts in a drone swarm-based SAR system, which provide recommendations for future designers of such systems. The findings in the present paper further underline the complexity of partnering SAR professionals with artificial intelligence, and help to guide future research in a direction where the skills and potentials of both sides are allowed to flourish.

ACKNOWLEDGEMENTS

We thank all participants for taking time out of their day to partake in our study. We also acknowledge that this work has benefited greatly from the discussions we have had with the domain experts from DEMA and Robotto, who openly shared their time and knowledge with us.

REFERENCES

[1] Ashraf Abdul, Christian Von Der Weth, Mohan Kankanhalli, and Brian Y. Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. *Conference on Human Factors in Computing Systems - Proceedings* (4 2020), 1–14. https://doi.org/10.1145/3313831.3376615

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (9 2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[3] Ankit Agrawal, Sophia J. Abraham, Benjamin Burger, Chichi Christine, Luke Fraser, John M. Hoeksema, Sarah Hwang, Elizabeth Travnik, Shreya Kumar, Walter Scheirer, Jane Cleland-Huang, Michael Vierhauser, Ryan Bauer, and Steve Cox. 2020. The Next Generation of Human-Drone Partnerships: Co-Designing an Emergency Response System. *Conference on Human Factors in Computing Systems - Proceedings* (4 2020), 1–13. https://doi.org/10.1145/3313831.3376825

[4] Abdulla Al-Kaff, David Martín, Fernando García, Arturo de la Escalera, and José María Armingol. 2018. Survey of computer vision algorithms and applications for unmanned aerial vehicles. *Expert Systems with Applications* 92 (2 2018), 447–463. https://doi.org/10.1016/J.ESWA.2017.09.033

[5] Ma'Moun Al-Smadi, Khairi Abdulrahim, and Rosalina Abdul Salam. 2016. Traffic surveillance: A review of vision based vehicle detection, recognition and tracking. *International Journal of Applied Engineering Research* 11, 1 (2 2016), 713–726. https://www.researchgate.net/publication/298711223_Traffic_surveillance_A_review_of_vision_based_vehicle_detection_recognition_and_tracking

[6] Ross Arnold, Jonathan Jablonski, Benjamin Abruzzo, and Elizabeth Mezzacappa. 2020. Heterogeneous UAV Multi-Role Swarming Behaviors for Search and Rescue. *Proceedings - 2020 IEEE International Conference on Cognitive and Computational Aspects of Situation Management, CogSIMA 2020* (8 2020), 122–128. https://doi.org/10.1109/COGSIMA49017.2020.9215994

[7] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (6 2020), 82–115. https://doi.org/10.1016/J.INFFUS.2019.12.012

[8] Antonio Brunetti, Domenico Buongiorno, Gianpaolo Francesco Trotta, and Vitoantonio Bevilacqua. 2018. Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing* 300 (7 2018), 17–33. https://doi.org/10.1016/J.NEUCOM.2018.01.092

[9] J. Cacace, A. Finzi, V. Lippiello, M. Furci, N. Mimmo, and L. Marconi. 2016. A control architecture for multiple drones operated via multimodal interaction in search & rescue mission. *SSRR 2016 - International Symposium on Safety, Security and Rescue Robotics* (12 2016), 233–239. https://doi.org/10.1109/SSRR.2016.7784304

[10] Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. 2022. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior* 127 (2 2022). https://doi.org/10.1016/J.CHB.2021.107018

[11] Andreas Daugbjerg Christensen, Andreas Skjoldgaard Andersen, Shpend Gjela, and Philip Michaelsen. 2023. *Interfaces for Live Video-streams from Search and Rescue Drone Swarms.* Technical Report.

[12] Fred D. Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly: Management Information Systems* 13, 3 (9 1989), 319–340. https://doi.org/10.2307/249008

[13] Akuadasuo Ezenyilimba, Margaret Wong, Alexander Hehr, Mustafa Demir, Alexandra Wolff, Erin Chiou, and Nancy Cooke. 2022. Impact of Transparency and Explanations on Trust and Situation Awareness in Human–Robot Teams. *Journal of Cognitive Engineering and Decision Making* 17, 1 (11 2022), 75–93. https://doi.org/10.1177/15553434221136358

[14] David Gunning. 2017. *Explainable Artificial Intelligence (XAI).* Technical Report. Defense Advanced Research Projects Agency (DARPA).

[15] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (10 2006), 904–908. https://doi.org/10.1177/154193120605000909

[16] Maria-Theresa Oanh Hoang, Andreh Bassam Bahodi, and Rasmus Skov Buchholdt. 2022. *User Interface Design for UAV Swarms in Search and Rescue.* Technical Report. Aalborg University, Aalborg.

[17] Maria Theresa Oanh Hoang, Niels Van Berkel, Mikael B. Skov, and Timothy R. Merritt. 2023. Challenges and Requirements in Multi-Drone Interfaces. *Conference on Human Factors in Computing Systems - Proceedings* (4 2023), 1–9. https://doi.org/10.1145/3544549.3585673

[18] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2013. The Effect of Explanations on Perceived Control and Behaviors in Intelligent Systems. *Conference on Human Factors in Computing Systems - Proceedings* (4 2013), 181–186. https://doi.org/10.1145/2468356.2468389

[19] Jean Paul Imbert, Helen M. Hodgetts, Robert Parise, François Vachon, Frédéric Dehais, and Sébastien Tremblay. 2014. Attentional costs and failures in air traffic control notifications. *Ergonomics* 57, 12 (12 2014), 1817–1832. https://doi.org/10.1080/00140139.2014.952680

[20] Alexander John Karran, Théophile Demazure, Antoine Hudon, Sylvain Senecal, and Pierre Majorique Léger. 2022. Designing for Confidence: The Impact of Visualizing Artificial Intelligence Decisions. *Frontiers in Neuroscience* 16 (6 2022). https://doi.org/10.3389/FNINS.2022.883385

[21] Peter Kearney, Wen Chin Li, and John J.H. Lin. 2016. The impact of alerting design on air traffic controllers' response to conflict detection and resolution. *International Journal of Industrial Ergonomics* 56 (11 2016), 51–58. https://doi.org/10.1016/J.ERGON.2016.09.002

[22] Vincenzo Lomonaco, Angelo Trotta, Marta Ziosi, Juan de Dios Yáñez Ávila, and Natalia Díaz-Rodríguez. 2018. Intelligent Drone Swarm for Search and Rescue Operations at Sea. (11 2018). https://arxiv.org/abs/1811.05291v1

[23] Angela Mastrianni, Aleksandra Sarcevic, Lauren Chung, Issa Zakeri, Emily Alberto, Zachary Milestone, Ivan Marsic, and Randall S Burd. 2021. Designing Interactive Alerts to Improve Recognition of Critical Events in Medical Emergencies. *Designing Interactive Systems Conference 2021* (2021), 864–878. https://doi.org/10.1145/3461778.3462051

[24] Denys J C Matthies, Laura Milena Daza Parra, and Bodo Urban. 2018. Scaling Notifications Beyond Alerts: From Subtly Drawing Attention up to Forcing the User to Take Action. *Adjunct Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (8 2018), 45–47. https://doi.org/10.1145/3266037.3266096

[25] Balmukund Mishra, Deepak Garg, Pratik Narang, and Vipul Mishra. 2020. Drone-surveillance for search and rescue in natural disaster. *Computer Communications* 156 (4 2020), 1–10. https://doi.org/10.1016/J.COMCOM.2020.03.012

[26] Philipp Müller, Sander Staal, Mihai Bâce, and Andreas Bulling. 2022. Designing for Noticeability: Understanding the Impact of Visual Importance on Desktop Notifications. *Conference on Human Factors in Computing Systems - Proceedings* (4 2022), 1–13. https://doi.org/10.1145/3491102.3501954

[27] Li Qingqing, Jussi Taipalmaa, Jorge Pena Queralta, Tuan Nguyen Gia, Moncef Gabbouj, Hannu Tenhunen, Jenni Raitoharju, and Tomi Westerlund. 2020. Towards Active Vision with UAVs in Marine Search and Rescue: Analyzing Human Detection at Variable Altitudes. *2020 IEEE International Symposium on Safety, Security, and Rescue Robotics, SSRR 2020* (11 2020), 65–70. https://doi.org/10.1109/SSRR50563.2020.9292596

[28] Jorge Peña Queralta, Jenni Raitoharju, Tuan Nguyen Gia, Nikolaos Passalis, and Tomi Westerlund. 2020. AutoSOS: Towards Multi-UAV Systems Supporting Maritime Search and Rescue with Lightweight AI and Edge Computing. (5 2020). https://arxiv.org/abs/2005.03409v1

[29] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (4 2018), 1–13. https://doi.org/10.1145/3173574.3173677

[30] Sasa Sambolek and Marina Ivasic-Kos. 2021. Automatic person detection in search and rescue operations using deep CNN detectors. *IEEE Access* 9 (3 2021), 37905–37922. https://doi.org/10.1109/ACCESS.2021.3063681

[31] Wojciech Samek and Klaus Robert Müller. 2019. Towards Explainable Artificial Intelligence. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 11700 LNCS. Springer Verlag, 5–22. https://doi.org/10.1007/978-3-030-28954-6{_}1

[32] Corina Sas, Steve Whittaker, Steven Dow, Jodi Forlizzi, and John Zimmerman. 2014. Generating implications for design through design research. *Conference on Human Factors in Computing Systems - Proceedings* (2014), 1971–1980. https://doi.org/10.1145/2556288.2557357

[33] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kuhl, and Michael Vossing. 2022. A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. *AIES 2022 - Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (7 2022), 617–626. https://doi.org/10.1145/3514094.3534128

[34] Jürgen Scherer, Saeed Yahyanejad, Samira Hayat, Evşen Yanmaz, Vladimir Vukadinovic, Torsten Andre, Christian Bettstetter, Bernhard Rinner, Asif Khan, and Hermann Hellwagner. 2015. An autonomous multi-UAV system for search and rescue. *DroNet 2015 - Proceedings of the 2015 Workshop on Micro Aerial Vehicle Networks, Systems, and Applications for Civilian Use* (5 2015), 33–38. https://doi.org/10.1145/2750675.2750683

[35] Tjeerd A.J. Schoonderwoerd, Emma M.van Zoelen, Karel van den Bosch, and Mark A. Neerincx. 2022. Design patterns for human-AI co-learning: A wizard-of-Oz evaluation in an urban-search-and-rescue task. *International Journal of Human-Computer Studies* 164 (8 2022). https://doi.org/10.1016/J.IJHCS.2022.102831

[36] Niels Van Berkel, Omer F. Ahmad, Danail Stoyanov, Laurence Lovat, and Ann Blandford. 2020. Designing Visual Markers for Continuous Artificial Intelligence Support: A Colonoscopy Case Study. *ACM Transactions on Computing for Healthcare* 2, 1 (12 2020), 1–24. https://doi.org/10.1145/3422156

[37] Niels Van Berkel, Jeremy Opie, Omer F. Ahmad, Laurence Lovat, Danail Stoyanov, and Ann Blandford. 2022. Initial Responses to False Positives in AI-Supported Continuous Interactions: A Colonoscopy Case Study. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 12, 1 (3 2022), 1–18. https://doi.org/10.1145/3480247

[38] Christopher Wickens and Angela Colcombe. 2007. Dual-Task Performance Consequences of Imperfect Alerting Associated With a Cockpit Display of Traffic Information. *Human Factors* 49, 5 (10 2007), 839–850. https://doi.org/10.1518/001872007X230217

[39] Jiaming Yan, Zhengyu Peng, Hong Hong, Hui Chu, Xiaohua Zhu, and Changzhi Li. 2018. Vital-SAR-Imaging With a Drone-Based Hybrid Radar System. *IEEE Transactions on Microwave Theory and Techniques* 66, 12 (12 2018), 5852–5862. https://doi.org/10.1109/TMTT.2018.2874268

[40] Wen Bo Yu, Daniel MacCormick, Loutfouz Zaman, and Pejman Mirza-Babaei. 2019. Getting the player's attention: Comparing the effectiveness of common notification types in task management games. *CHI PLAY 2019 - Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play* (10 2019), 797–804. https://doi.org/10.1145/3341215.3356300

[41] Yunfeng Zhang, Q. Vera Liao, and Rachel K.E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (1 2020), 295–305. https://doi.org/10.1145/3351095.3372852