Beyond the Wizard of Oz: Negative Effects of Imperfect Machine Learning to Examine the Impact of Reliability of Augmented Reality Cues on Visual Search Performance

Aditya Raikwar, Domenick Mifsud, Christopher D. Wickens, Anil Ufuk Batmaz, Amelia C. Warden, Brendan Kelley, Benjamin A. Clegg, Francisco R. Ortega

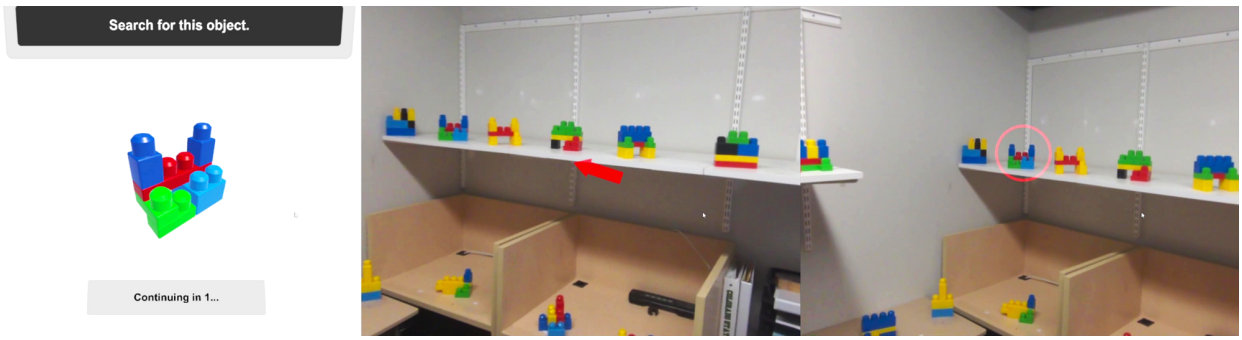

Fig. 1: Snapshots of one trial (Left panel: Target image is shown to the participants to remember; Center panel: Participants search the 3D environment for the designated object, there may be a cue or no cue depending on the trial condition; Right panel: selecting or rejecting the suggested object)

Abstract—Despite knowing exactly what an object looks like, searching for it in a person's visual field is a time-consuming and error-prone experience. In Augmented Reality systems, new algorithms are proposed to speed up search time and reduce human errors. However, these algorithms might not always provide 100% accurate visual cues, which might affect users' perceived reliability of the algorithm and, thus, search performance. Here, we examined the detrimental effects of automation bias caused by imperfect cues presented in the Augmented Reality head-mounted display using the YOLOv5 machine learning model. 53 participants in the two groups received either 100% accurate visual cues or 88.9% accurate visual cues. Their performance was compared with the control condition, which did not include any additional cues. The results show how cueing may increase performance and shorten search times. The results also showed that performance with imperfect automation was much worse than perfect automation and that, consistent with automation bias, participants were frequently enticed by incorrect cues.

Index Terms—Augmented Reality, Visual Search, Automation Bias, Imperfect Cues

1 Introduction

Visual search involves actively scanning through perceptual information to locate a specific target. For humans in complex visual environments, the search process will be effortful and serial [23, 37, 42]. Examples include a pilot scanning for potential hazards [1], a radiologist searching for a tumor or fracture [43], and a quality inspector searching a product for defects [9, 11]. In such real-world visual search settings, delays and errors in identifying targets can have profound implications. Support for visual search is possible through the use of Augmented Reality (AR) Head-Mounted Displays (HMDs) to cue individuals to targets identified through machine learning solutions. Users can see cues to target objects in relation to their real-world surroundings, which could be necessary for finding objects that might be outside their current field of view, hidden from a certain angle, or behind other things. Cues can provide essential contextual awareness, giving users a sense of where objects are located within their surroundings. In addition, these cues contribute to effective navigation and reduce the cognitive load associated by helping users perceive spatial relationships [7]. By enhancing immersion and increasing efficiency, visual cues serve as invaluable guides for visual search tasks in complex task environments, ensuring that users can quickly and accurately locate objects.

While benefits to both latency and accuracy from AR cues in visual search situations have been shown [4, 12], examining the impact of errors in the cues provided on performance is vital. Even where automated systems might outperform human operators in recognizing and recording the locations of targets, errors within such systems are still possible. In exploring the impact of imperfect cueing, Warden et al. found that when given correct arrow cues to the location of a target, as expected, participants were faster and more accurate in a visual search task [40]. However, incorrect arrow cues also produced higher error rates and longer search times than the uncued cases. This cost to accuracy suggests that participants were over-relying on the visual

- *Aditya Raikwar is with Colorado State University. E-mail: adirar@colostate.edu*
- *Domenick Mifsud is with Georgia Institute of Technology. E-mail: dmifsud3@gatech.edu*
- *Christopher D. Wickens is with Colorado State University. E-mail: chris.wickens@colostate.edu*
- *Anil Ufuk Batmaz is with Concordia University. E-mail: ufuk.batmaz@concordia.ca*
- *Amelia C. Warden is with Colorado State University. E-mail: acwarden@colostate.edu*
- *Brendan Kelley is with Colorado State University. E-mail: brendan.kelley@colostate.edu*
- *Benjamin A. Clegg is with Montana State University. E-mail: benjamin.clegg@montana.edu*
- *Francisco R. Ortega is with Colorado State University. E-mail: fortega@colostate.edu*

Manuscript received 4 October 2023; revised 17 January 2024; accepted 24 January 2024. Date of publication 4 March 2024; date of current version 15 April 2024. This article has supplementary downloadable material available at https://doi.org/10.1109/TVCG.2024.3372062, provided by the authors. Digital Object Identifier no. 10.1109/TVCG.2024.3372062

cues. Such work underlines the value of understanding how dependent users are on signals and how their performance is affected by imperfect signals. The choice of cue to deploy in a particular context ultimately might be determined not only by the magnitude of the benefits they provide when correct but also by the ability of humans to recognize when the cues are incorrect and overcome the tendency towards **automation bias** [22]. Automation Bias is a cognitive bias that people employ during decision-making when they rely on decisions and inferences of automation rather than seek information to make their own judgments and decisions [26].

Previous studies on visual cues, such as Warden et al. [40], used a Wizard-of-Oz (WoZ) approach to set up the detrimental effects of imperfect cues on user search performance. Such work offers valuable initial insight into the impact of incorrect cues on the accuracy of finding the target. However, human-specified errors in such systems introduce an essential limitation to the potential generalization of the findings. Errors can be determined either randomly – with incorrect cues occurring on a specified fraction of trials and directing people to a randomly selected incorrect distractor location; or systematically based on human input – with incorrect cues happening for ''hard" or ''easy" targets, and/or cues directing operators to distractors that are ''similar" or "dissimilar" to the actual target. The issue this creates is that whichever path is chosen to create the imperfect cues, actual errors from machine learning could be very different. Humans are able to correlate different variables even when those variables aren't present in the dataset, while machines can't do that, leading to different errors.

Evidence that the types of representations developed through deep learning differ from human representations (see for example, [8]), that humans and machines make different types of errors [3, 27], and attempts to offer additional insights into the machine learning recommendations through aspects like explainable Artificial Intelligence (AI) to reduce automation bias (e.g., [33]), all point to an array of potential mismatches between humans and automation. Such differences create ways in which imperfect cues generated within a machine learning system could diverge from evidence produced through WoZ scenarios. In addition, a machine learning system working in real-time has the potential to change the cued location as it obtains more information and updates its predictions. The influence of varying locations being cued through a single search, which provides a further source of information about the potential unreliability of the cue being offered, adds an additional justification to explore actual machine learning-driven cues rather than just simulated versions of them.

In this paper, we implemented a real-time object recognition system using a machine learning (ML) algorithm. We analyzed whether the results from visual cues from such a system differ from prior WoZ approaches. For this, we used an arrow as a visual cue to point out the objects. This also helped us gain insight into automation bias in ML. The expected consequences of automation bias are that people fail to respond to critical information or follow the aid (i.e., cues) when it is wrong. This bias is pervasive in situations where people trust technology to be more accurate and efficient than human judgment [26]. The findings are especially relevant in safety-critical applications where automation bias can lead to errors or overlooked information. For example, in autonomous vehicles, medical diagnosis, or aviation, understanding how users respond to automation cues is crucial for system design and training.

Our research aims to address the convergence of human performance and computer technology in cued search by (1) examining the effects of visual cues for searching in AR HMDs and (2) considering the imperfections of automation, such as vision-based AI, that may make the AI-based cueing imperfectly reliable. The visual cues are displayed on the near-field, e.g., within arm's reach. The far field is where the target objects are located (about 1 meter away from the user). The distinction between near- and far-field refers to the positioning in the real world, not the display on the video pass-through headset. Refer to Fig. 1 for step-by-step visuals of a single trial.

The contributions of this research paper include:

1. Increasing the external validity given a more realistic scenario compared to a WoZ study for real-time AR HMDs cueing using ML during a visual search task prevalent in this type of research.
2. Demonstrated that the automation bias is high when using a visual cue. Due to the higher reliability of the imperfect visual cue, the magnitude of automation bias was greater using ML than in the WoZ study.
3. Demonstrated that the automation bias in WoZ studies of AR HMD cueing is replicated when using an AR HMD system using ML.
4. Released open-source code and ML training dataset (https://github.com/NuiLab/ML-Imperfect-Cueing), the dataset, and the ML methodology used to train blocks using virtual objects as opposed to pictures. We hope this will help researchers replicate studies and conduct other studies related to visual search.

2 Related Work

The time required to perceive and interpret a particular cue is denoted by its Cue Effectiveness Value (CEV) [29–31, 45]. The CEV considers the time needed to perceive and interpret a cue and orient attention from the cue to the cued target. This is an essential factor to consider when the aim is to reduce the search time using the cue [29–31, 41, 45]. CEV can be predicted and modeled by understanding the cognitive processes involved in visual search. Based on previous studies [39, 40], we have selected an arrow as the cue that benefits visual search performance the most. One crucial factor influencing these processes is whether the cue is exogenous or endogenous [31, 41]. An *endogenous cue* provides information about the target's location but does not appear at that location itself. For example, an image of the target cue that represents the target's appearance but does not directly indicate where attention should be directed is an endogenous cue. In contrast, an *exogenous cue* directs attention to the target's location, such as a flashing highlight near the object in the world (can be presented on an HMD). Exogenous cues have limitations of only being effective if the target is located in the field of view and may distract individuals when they are not actively searching for that specific target. We used an arrow cue to direct attention to the objects to overcome these limitations. We used an arrow because it can point to objects outside the field of view and is not as distracting as a flashing light. These types of cues can be referred to as world-referenced.

The design of AR HMDs can utilize world-referenced cueing and enables the creation of exogenous cues independent of the head's orientation. This cueing method involves placing a cue on an HMD to constantly overlay or point directly at the target, regardless of how the head is turned. This is achieved by presenting the cue on the display in world-referenced coordinates, which are continuously updated on the display as the head moves horizontally or vertically. In the present investigation, we used this type of cue in the form of an AR arrow that consistently indicates the direction of the target.

It is essential to consider the user's prior knowledge about the search target, as detailed information regarding the target may impact visual search tasks. In some cases, a user may only have general knowledge of the target, such as hazards [1] or physical injury [43, 44], which require the user to identify something that may warp or change on a case by case basis. In other cases, the user may have greater knowledge about a target or may have been prompted to search for a specific target. This may be a pedestrian [28], which may differ in some regards but have common, well-known attributes, office tools or supplies [19], or specific virtual geometries [16]; all of which vary to a lesser degree.

The cues used in previous studies, e.g., [16, 39], were always accurate, but in real-world situations, cues are generated through automation inferences based on what the system considers the target. Computer vision using ML algorithms can achieve this by comparing the features of each object in the search field, like humans, with a template of what the true target looks like. However, like many other functions of automation, this inference is not always accurate, as various factors such as lighting conditions, pattern complexity in the search scene, and the reliability of the machine-vision system can affect and degrade its performance [2, 6, 10, 13, 17, 35].

The imperfect reliability of automation has significant implications for human performance. Imperfectly reliable automation can impact how much humans rely on the automation's decision of what should be attended to, and researchers have examined these factors in relation to human trust in imperfect automation [13, 17, 18, 34]. Several studies have specifically examined this issue in the context of visual search and target cueing, suggesting that cueing benefits are reduced as reliability is reduced [2, 6, 10, 21, 48]. Although Mifsud et al. [21] have conducted a study closely related to visual cues, only Yeh et al. [47], and Warden et al. [40], to our knowledge, have done so in the specific context of AR HMD target cueing (without the use of ML).

The above research has yielded two critical findings regarding human reliance on imperfect automation. First, a vast amount of research in human-automation interaction has examined the consequences of imperfectly reliable automation (see [13, 32, 35] for a summary). Such research has typically revealed that many participants follow automation's advice or attention guidance infrequently when automation is wrong, referred to as the automation bias [26]. Some research suggests that the automation bias is particularly prevalent because the guidance is offered in highly realistic AR format [21]. In the cueing work from Warden et al. [39], the search guidance cues were 100% reliable. A follow-up study examined the automation bias by conducting an experiment that used WoZ [40]. The study used WoZ methodology to examine the impacts of unreliable cues, which suggested that people exhibit an automation bias in visual search using AR HMDs. Three cues were provided: a mini-map, an icon image, and an arrow pointing to the object. The more realistic cue (i.e., the arrow cue) exhibited the highest automation bias. With this in mind, we designed the current study presented here to look at automation bias using real-time ML (i.e., no WoZ) to understand if there were differences, given the lack of control in the accuracy of the ML using the arrow cue.

Moreover, it has been found that the impact of cue imperfections can vary depending on the type of cue used. Specifically, the more reliable an AR cue is, the more it can assist users when correct. But it can also lead to more significant human errors caused by the automation bias when it is wrong. This prediction is somewhat supported by research on imperfect AR target cueing, which presented an 83% reliable cue pointing to a potential target object [40]. The study found that the AR HMD world-reference ego-centric (i.e., a cue that encodes the location of a target with respect to the viewer) arrow cue was more effective when the cue was correct, but more problematic when the cue was wrong, suggesting that the more immersive and world-referenced cue amplified the automation bias. This tendency is consistent with the notion that a more immersive cue leads to increased *attentional tunneling* [36, 46] ignoring the raw data in the real environment outside the HMD. Supporting this causal assumption, researchers have found that AR HMD cueing reduced the detection of other non-cued but high-priority threats in the scene [46, 47].

As stated above, our work differs from previous work, including recent findings from Warden et al. [39, 40], which used the WoZ method. In this paper, we used an ML algorithm for external validation to investigate the negative effects of imperfect visual cues. Additionally, our research addresses a longstanding inquiry concerning the WoZ method's efficacy in the context of visual cues. Additionally, we showed that ML algorithms produce a higher automation bias than the WoZ method.

3 Methodology

3.1 Hypotheses

We hypothesized that the interaction of the exogenous type of cue and the impact of cue imperfections would be significant, such that the benefits of cues would be greater for participants in the perfect automation condition compared to those in the imperfect automation condition. Overall, this design allowed us to test multiple hypotheses and explore the effects of automation quality and cues on human visual search.

H_1 **The addition of cues (perfect/imperfect) produces higher accuracy performance relative to the control condition with no cues.** When a visual stimulus is available on the screen, the participants will pay attention to it, especially when it is generated by computers. This will increase participants' awareness, and they reach higher accuracy [14].

H_2 **The addition of cues produces lower response times (i.e., users will find them faster) relative to the control condition with no cues.** The stimulus directs the attention of the users towards the target, reducing the search time.

H_3 **The benefits of imperfect cueing will be restricted to the frequent occasions when the cue is correct. In rare trials, when the cue is wrong, performance will be worse than when there is no cue at all.** Previous studies showed that users tended to follow the instructions and assistance provided by ML [5]. When there is an error in the ML algorithm, it will take time for users to realize that the system failed, and they must change their decisions. This process would decrease the overall search performance.

3.2 Participants

A total of 53 participants attended our studies. Each participant was assigned either of 2 conditions (perfect or imperfect cueing). There were 25 participants (17 male, 7 female, and 1 non-binary) with imperfect cueing and 28 (16 male, 12 female) with perfect cueing. Participants consisted of students and staff from Colorado State University, as well as people who were not affiliated with the university. There were 47 participants still studying at a university. The participants' ages ranged between 18 and 55 (M = 25.77, STD = 7.63). Among the 53 participants, 78.15% of the people reported that they didn't play any games regularly, and 84.62% of participants had either used AR or VR in the past. Participants received compensation in the form of $20 or class credits. None of the participants had participated in a prior study of cueing in our lab. All participants had normal or corrected-to-normal vision. This study was approved by the university IRB.

3.3 Materials

Fig. 2: OAK-1 camera mounted on XR-3.

Participants completed the experiment using the Varjo XR-3 [38] video pass-through AR HMD. The Varjo XR-3 headset can display 3D content over a 90Hz video feed of the real-world environment taken by

the headset's cameras. The field of view (FOV) of the device is 115° by 90°. The Varjo XR-3 was selected due to its FOV. To track the objects, an OAK-1 [24] camera was mounted to the front of the Varjo XR-3 (see Fig. 2). The OAK-1, developed by OpenCV [25], has dedicated hardware for running neural networks on the device and has a field of view of 68.8° by 42.75°. The camera weighs 53.1g with dimensions of 36 by 54.5 by 27.8 mm and did not block the Varjo sensors. The neural network used was YOLOv5-Nano, which was trained from scratch on a synthetic dataset of 40,000 images. The synthetic dataset was created in Unity 2020.3.27 using 3D models of the 40 target objects + 3 practice objects. The study software was developed on Intel core i9-9900K 3.60GHz, 64GB RAM, and RTX 2080Ti graphics card.

3.4 Machine Learning for Imperfect Cueing

A convolutional neural network called YOLOv5 [15] was trained on a fully synthetic dataset with 43 classes. These 43 classes were the target objects (3 practice objects and 40 trial objects). The model never saw a real picture of the targets it searched for until the classification phase. The system used for real-time object detection is YOLOv5-Nano due to its speed and ease of deployment on the OAK camera [24]. The external camera is used to capture the scene, perform neural network inference on the device, and then finally send the coordinates of the objects to the headset. The frames captured by the camera are first down-scaled to 576 x 374 before being sent to the model for inference.

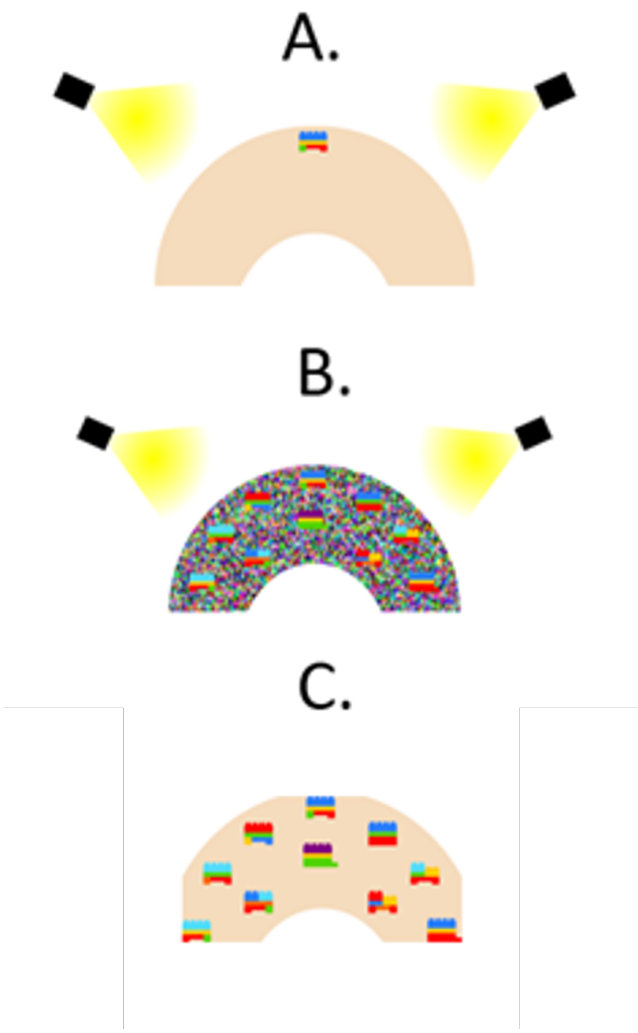

Fig. 3: Tweaking different parameters A. Not being exposed to all blocks at once, B. confusing background images in the training set, C. not enough lighting in the training set.

40,000 synthetic images were generated using a custom Unity program. The program took 3D models of target objects as input. It created a dataset consisting of images of these models in different orientations and with different backgrounds. The ability to tweak specific parameters allowed us to test subjects' responses to different errors created by AI such as not being exposed to all blocks at once (see Fig. 3A), confusing background images in the training set (see Fig. 3B), or not enough lighting in the training set (see Fig. 3C). The objective was to have a classification accuracy of at least 83% with the understanding that in a live system, this would never be constant. During the study, the average recognition across all the subjects was 88.9%. The recognition rate reported is the rate observed during actual trials, not during the testing phase.

The model took in 30,000 training images. The remaining 10,000 images were used as validation images. The model was trained for 3 hours over 300 epochs. Neural network inference on the OAK camera took 49.5ms on average (3.4ms Std Dev). Because of this, we limited the speed of the camera to 20fps. The data is sent over sockets from the OAK camera to Unity, where it is rendered for the headset. This process takes around 10ms, so there is very little delay on the system. The decision for YOLOv5 was to have a real-time recognition system that would be able to run in the computer described in Sec. 3.3.

During the actual experiment, the model continuously ran the detection algorithm. This meant that if initially, the model predicted the wrong location for the target object, it would correct itself if the user looked in the correct direction (i.e., the object is in the FOV of the camera). Pilot tests were conducted to check the accuracy of the model.

The coordinates of each object in the 3D environment were stored in the system at the start of the experiment. When the ML algorithm detected the target object in the frame, the location was projected out of the frame in the forward vector direction onto the 3D environment. This gave an approximate position of the target image. Since the coordinates of each object are stored in the system, the closest coordinate was selected as the target object. A threshold of 13 cm in the 3D environment was selected based on pilot study data.

3.5 Task

Participants completed a visual search task in 3D space in which they were asked to locate real-world objects (Mega Bloks [20], as shown in Fig. 4b.), with and without the aid of a target cue presented via the Varjo XR-3 AR HMD video-pass through. Participants were told in the instructions that the system is not 100% reliable and may fail and that they should confirm the suggestion presented by the system. Each trial consisted of three steps (see Fig. 1):

1. Target image is shown - an image of the target object is shown to the participants, and they press a button to continue when they feel comfortable recalling the object.
2. Searching - participants search the 3D environment for the designated object; there may be a cue or no cue depending on the trial condition.
3. In case of cue condition, selecting or rejecting the suggested object by the cue and searching for the target on their own. The target cue condition consisted of a 3D arrow pointing to the target object's location in 3D space.

The forward vector on the headset is the cursor. When no cues were presented, the center of the screen had a red dot representing the cursor. The red dot is replaced by cues in the cue conditions. When the participants looked at an object, the cursor changed to a circle enveloping the object. The participants held a Vive controller whose trigger button was used for selecting the target. The circle indicated that pressing the trigger would select the object. The red circle represents the system's predicted target. An ML selection was considered as rejected when the participants did not select the suggested object.

This was a mixed-subject design where the groups were split based on different levels of cue reliability (between subjects), and each group had two conditions, cue versus no cue (within-subjects, repeated measures). The two levels of cue reliability were: 1) the locations of all

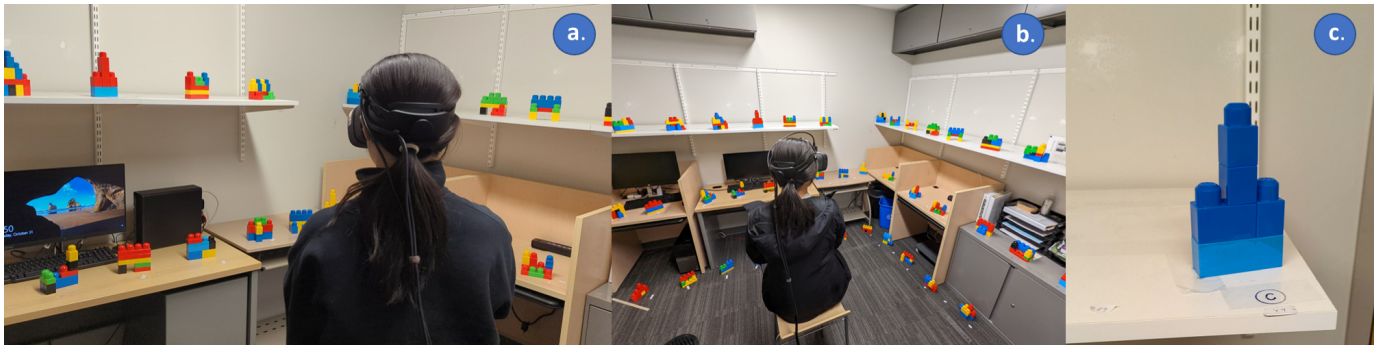

Fig. 4: a. Head-level view from behind a participant; b. Top-level view from behind a participant; c. Example object used in the study.

objects were known, and the cue always pointed at the correct object, and 2) the objects were located using ML and had an average accuracy of 88.9%. Participants did not receive a cue to help find the target object for the no-cue condition (i.e., the control condition). Thus, participants in the two groups had identical experiences in the no-cue condition. The search field of the room incorporated 180° surrounding the participant when looking forward from the chair in which they were seated. A total of 40 potential target objects + 3 practice objects were uniformly distributed across the 180° search field in the horizontal direction and approximately 15° in the vertical direction relative to the chair. Objects were placed at three different height levels as shown in Fig. 4: ground, table, and shelf level (separation between levels was 28in).

3.6 Design and Procedures

First, participants signed consent forms and verified their vision. Before starting the experiment, the participants were asked to complete a pre-experiment survey. Next, participants were seated in the middle of the room and assigned either cue with ML or perfect cueing. Every participant was assigned the condition of no cues. The participants completed three practice trials. Participants were presented with either cue or no cue in blocks of 40 trials. The order of these two blocks was counterbalanced across participants. The 40 trials within each block presented 40 different target images in random order. Once the practice trials were done, participants had to find 40 target objects. After attempting to find all 40 objects, the next condition was presented (if a cue was presented first, then no cue for the second condition and vice versa.) In the two blocks, the 40 target objects were the same (40 objects with no cue condition and the same 40 objects with cue condition). In the end, the participants were asked to complete a post-experiment survey. The entire experiment consisted of 86 trials and lasted approximately 35 minutes.

4 Results

This study employed a 2 (cue reliability level) x 2 (cue condition) mixed analysis of variance (MANOVA) design to explore the effects of automation reliability on performance and response time. Specifically, the design included two within-subject conditions of cued and not cued and two between-subject conditions of perfect and imperfect automation cueing. The within-subject factor of cued or not cued refers to whether or not participants were given a cue to locate the objects. The cue was intended to guide participants toward more accurate performance. Automation reliability for imperfect conditions was significantly lower than perfect conditions ($t(51) = 8.03$, $p < .001$; mean perfect automation = 100%, mean for imperfect automation = 88.9%).

4.1 Accuracy

The effects of cue reliability and cue condition on the accuracy of finding the objects are shown in Fig. 5. The data was normalized from 0 to 1 (e.g., 0.9 represents 36 correct selections out of 40) on the y-axis. The x-axis has 4 conditions grouped into 2 groups: ML cues and perfect cues. Each of these groups has 2 conditions, with cue and without cue.

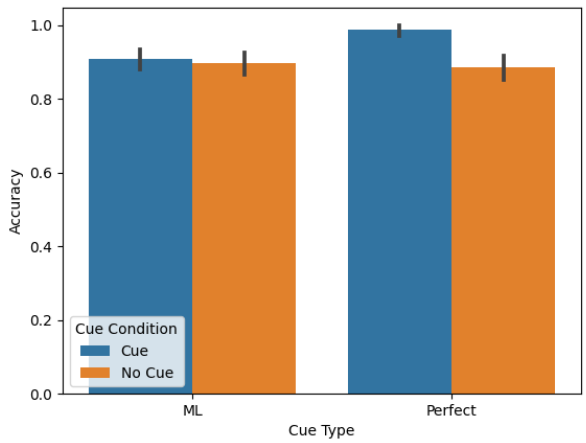

Fig. 5: Percent accuracy of finding the target when the cues are presented (Blue) vs. no cues (Orange); Different cue conditions (ML vs. Perfect Cue); The bars are standard error bars.

A mixed model ANOVA on these data revealed that there was a significant benefit of cueing (Wilks' lambda = .67, $F(1,51) = 24.83$, $p < .001$), (No Cue: M = 89%; Cue: M = 95%). This effect supports H_1, "The addition of cues (perfect/imperfect) produces higher accuracy performance relative to the control condition with no cues." The ANOVA also revealed a significant effect of Groups ($F(1,51) = 4.07$, $p < .05$).This effect can be further understood in the context of the highly significant interaction between cueing and groups (Wilks' lambda = .76, $F(1,51) = 16.15$, $p < .001$). The interaction revealed no difference in accuracy between the two groups in no cue search but a large cueing benefit for the perfect (M = 98.7%) over the imperfect (M = 90.7%) in cued search ($t(51) = 4.98$, $p < .001$). This result supports H_3, "Imperfect cueing automation negatively impacts the overall search performance due to high erroneous searches by humans on those imperfect trials."

When comparing accuracy with imperfect cues (M = 90.7%, SD = 2.85) vs. no cues (M = 89.6%, SD = 3.16), there was no significant difference ($t(24) = 1.05$, $p = .31$). On the other hand, for the perfect cueing group, accuracy with the cue (M = 98.75%), was significantly higher than with no cue (M = 88.48%; $t(27) = 5.3$, $p < .001$). Thus, the cueing benefit to overall accuracy was only observed when the cue was perfect.

Although there was a relatively large number of unique items to search for, one question is whether performance changed as a function of encountering search for the same objects twice across the experiment. Here we looked at changes in performance from Block 1 to Block 2, which includes both potential changes from a second exposure to items in the search task as well as any changes in general task learning across

the session. When comparing the accuracy of finding objects, aided by imperfect cues, in Block 1 (M = 90.4%, SD = 2.82) compared to Block 2 (M = 89.9%, SD = 3.19), there was no significant difference (t(24) = 0.23, p = .82). When comparing the accuracy of finding objects, aided by perfect cues, in Block 1 (M = 94.28%, SD = 4.25) compared to the Block 2 (M = 92.95%, SD = 2.61), there was no significant difference (t(27) = 0.49, p = .62). Given that there was no evidence in this analysis of changes across the experiment, this is congruent with viewing the specific Lego Block shapes twice did not have any significant influence on performance.

4.2 Response Time

Response time for the four conditions is plotted in Fig. 6. A corresponding MANOVA to that carried out with accuracy revealed a main effect of faster responses for cued trials (M = 5.9s) than without (9.4s) (Wilks' lambda = .43, F(1,51) = 66.67, $p < .001$). This effect supports H_2, "The addition of cues produces lower response times relative to the control condition with no cues." There was also a significant difference between the two groups (F(1,51) = 9.87, $p < .005$, partial eta squared = .16); however, as with accuracy, this difference can best be interpreted in the context of the significant interaction between the two variables (Wilks' lambda = .68, F(1,51) = 23.81, $p < .001$). Here, pairwise comparisons between response time in the no-cue conditions (the two orange bars in Fig. 6) revealed no difference, but between the cued trials, more rapid responses were seen for those who had a perfect cue (M = 3.74s) compared to those for whom the cue was imperfectly reliable (M = 9.2s) (t(51) = 4.90, $p < .001$).

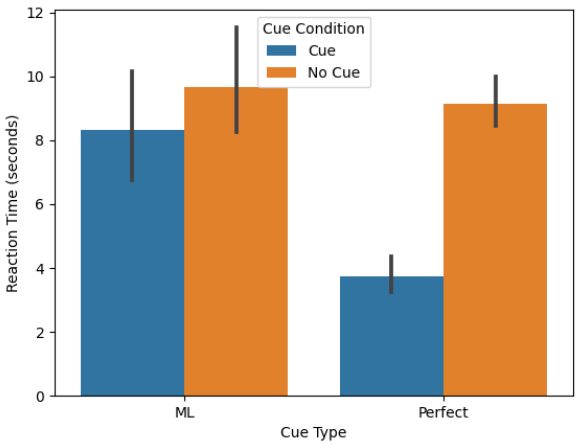

Fig. 6: Response times of finding the target when the cues are presented (Blue) vs. no cues (Orange); Different cue conditions (ML vs. Perfect Cue); the bars are standard error bars.

When comparing response time when imperfect cues were presented (M = 8.31s, SD = 3.98) vs. no cues (M = 9.67s, SD = 4.57), there was a small but significant benefit of cueing (t(24) = -2.12, $p < .05$). In contrast, when presented with a perfect cue (M = 3.74s, SD = 1.49) vs. no cue (M = 9.16s, SD = 2.18), there was a large and significant benefit of cueing (t(27) = -10.13, $p < .001$). Thus, the cueing benefit to overall RT was observed for both perfect and imperfect cued conditions.

Although there was a relatively large number of unique items to search for, one question is whether performance changed as a function of encountering search for the same objects twice across the experiment. Here, we looked at changes in performance from Block 1 to Block 2, which includes both potential changes from a second exposure to items in the search task as well as any changes in general task learning across the session. When comparing the response time of finding objects, aided by imperfect cues, in Block 1 (M = 9.07 sec, SD = 5.06) compared to Block 2 (M = 8.91 sec, SD = 3.47), there was no significant difference (t(24) = 0.23, p = .81). When comparing the response time of finding objects, aided by perfect cues, in Block 1 (M = 6.02 sec, SD = 3.33) compared to Block 2 (M = 6.88 sec, SD = 3.18), there was no significant difference (t(27) = 0.74, p = .47). Given that there was no evidence in this analysis of changes across the experiment, this is congruent with viewing the specific Lego Block shapes twice did not have any significant influence on performance.

4.3 The Speed-Accuracy Trade-off Function

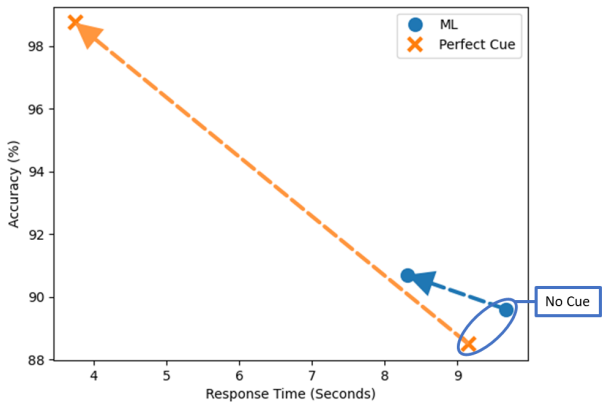

Fig. 7: In the speed-accuracy trade-off space, the combined impacts of imprecise cueing.

The combined effects of imperfect cueing on accuracy and response time – that is, on overall performance as stated in H_3 - are represented in the speed-accuracy trade-off space in Fig. 7. Within the figure, high performance (high accuracy, fast responses) can be seen in the upper left region, while poorer performance is in the lower right corner. Figure 7 first illustrates, with high prominence, the general equivalence in both measures of the no cue condition, the two circled points. Thus, any concerns about fundamental differences in search performance abilities between the two populations are eliminated. The differential impact of cueing on the two groups is illustrated by the two arrows. When the cues are perfect, their benefit to both dimensions of performance is pronounced, as illustrated by the long dashed orange arrow pointing to the upper left in Fig. 7. But for the imperfect cueing group, with this relatively small drop in cueing reliability, from 100% to 88.9%, the cues' benefit to accuracy is eliminated, and its benefit to processing speed, while not eliminated, is greatly reduced, as illustrated by the short dashed blue arrow.

4.4 Automation Bias

Section 1 described automation bias as the tendency to automatically follow the automation's recommendation (here, regarding where the target is). For statistical reasons, this tendency is revealed most clearly by examining participant responses when automation is wrong.

In Fig. 8, we have used the same speed-accuracy space as in Fig. 7, to depict the automation bias observed in the current data for the imperfect automation group. In the upper left is depicted the excellent performance when the cue was correct. In the lower right is the performance when the cue was wrong. The figure reveals the large and statistically significant loss in accuracy from 88.3% to 33.1% (t(23) = 8.15, $p < .001$) as well as the large increase in response time, from 7.3s to 16.32s (t(25.7) = 5.04; $p < .01$).Collectively, both of these effects further confirm H_3 "Imperfect cueing automation negatively impacts the overall search performance due to high erroneous searches by humans on those imperfect trials." Not only does performance suffer on all search trials with imperfect automation cueing, as shown in Fig. 7, but time and accuracy performance is particularly reduced when automation errors and the wrong object are cued.

The finding of the response time delay when automation is wrong illuminates the cognitive strategies employed by the participants. This

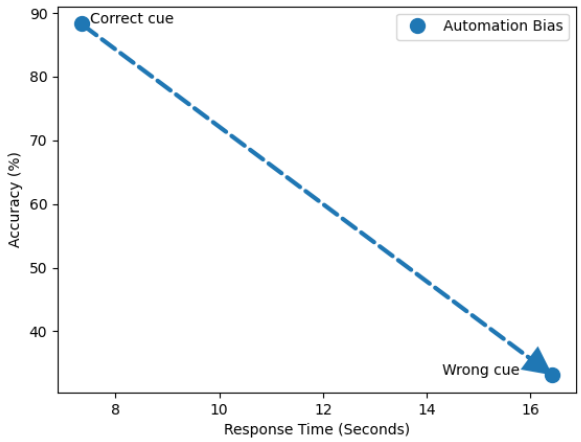

Fig. 8: In the speed-accuracy trade-off space, automation bias.

delay indicates participants did not always "blindly follow" the automation recommendation. Rather, when they saw a cue they thought was wrong, they tried to figure out the correct target (based on their imperfect memory of the image viewed at the beginning of the trial) but failed. Following that, they followed the automation recommendation nevertheless. This whole cycle took them an additional average of 11.4s.

The two strategies to be followed when the cue is wrong would seemingly produce a different pattern of effects of response time and accuracy. "Blindly following" (the automation bias) would keep response time at its very short value typical of the control condition (8.3s) or the cue-correct condition (7.3s), and produce a near 100% error rate. In contrast, noticing the error and triggering a "re-search" for the correct target would produce this prolonged delay. Accuracy would be improved but would be far from perfect because, by this time, the participant's memory of what the true target looked like would be degraded. We cannot estimate the proportion of trials on which each of these two strategies was deployed, nor the extent to which different participants or different trials for the same participant contributed to this difference in strategy choice, as this study was not intended for such analysis.

5 Discussion

The present study investigated the impact of cue reliability on the accuracy and response time during a visual search task presented in an AR environment. It was hypothesized that the addition of cues would produce higher accuracy performance (H_1) and lower response times (H_2) compared to the control condition with no cues consistent with the prior findings of [21, 26]. The study results provided partial support for the two previously presented hypotheses. Specifically, when the cue was consistently accurate, it significantly improved both speed and accuracy. However, in the case of imperfect cues (as observed in the ML condition), there was a significant benefit only to speed (reduced RT) but not to accuracy. Furthermore, the study also revealed that search performance is highly impacted by imperfect cueing automation. This was shown both by the lower performance and particularly longer time with imperfect-than-perfect cueing and by highly erroneous searches by participants on the trials when automation was wrong.

A task that required the integration of near domain (display cue) information with far domain (search field) information supported a 5.95% increase in the performance (accuracy) of searching for the target objects. The study's first hypothesis (H_1 - "The addition of cues (perfect/imperfect) produces higher accuracy performance relative to the control condition with no cues") was supported, as the addition of cues produced higher accuracy performance relative to the control condition with no cues. The use of cues can guide participants toward more accurate performance and lead to faster response times. This finding is consistent with previous research that has demonstrated the efficacy of AR cues in improving visual search performance (e.g., [21, 47]).

In comparing the different groups, there was only a 1.1% increase in accuracy performance when the cueing was not reliable. The observation that imperfect cues did not enhance accuracy could be explained by participants harboring doubts about the reliability of these cues. In such instances, they opted to rely on their own flawed perceptual judgment instead. In the case of perfect cue vs. no cue condition, we saw an increase of 10.27% in accuracy performance with the perfect cue. Looking at the mean accuracy performance of the group with perfect cues (M = 98.75%), the results suggest that when cues are perfect, the users tend to follow the cue. The participants who did not perform perfectly were probably influenced by the statement in the instructions about the system not being completely reliable. So, when they may have encountered objects they were unsure of, their trust in the system wavered.

The study's second hypothesis (H_2 - "the addition of cues produces lower response times relative to the control condition with no cues") was fully supported, as the addition of cues produced lower response times relative to the control condition with no cues. Participants who received cues responded (found the objects) significantly faster than those who did not receive cues. The use of cues can guide participants to find objects more rapidly. This finding is also consistent with previous research that has shown the efficacy of AR cues in reducing response times in visual search tasks (e.g., [39]).

Regarding H_3 – "The benefits of imperfect cueing will be restricted to the frequent occasions when the cue is correct. In rare trials when the cue is wrong, performance will be worse than when there is no cue at all." – there was an increase of 5.42s in response time when the cueing was not reliable. Such an increase observed is attributable to two factors: (1) the large increase in 11% of the trials when the automation was wrong (see Fig. 8), associated with double checking the ML-inferred target, and (2) the general slowing on all trials, associated with the appropriately increased caution by participants who realized the imperfection of the cueing automation. Hence, our hypothesis H_3 was supported.

The use of ML can simplify a lot of tasks and may be essential when the precise nature of the target cannot be specified in advance, but the reliability of such ML is not perfect. While ML can be a powerful tool for automating the detection of target objects, it is important to consider the potential for error. In this study, the average error of the ML model was 11.1%. Looking at just the accuracy of searching for the target, we may be able to interpret that the participants were blindly following the cues presented by the system in most cases (88 out of 112), even when we told each participant that the system may fail regardless of the condition (i.e., perfect vs. imperfect). This led to similar results between the imperfect cue and no cue conditions. But by exploring the results from their response times, we can paint a different picture. There was an increase in response times when the cue was wrong. The participants were able to figure out that the cue provided was false and started looking for the correct target. Some participants found the correct target after that, but most failed. This additional search took, on average, an additional 5 seconds.

Another interesting analysis is how our study compared the recent findings of incorrect cueing using WoZ by Warden et al. [40] to a new system with real-time ML cueing. Our study used a Varjo XR-3 AR video pass-through, which has a much bigger FOV compared to Hololens 2 used by Warden et al. [40]. Another difference is that we used a higher ML accuracy rate of 89% versus a lower accuracy (yet constant by using WoZ) of 83% accuracy. The procedures and participant population demographics were similar between the two studies. Human performance accuracy with the imperfect cues there was 92%. Using our approach with real-time ML was 90.7%. In both studies, the mean response time in the imperfect cueing condition was approximately 8 seconds.

Most critically, the drop in accuracy when the cue was wrong, reflecting the automation bias, was large and highly significant in both studies:

a drop to 15% was reported by Warden et al. [40] compared to 31% of our study. This underlines that a real system not only reproduces the automation bias from a WoZ study but, in this case, is worse. Another analysis was the response time on those automation error trials; both studies also revealed a lengthening, indicating that participants did not just blindly follow the incorrect automation guidance.

Furthermore, these results support the WoZ methodology as a reliable experimental approach for studying user interaction. It suggests that the controlled environment of WoZ experiments can accurately simulate real-world interactions, indicating the robustness of this method. We remain highly confident that additional research with ML cueing will continue to produce a pattern of results that are essentially equivalent to those where the cueing error is experimentally imposed (i.e., using WoZ) when the accuracy is around 83% [40, 47].

5.1 Limitations

Although the study used different shapes and color combinations of the Mega Bloks™, making them easier to distinguish from the natural environment, a study with actual objects that belong to the environment contextually may provide further insight (e.g., a pen in a home office). Another limitation concerns the ecological validity of the study. Many cases of cued search in the real world do not require comparing objects with an image to be detected. For instance, airport agents scanning luggage for a weapon are not comparing items with an image to be detected; rather, they have a long-term memory representation of what a typical weapon may be. Yet, the objective of our study was to search items to which participants have had no previous exposure, and the current population had not participated in the prior study of Lego block search [40]. Also, we did not use experts in visual searches, such as radiologists personnel. Future work may examine the effects of AR cueing with visual search experts. Finally, while the study may improve external validity by using a real-time ML system for visual search, it limits its internal validity due to the classification accuracy of the ML model presented to each participant varied by a small number. In a WoZ study, the classification accuracy is kept constant. However, as described in the Result and Discussion Sections, our study validates the control WoZ studies done by other researchers (e.g., [40, 47]), which also allows for rapid prototyping and iterative design, potentially saving time and resources during the early stages of research and application development.

5.2 Design Implications

The study has several implications for the design of cueing systems in visual search tasks. First, the study highlights the potential benefits of using cues to improve search performance and reduce response times, which can inform the design of user experiences. Our results can be used in real-world applications, and developers can create more intuitive and user-friendly interfaces.

However, the study also highlights the potential risks associated with imperfect automation, which can negatively impact overall search performance. To mitigate these risks, designers of cueing systems should consider the reliability of the automation and provide clear feedback to users on the reliability of the cues.

Additionally, designers should consider the potential for automation bias and design cues that encourage users to actively engage in the search task rather than relying solely on automation.

Finally, the study has implications for the use of ML in the design of cueing systems. While ML can provide a powerful tool for automating the detection of target objects, it is important to consider the potential for error and design cues that take into account the limitations of ML systems.

6 Conclusion and Future Work

The presented research study provides insightful information on the potential benefits and risks of using cues to improve visual search efficiency. This is in the context of imperfect automation. The study emphasizes the necessity for cueing system designers to carefully evaluate the reliability of automation. Taking this into account, they can develop cues that motivate users to actively participate in the search task. The study emphasizes the potential for ML as an effective tool for automating the identification of target objects. However, it also emphasizes the necessity of considering limitations and the potential for errors when in the design of cueing systems.

In future work, the study may be expanded to look at different objects (real-world objects that contextually belong to the environment) and add walking to the search. In addition, introducing multiple target objects and multiple cues may present interesting findings. For example, if you have two cues, is it possible to decrease automation bias? We believe that this may be possible through a combination of cues that further increase the trust and reliability of the systems being used and, therefore, increase user performance. It will also be interesting to investigate how predictive cue information influences user trust when guiding through a sequence of tasks. We also plan to analyze the effect of cueing error sources with a within-subjects experiment in the future.

Acknowledgments

This work was supported by NSF awards 2327569, 2238313, and ONR N00014-21-1-2949, N00014-21-1-2580.

References

[1] N. Barbotin, J. Baumeister, A. Cunningham, T. Duval, O. Grisvard, and B. H. Thomas. Evaluating visual cues for future airborne surveillance using simulated augmented reality displays. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 213–221, 2022. doi: 10.1109/VR51125.2022.00040 1, 2

[2] M. L. Bartlett and J. S. McCarley. Benchmarking aided decision making in a signal detection task. *Human Factors*, 59(6):881–900, 2017. PMID: 28796974. doi: 10.1177/0018720817700258 2, 3

[3] L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, and A. van den Oord. Are we done with imagenet? *CoRR*, abs/2006.07159, 2020. 2

[4] F. Biocca, A. Tang, C. Owen, and F. Xiao. Attention funnel: Omnidirectional 3d cursor for mobile augmented reality platforms. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, 8 pages, p. 1115–1122. Association for Computing Machinery, New York, NY, USA, 2006. doi: 10.1145/1124772.1124939 1

[5] M. J. Bitner, A. L. Ostrom, and M. L. Meuter. Implementing successful self-service technologies. *Academy of management perspectives*, 16(4):96–108, 2002. 3

[6] M. M. Boskemper, M. L. Bartlett, and J. S. McCarley. Measuring the efficiency of automation-aided performance in a simulated baggage screening task. *Human Factors*, 64(6):945–961, 2022. PMID: 33508964. doi: 10.1177/0018720820983632 2, 3

[7] M. M. Chun. Contextual cueing of visual attention. *Trends in Cognitive Sciences*, 4(5):170–178, 2000. doi: 10.1016/S1364-6613(00)01476-5 1

[8] S. Dodge and L. Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, pp. 1–7, 2017. doi: 10.1109/ICCCN.2017.8038465 2

[9] C. Drury. *Visual search in industrial inspection*. Taylor & Francis London, 1990. 1

[10] J. Goh, D. A. Wiegmann, and P. Madhavan. Effects of automation failure in a luggage screening task: A comparison between direct and indirect cueing. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 49(3):492–496, 2005. doi: 10.1177/154193120504900359 2, 3

[11] A. Gramopadhye, C. Drury, X. Jiang, and R. Sreenivasan. Visual search and visual lobe size: can training on one affect the other? *International Journal of Industrial Ergonomics*, 30(3):181–195, 2002. doi: 10.1016/S0169-8141(02)00099-9 1

[12] S. J. Henderson and S. Feiner. Evaluating the benefits of augmented reality for task localization in maintenance of an armored personnel carrier turret. In *2009 8th IEEE International Symposium on Mixed and Augmented Reality*, pp. 135–144, 2009. doi: 10.1109/ISMAR.2009.5336486 1

[13] K. A. Hoff and M. Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434, 2015. PMID: 25875432. doi: 10.1177/0018720814547570 2, 3

[14] T. Honda, N. Hagura, T. Yoshioka, and H. Imamizu. Imposed visual feedback delay of an action changes mass perception based on the sensory prediction error. *Frontiers in psychology*, 4:760, 2013. 3

[15] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, K. Michael, TaoXie, J. Fang, imyhxy, Lorna, Z. Yifu, C. Wong, A. V,

D. Montes, Z. Wang, C. Fati, J. Nadar, Laughing, UnglvKitDe, V. Sonck, tkianai, yxNONG, P. Skalski, A. Hogan, D. Nair, M. Strobel, and M. Jain. ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, nov 2022. doi: 10.5281/zenodo.7347926 4

[16] R. Kumaran, Y.-J. Kim, A. E. Milner, T. Bullock, B. Giesbrecht, and T. Höllerer. The impact of navigation aids on search performance and object recall in wide-area augmented reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, article no. 710, 17 pages. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3544548.3581413 2

[17] J. LEE and N. MORAY. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10):1243–1270, 1992. PMID: 1516577. doi: 10.1080/00140139208967392 2, 3

[18] J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004. PMID: 15151155. doi: 10.1518/hfes.46.1.50 3

[19] W. Lu, H. B.-L. Duh, S. Feiner, and Q. Zhao. Attributes of subtle cues for facilitating visual search in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):404–412, 2014. doi: 10.1109/TVCG.2013.241 2

[20] Mega bloks. https://shop.mattel.com/collections/mega-bloks. Accessed: 2023-03-24. 4

[21] D. Mifsud, C. Wickens, M. Maulbeck, P. Crane, and F. R. Ortega. The effectiveness of gaze guidance lines in supporting jtac's attention allocation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 66(1):2198–2201, 2022. doi: 10.1177/1071181322661143 3, 7

[22] K. L. Mosier and L. J. Skitka. Automation use and automation bias. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 43(3):344–348, 1999. doi: 10.1177/154193129904300346 2

[23] U. Neisser. Visual search. *Scientific American*, 210(6):94–103, 1964. 1

[24] Oak-1. https://store.opencv.ai/products/oak-1. Accessed: 2023-03-24. 4

[25] Opencv. https://opencv.org/. Accessed: 2023-03-24. 4

[26] R. Parasuraman and D. H. Manzey. Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3):381–410, 2010. PMID: 21077562. doi: 10.1177/0018720810376055 2, 3, 7

[27] D. K. Patil. *Something Is Fishy! - How Ambiguous Language Affects Generalization of Video Action Recognition Networks*. PhD thesis, Colorado State University, 2022. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-03-08. 2

[28] M. T. Phan, I. Thouvenin, and V. Frémont. Enhancing the driver awareness of pedestrian using augmented reality cues. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1298–1304, 2016. doi: 10.1109/ITSC.2016.7795724 2

[29] M. I. Posner. Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32(1):3–25, 1980. doi: 10.1080/00335558008248231 2

[30] M. I. Posner. Orienting of attention: Then and now. *Quarterly Journal of Experimental Psychology*, 69(10):1864–1875, 2016. PMID: 25176352. doi: 10.1080/17470218.2014.937446 2

[31] M. I. Posner, C. R. Snyder, and B. J. Davidson. Attention and the detection of signals. *Journal of experimental psychology: General*, 109(2):160, 1980. 2

[32] R. Sargent, B. Walters, and C. Wickens. Meta-analysis qualifying and quantifying the benefits of automation transparency to enhance models of human performance. In *Human-Computer Interaction. Theoretical Approaches and Design Methods: Thematic Area, HCI 2023, Held as Part of the 25th HCI International Conference, HCII 2023, Hybrid Event, July 23–July 28, 2023, Proceedings, Part I*. Springer, 2023. In print. 3

[33] M. Schemmer, N. Kühl, C. Benz, and G. Satzger. On the influence of explainable ai on automation bias, 2022. 2

[34] A. Seeliger, R. P. Weibel, and S. Feuerriegel. Context-adaptive visual cues for safe navigation in augmented reality using machine learning. *International Journal of Humanâ€“Computer Interaction*, 0(0):1–21, 2022. doi: 10.1080/10447318.2022.2122114 3

[35] B. Shneiderman. *Human-centered AI*. Oxford University Press, 2022. 2, 3

[36] B. V. Syiem, R. M. Kelly, J. Goncalves, E. Velloso, and T. Dingler. Impact of task on attentional tunneling in handheld augmented reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, article no. 193, 14 pages. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3411764.3445580 3

[37] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980. doi: 10.1016/0010-0285(80)90005-5 1

[38] Varjo xr-3. https://varjo.com/products/xr-3/. Accessed: 2023-03-24. 3

[39] A. C. Warden, C. D. Wickens, D. Mifsud, S. Ourada, B. A. Clegg, and F. R. Ortega. Visual search in augmented reality: Effect of target cue type and location. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 66(1):373–377, 2022. doi: 10.1177/1071181322661260 2, 3, 7

[40] A. C. Warden, C. D. Wickens, D. Rehberg, F. R. Ortega, and B. A. Clegg. Fast, accurate, but sometimes too-compelling support: The impact of imperfectly automated cues in an augmented-reality head-mounted display on visual search performance. *IEEE Transactions on Human-Machine Systems*, pp. 1–12, 2023. doi: 10.1109/THMS.2023.3302152 1, 2, 3, 7, 8

[41] C. D. Wickens, J. S. McCarley, and R. S. Gutzwiller. *Applied attention theory*. CRC press, 2023. 2

[42] J. M. Wolfe. Guided search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, 28(4):1060–1092, 2021. 1

[43] C.-C. Wu and J. M. Wolfe. Eye movements in medical image perception: A selective review of past, present and future. *Vision*, 3(2):32, Jun 2019. doi: 10.3390/vision3020032 1, 2

[44] C.-C. Wu and J. M. Wolfe. Eye movements in medical image perception: A selective review of past, present and future. *Vision*, 3(2):32, Jun 2019. doi: 10.3390/vision3020032 2

[45] S. Yantis. Stimulus-driven attentional capture. *Current Directions in Psychological Science*, 2(5):156–161, 1993. 2

[46] M. Yeh, J. L. Merlo, C. D. Wickens, and D. L. Brandenburg. Head up versus head down: The costs of imprecision, unreliability, and visual clutter on cue effectiveness for display signaling. *Human Factors*, 45(3):390–407, 2003. PMID: 14702991. doi: 10.1518/hfes.45.3.390.27249 3

[47] M. Yeh and C. D. Wickens. Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors*, 43(3):355–365, 2001. PMID: 11866192. doi: 10.1518/001872001775898269 3, 7, 8

[48] M. Yeh, C. D. Wickens, and F. J. Seagull. Target cuing in visual search: The effects of conformality and display location on the allocation of visual attention. *Human Factors*, 41(4):524–542, 1999. PMID: 10774124. doi: 10.1518/001872099779656752 3