The Cognitive Effects of Machine Learning Aid in Domain-Specific and Domain-General Tasks

Kristin Divis
Sandia National Laboratories[1]
kmdivis@sandia.gov

Breannan Howell
Sandia National Laboratories
bchowel@sandia.gov

Laura Matzen
Sandia National Laboratories
lematze@sandia.gov

Mallory Stites
Sandia National Laboratories
mcstite@sandia.gov

Zoe Gastelum
Sandia National Laboratories
zgastel@sandia.gov

Abstract

With machine learning (ML) technologies rapidly expanding to new applications and domains, users are collaborating with artificial intelligence-assisted diagnostic tools to a larger and larger extent. But what impact does ML aid have on cognitive performance, especially when the ML output is not always accurate? Here, we examined the cognitive effects of the presence of simulated ML assistance—including both accurate and inaccurate output—on two tasks (a domain-specific nuclear safeguards task and domain-general visual search task). Patterns of performance varied across the two tasks for both the presence of ML aid as well as the category of ML feedback (e.g., false alarm). These results indicate that differences such as domain could influence users' performance with ML aid, and suggest the need to test the effects of ML output (and associated errors) in the specific context of use, especially when the stimuli of interest are vague or ill-defined.

1. Introduction

The breadth of applications of machine learning (ML) intelligent systems in predictive decision-making scenarios is substantial, and users are interfacing with these assistant systems in a multitude of ways. This human-ML collaboration is seen in a variety of fields, from medical diagnosis and treatment [1,2,3], to automated driving systems [4,5], to threat detection [6]. ML assistance has also been utilized in aiding visual search and predicting the presence of targets in modalities such as baggage screens [7,8,9] and simulated combat images [10]. In the field of nuclear safeguards, ML diagnostic tools could also be applied to assist human users and enhance performance [11,12]. No model performs at 100% accuracy however, making the examination of how different errors from the ML diagnostic algorithm tools affect human decision making and performance crucial.

As in many of the application areas noted above, we are focused on ML assistance in the context of target detection (e.g., "Is there a tumor in this medical image?") rather than other common ML applications such as natural language processing. Previous research has found ML model errors do have a significant effect on human performance and trust in the system. The algorithms' performance on any given trial falls into four classifications drawn from classic signal detection theory [13]: *Hit*, *Miss*, *False Alarm* (*FA*), or *Correct Rejection* (*CR*).[2] See Table 1. A *Hit* occurs when there is an item of interest in an image that is correctly identified by the ML algorithm, while a *CR* is correct indication that no target is present. Algorithmic errors fall into the general categories of *Miss* and *FA*. A *Miss* occurs when there is an item of interest in the image, but the ML algorithm fails to identify it; while a *FA* occurs when the algorithm incorrectly identifies the presence of an item of interest when none exists in the image. In situations with multiple targets, these errors can combine (e.g., a "*Miss+FA*" occurs when the algorithm identifies the *wrong* item of interest in an image). Some researchers have found that human trust and use of a system is directly related to the system's performance in

[1] Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. SAND2021-11316 J.

[2] Note that these performance terms differ from what is commonly used in the ML development community (true positive, false negative, false positive, and true negative, respectively), but the signal detection terms are more commonly associated with user studies and therefore the terms that we adopted to describe this research.

URI: https://hdl.handle.net/10125/79358
978-0-9981331-5-7

these four categories [14,15]. Rice [10] found that an increase in automation reliability resulted in an overall improvement in human–automation performance relative to less reliable automation. The researchers also found automation false alarms primarily resulted in a degradation of compliance, while misses predominately resulted in a degradation of reliance. Reduced reliance in a system is a behavior typically associated with target-absent events (*CR*).

		Ground Truth:	
		Target Present	Target Absent
ML Output:	Target Present	*Hit*	*False Alarm (FA)*
	Target Absent	*Miss*	*Correct Rejection (CR)*

Table 1: Classic signal detection theory categories within the context of target presence ground truth and ML output.

While there is a breadth of research in trust in automation, the body of research lacks depth. One major drawback to many of these experimental designs is that they are highly complex, and therefore filled with possible confounding variables. Paradigms have included docking spaceships, running power plants, and finding targets in complex visual searches such as x-rays and combat aerial photographs [14,15,10]. These paradigms are context-specific and difficult to apply to other domains of human-machine teaming.

Different contexts might also impact how humans use ML outputs. In target detection tasks, ML systems are often intended to aid humans' visual search processes. Much of the research on human visual search has used simple laboratory stimuli. While some research suggests visual search relies on the same active scanning processes for both simple laboratory stimuli and complex, applied stimuli [16], there is other evidence that these differences in visual search paradigms could significantly impact human performance. This contrasting research argues that visual search performance during traditional, simple search tasks is not comparable to visual search performance in more real-world applications [17]. These differences in human performance by visual task type have not been investigated in relation to ML algorithms that are intended to support human visual search tasks.

The current study addresses these knowledge gaps using a domain-specific target detection task relevant to the field of international nuclear safeguards (Experiment 1), as well as a classic domain-general visual search task (Experiment 2). Simulated ML target detection indicators are implemented to investigate the effects of different ML error types on human decision making and performance for both real world images and a basic visual search task.

2. Experiment 1: Nuclear Cooling Towers

Experiment 1 investigated the effects of accurate and inaccurate algorithmic outputs on human performance in a visual search and decision making task using real-world images related to the domain of international nuclear safeguards. International nuclear safeguards are a set of measures and procedures implemented through bilateral and multilateral agreements between countries and the International Atomic Energy Agency (IAEA) to verify that a country's nuclear fuel cycle programs are not being misused for the development of nuclear weapons production. While many of the safeguards measures are verified in the field through activities such as material sampling, nondestructive assay of nuclear materials, and confirmation of facility design, others are conducted by analysts at the IAEA Headquarters in Vienna, Austria. An example of a Headquarters-based task is the collection and analysis of open-source text and imagery to confirm a country's declared nuclear operations (see [12])

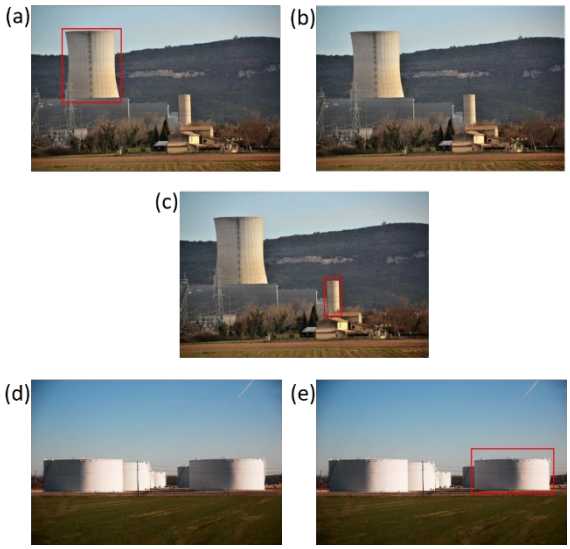

Figure 1: Examples of each type of ML category for Experiment 1: (a) *Hit*, (b) *Miss*, (c) *Miss+False Alarm*, (d) *Correct Rejection*, and (e) *False Alarm*. Images available via creative commons license [18,19].

Participants in Experiment 1 were tasked with identifying nuclear cooling towers in a set of stimuli that consisted of images of buildings and non-buildings, with and without nuclear cooling towers. Some participants received simulated ML assistance in the form of bounding boxes that were placed on the images. The participants were told that the bounding boxes were produced by a ML algorithm and that they were intended to help them find the nuclear cooling towers.

They were not given details on the type of ML algorithm used to create the output. In reality, the bounding boxes were added by the experimenters and manipulated such that they produced one of the following ML performance categories: *Hit*, *CR*, *Miss*, *FA*, or *Miss+FA* (see Figure 1 for examples).

Another subset of participants did not receive simulated ML assistance. They had to rely on their own visual search process to locate and identify nuclear cooling towers in the images.

We tested how the presence or absence of the simulated ML outputs and the different categories of ML performance affected the participants' accuracy and response times (RTs) for the task.

2.1 Methods

In this task, participants were shown photos and asked to indicate whether a nuclear cooling tower was present or absent. All photos were collected from Flickr and included images both with and without nuclear cooling towers (and were part of the set of images curated for [12]). In order to select an appropriate subset of stimuli for Experiment 1, we first tested a larger set of photos in a norming study. We then down-selected to 240 photos in which accuracy performance for the nuclear cooling tower detection task was not at ceiling (i.e., near perfect performance). The photos were balanced in Experiment 1 so that 60% of photos contained a nuclear cooling tower. The task took about 15 minutes to complete.

Participants who received the "ML assistance" were shown images that contained mock ML outputs. When the "ML" identified a target in an image, a red bounding box was placed around that item in the image. For target absent images, no bounding box was shown. No real ML algorithms were used—instead the bounding boxes were created by the research team to maximize experimental control. The simulated ML outputs were correct 80% of the time (112 *Hit* trials and 80 *CR* trials), with the remaining 20% of the trials being equally divided between three types of ML errors (16 each *Miss*, *FA*, and *Miss+FA* trials). All stimuli were counterbalanced across 6 between subject conditions, so that the same underlying target present image could appear as a *Hit*, *Miss*, or *Miss+FA*, and the same underlying target absent image could appear as a *CR* or *FA*.

A final control condition included the same images with no ML output present. Participants who did not receive ML assistance saw the same images as participants in the other condition, but without any bounding boxes.

Data were collected online using the Amazon Mechanical Turk (AMT) platform with the following criteria for participants: located in the USA, at least 95% prior approval rating, and previously completed at least 1,000 tasks on AMT. All tasks were identical from a participant's perspective during selection (and participants could only see and complete the task once), helping to ensure the only differences between conditions were those we directly manipulated. We chose to target a general population sample rather than a professional population for this task due to: (1) access to qualified participants (especially during the COVID-19 global pandemic) and (2) number of participants available. AMT has hundreds of qualified workers, whereas there are a limited number of professional nuclear safeguards analysts. This approach allows us to first establish effects in a general population before testing the resultant strongest hypotheses on the limited professional population in later studies.

A total of 284 participants signed up for this task; data from 51 participants were removed from analysis due to attention check trial accuracy below 80% or overall accuracy below 60%. Two hundred participants provided data for conditions with simulated ML output; 33 participants provided data for the condition without ML output. Additional individual trials (e.g., for a single image) were removed if the RT was above three standard deviations of that participant's median RT.

2.2 Results

All statistical tests reported here were held at an $\alpha = .05$ level (95% confidence interval). A mixed effects model (see [20]) predicting accuracy from the fixed effects of nuclear cooling tower target (present or absent) and ML output (present or absent), along with random intercepts for participant and underlying image (regardless of presence of additional ML output) revealed a significant crossover interaction between target presence and ML output ($Z = 21.48$, $p < .001$). The simple effects showed significantly higher performance on images with nuclear cooling towers than those without nuclear cooling towers when ML output was provided ($Z = 6.04$, $p < .001$) but significantly *lower* performance on images with nuclear cooling tower images than those without nuclear cooling towers when ML output was *not* provided ($Z = 2.38$, $p = .017$). See Figure 2a.

Next, we examined accuracy performance on each type of ML output category (*CR*, *FA*, *Hit*, *Miss*, and *Miss+FA*) for conditions in which ML output was provided. The equivalent for the control condition with no ML output is the simple effect of target presence (target present "miss" and target absent "correct rejection") in the analysis above. A mixed effects model predicting accuracy for trials with ML output from the fixed effect of ML category with random intercepts for

participant and underlying image (with Tukey correction for multiple comparisons) revealed significant differences ($Z > 1.64$, $p < .05$) for all ML performance class comparisons except the difference between *FA* and *Miss* ($Z = 2.63$, $p = .053$). *Hit* images had the highest performance, followed by *CR* and then *Miss+FA*. *Miss* and *FA* images led to the lowest performance. See Figure 2b.

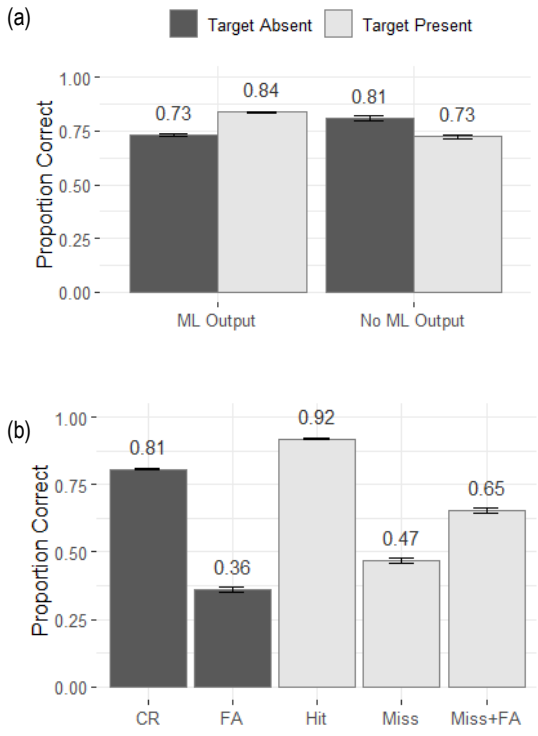

Figure 2: Proportion correct in Experiment 1 by (a) presence of target and ML output and (b) ML category for ML output conditions only. Error bars represent standard error of the mean calculated on a trial-by-trial basis.

We also examined the difference in RTs across the ML categories for correct responses. A mixed effects model predicting RT for accurate trials from the fixed effect of ML category with random intercepts for participant and underlying image (with Tukey correction for multiple comparisons) revealed significant differences ($Z > 1.64$, $p < .05$) between all image comparisons, with the exception of *Miss+FA* relative to *CR* ($Z = 2.46$, $p = .091$). When participants responded correctly, their responses were fastest for *Hits*, somewhat slower for *Misses*, and slowest for *FAs*. Their average response times for *CR* and *Miss+FA* trials were very similar to one another and fell between the average RTs for the *Miss* and *FA* trials. See Figure 3.

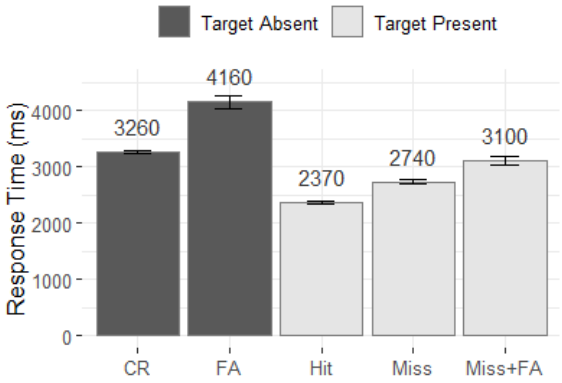

Figure 3: Response time in Experiment 1 by ML category for correct trials. Error bars represent standard error of the mean calculated on a trial-by-trial basis.

3. Experiment 2: T/L search

To investigate whether these findings also extrapolate to a basic visual search task, a second study was conducted. The experimental design in Experiment 2 remained consistent with Experiment 1. However, instead of searching for and identifying nuclear cooling towers in real-world images, the participants in Experiment 2 were tasked with searching for a "T" with a perfectly centered crossbar among a set of "L"s (asymmetrical distractors with offset crossbars) [21]. Also known as a conjunction search task, the T/L visual search task is a widely-used laboratory task that controls for the confounding variables that can occur in visual search tasks that use more naturalistic and complex stimuli. The T/L task is commonly used as a domain general analog of visual search and decision making for both novices and professionals from fields such as baggage screening and reviewing medical images (e.g., [22]). Once again, performance was assessed via accuracy and RT for both correct (*Hit* and *CR*) and incorrect (*Miss, FA, and Miss+FA*) ML indicators with the T/L-letter stimuli, as well as a control condition in which no ML output was provided.

3.1 Methods

The methodology for Experiment 2 was similar to that in Experiment 1, except the stimuli from the T/L task had blue bounding boxes when ML output was present (see Figure 4). The underlying stimuli were in grayscale, with letters in four shades of gray appearing on a cloudy background. Each image had 10 letters and the letters could appear in any orientation. One of the

letters was a perfect "T" on 60% of the trials. The mock ML output was correct 80% of the time and had the same proportion of error types as Experiment 1.

Participants saw a total of 120 images, with half of the participants receiving ML output and the other half in a control condition with no ML output. When the ML output was present, there were a total of 56 *Hit*, 40 *CR*, 8 *Miss*, 8 *FA*, and 8 *Miss+FA* trials.

Data were collected on the AMT online platform with similar data quality checks and cleaning as in Experiment 1. Data from a total of 72 participants were included in our analysis.

3.2 Results

Statistical analyses were conducted in the same manner as in Experiment 1. A mixed effects model predicting accuracy from the fixed effects of presence

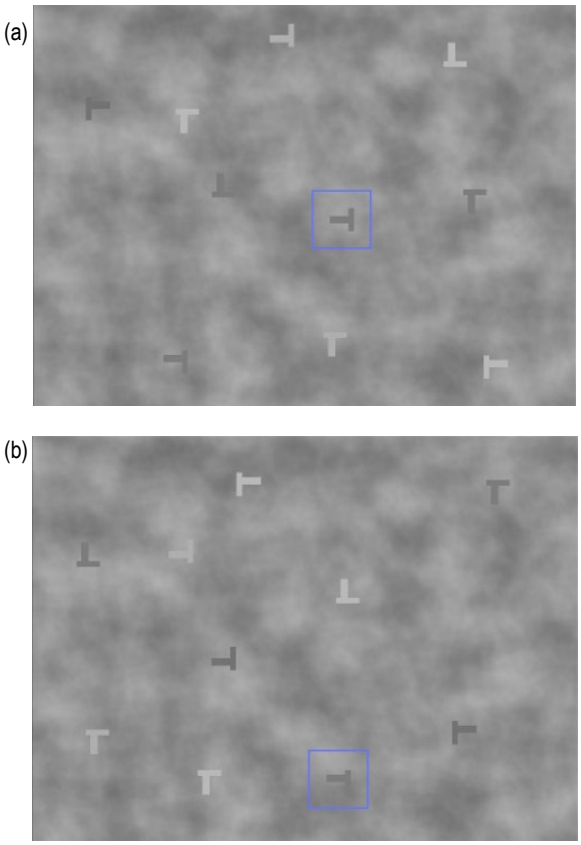

Figure 4: Examples of stimuli for Experiment 2 T/L task for (a) *Hit* (blue bounding box around perfect "T") and (b) *False Alarm* (blue bounding box around "L")

of a perfect "T" (target present *vs.* target absent) and ML aid (ML output *vs.* no ML output), along with random intercepts for participant and image, revealed significant main effects for the two fixed effects conditions. Accuracy performance was higher with than without ML output ($Z = 3.27$, $p < .001$). Accuracy performance was also higher for target absent trials relative to target present trials ($Z = 5.26$, $p < .001$). This pattern held regardless of whether ML aid was present. See Figure 5a.

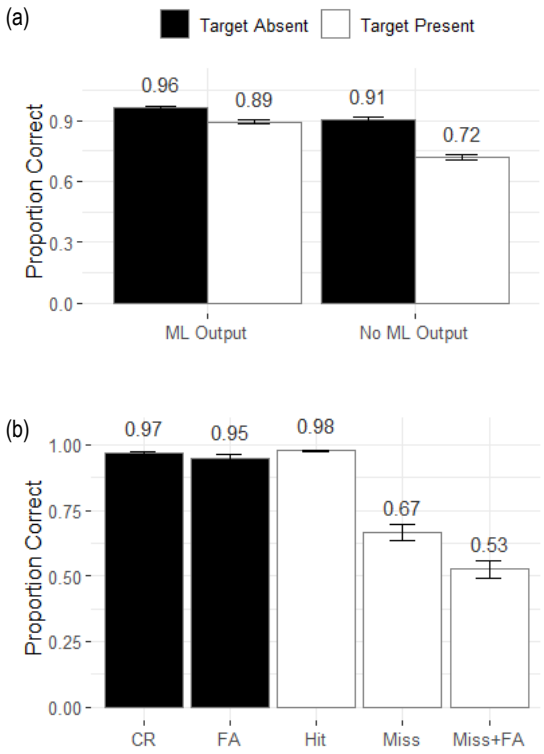

Figure 5: Proportion correct in Experiment 2 by (a) presence of target and ML output and (b) ML category for ML output conditions only. Error bars represent standard error of the mean calculated on a trial-by-trial basis.

Next, we compared performance across ML categories when ML aid was provided. A mixed effects model predicting accuracy for conditions in which ML output was provided from the fixed effect of ML category (*CR*, *FA*, *Hit*, *Miss*, *Miss+FA*), along with random intercepts for participant and base image revealed significantly lower performance for *Miss* or *Miss+FA* images relative to *CR*, *FA*, or *Hit* images (all $p < .05$, with Tukey correction for multiple comparisons). See Figure 5b.

We also investigated the effect of ML category on RT performance for correct trials. A mixed effects model predicting RT on accurate trials from the fixed effect of ML category with random intercepts for participant and underlying image (with Tukey correction for multiple comparisons) revealed significantly faster responses for *Hit* trials relative to all other types ($Z > 1.64$, $p < .05$) and significantly faster

responses to *Miss* trials relative to all other types ($Z > 1.64$, $p < .05$) except *Hit* trials. See Figure 6.

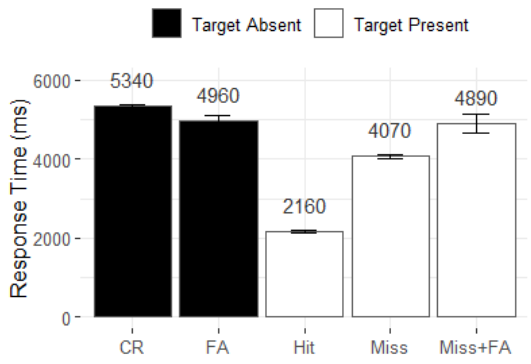

Figure 6: Response time in Experiment 2 by ML category for correct trials. Error bars represent standard error of the mean calculated on a trial-by-trial basis

4 Discussion

In this study, we found that aid from a simulated ML algorithm influenced decision makers' accuracy and response times. These differences were found in both a domain-specific international nuclear safeguards task where we asked participants to identify photos containing a nuclear cooling tower, as well as in a classic domain-general visual search task in which participants were asked to identify a perfect "T" amidst a field of asymmetrical "L"s. However, the nature of the differences in performance varied across the domain-specific and domain-general tasks.

The presence or absence of ML output led to a different pattern of accuracy performance for the two different tasks. For the T/L task, participants had higher accuracy for the target absent trials than for the target present trials, regardless of ML condition. Adding ML outputs improved participants' accuracy for both target present and target absent trials. The benefit of the ML outputs was higher for the target present trials (likely because the ML *Hits* made it easier to find many of the targets). However, participants still performed best on the target absent trials.

A very different pattern of performance appeared for the nuclear cooling tower task. Once again, participants performed better on target present trials when ML aid was provided. However, providing ML outputs in the nuclear cooling tower task led to *worse* performance on the target absent trials. The breakdown of performance by ML output category shows that this decline in accuracy is driven by the trials where the ML output was a *FA* (i.e., when the ML output incorrectly indicated that a nuclear cooling tower was present for a target absent trial). In the T/L task, participants rarely endorsed the ML's *FA*s. They correctly indicated that there was not a "T" present on 95% of the *FA* trials. In contrast, the participants frequently went along with the incorrect ML output in the nuclear cooling tower study, responding correctly on only 36% of *FA* trials. The participants also had very long average RTs for the *FA* trials in the nuclear cooling tower study, indicating that it took them a relatively long time to decide on a response. However, they ultimately agreed with the ML output in most cases.

When considering the other two types of ML errors (*Miss* and *Miss+FA*), we also see different patterns of performance across the two experiments. In the T/L task, participants had numerically higher accuracy for the *Miss* trials than for the *Miss+FA* trials, although this difference was not statistically significant. In the nuclear cooling tower task, participants had significantly higher accuracy for the *Miss+FA* trials than for the *Miss* trials. Both of these trial types contain a target, but the ML output is either absent or placed incorrectly around a different item in the image. In the T/L task for *Miss+FA* trials, participants can quickly evaluate the letter inside of the bounding box and determine that it is not a perfect "T". If they do not search the rest of the image to see if a "T" is present elsewhere, they will answer incorrectly. Participants performed much better on items in this condition in the nuclear cooling tower task. In this case, there are two routes that may have led them to the correct answer. First, they may have searched the rest of the image and spotted the nuclear cooling tower (similar to participants who answered correctly in the T/L task by searching the rest of the image to find the "T"). Alternatively, the participants in the nuclear cooling tower experiment may have trusted the ML output and responded accordingly. As with the *FA* trials, they may have been more likely to go along with what the model was telling them.

While complex visual search tasks are assumed to rely on the same visuospatial abilities as simple visual search, there are many factors that may be leading to these differences in our current findings. Researchers have argued that the decision-making component of search and decision tasks may play a stronger role in complex visual stimuli detection tasks [23]. As seen in complex visual search tasks such as X-ray image inspection [24,25], participants in the current study could have been relying more heavily on top-down processes of object recognition decision making to identify more difficult cooling tower targets. When comparing a simple conjunction task to an X-ray inspection task, researchers found different underlying visual-cognitive processes were predictive of performance, as well as little overlap in performance in the two tasks [17]. These differences in visual-cognitive abilities and top-down processing demands could be

contributing to our current findings associated with performance in response to ML aid in our two tasks. This highlights the importance of studying similarities and differences between domain-general and domain-specific tasks. The T/L task is one of the most commonly used domain-general analogs for professional visual search applications, but the highly controlled nature of the task can also lead to reduced generalizability to more realistic domain-specific tasks.

What makes the *FA* images more difficult to detect in the nuclear cooling task (lower accuracy and slower response times) compared to the T/L task? Participants' familiarity with the target item, as well as the level of ambiguity of the stimuli, may play a role. For our stimuli set, the degree of difference between a perfect "T" and an "L" is always large and quite clear. However, our more realistic application using nuclear cooling tower images is not always as clear. Our participants are unlikely to be as familiar with nuclear cooling towers as they are with the letter "T", and therefore may not have as strong of a mental model of just what constitutes a nuclear cooling tower (relative to a perfect "T"). Furthermore, the photos of nuclear cooling towers are also much more complex and varied. These concerns are especially important for less prototypical examples of nuclear cooling towers. What if the nuclear cooling tower is partially dismantled? Or if the photo captures a plume above a building on the horizon that looks very similar to plumes seen with nuclear cooling towers? A common—and critical—aspect of these more realistic, domain-specific applications is that the signals are often noisier and more ambiguous than typically seen in highly controlled, domain-general lab tasks.

In both the nuclear cooling tower and T/L tasks, participants must search for a potential target candidate. But in the nuclear cooling task, *deciding* whether a target candidate is indeed the sought-after item plays a large role [see 26]. The longer RTs for *FA* images are consistent with pausing longer to decide whether the item identified by the ML algorithm is in fact a nuclear cooling tower. Despite the pause, frequently participants ultimately agreed with the inaccurate ML output, as indicated by the low accuracy performance for *FA* images.

In a classic study on trust in automation, Lee and Moray [27] noted that operators tend to rely on automation (in our case, ML aid) more when trust exceeds self-confidence. We posit that identifying a nuclear cooling tower is inherently more ambiguous and challenging than identifying a perfect "T" amidst asymmetrical "L"s. When given ML output for nuclear cooling towers, participants may put more trust in that output (due to their lack of self confidence in knowing exactly what constitutes a nuclear cooling tower). Therefore, they may also more heavily rely on the ML aid *and* use it to help tailor their understanding of appropriate nuclear cooling tower characteristics.

The need to develop a mental model to better understand edge cases around what constitutes a classification type (e.g., "nuclear cooling tower") is not limited to tests of the general population. Professionals using ML aid to better perform their tasks must also develop a mental model of key traits during training and through experience. If a newly minted nuclear safeguards professional works with ML aid while learning to inspect images, she may rely on that model output to help her learn how she *should* be categorizing items (e.g., she might incorrectly add features from false alarms to her mental model). Further research is needed to tease out the consequences of ML errors for varying levels of expertise and types of tasks.

For example, future studies could delve into the role of ambiguity of and familiarity with the stimuli by (1) running a similar study with nuclear cooling tower images using participants who are quite familiar with identifying nuclear cooling towers (e.g., from the nuclear safeguards community) and (2) running a similar study with the T/L task where the difference between a perfect "T" and "L" is reduced, making it much more difficult to perceive whether the letter stem is perfectly centered ("T") or slightly offset ("L"). Future work could also delve into the role of more complex and noisy images (e.g., providing variants of nuclear cooling towers in more simplistic, controlled stimuli).

By understanding the underlying causes and roles of differences between applied, domain-specific tasks and highly controlled, domain-general tasks, we can better understand and predict the cognitive effects of ML errors for target detection tasks in new application spaces.

10. References

[1] K. Goddard, A. Roudsari, and J.C. Wyatt, "Automation bias: Empirical results assessing influencing factors", *International Journal of Medical Informatics, 83(5)*, 2014, pp. 368-375.

[2] M. Jacobs, M.F. Pradier, T.H. McCoy, R.H. Perlis, F. Doshi-Velez, and K. Z. Gajos, "How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection", *Translational Psychiatry*, *11*(1), 2021, pp. 1-9.

[3] B. Vasey, S. Ursprung, B. Beddoe, E. H. Taylor, N. Marlow, N. Bilbro, P. Watkinson, and P. McCulloch, "Association of clinician diagnostic performance with machine learning–based decision support systems: A systematic review", *JAMA Network Open*, *4*(3), 2021, p. e211276.

[4] M. Körber, E. Baseler, and K. Bengler. "Introduction matters: Manipulating trust in automation and reliance in automated driving", *Applied Ergonomics, 66*, 2018, pp. 18-31.

[5] H. Azevedo-Sa, H. Zhao, C. Esterwood, X.J. Yang, D.M. Tilbury, and L.P. Robert Jr, "How internal and external risks affect the relationships between trust and driver behavior in automated driving systems", *Transportation Research Part C: Emerging Technologies*, *123*, 2021, p. 102973.

[6] N. Du, K.Y. Huang, and X.J. Yang, "Not all information is equal: Effects of disclosing different types of likelihood information on trust, compliance and reliance, and task performance in human-automation teaming", *Human Factors, 62(6)*, 2020, pp. 987-1001.

[7] A. Chavaillaz, A. Schwaninger, S. Michel, and J. Sauer, "Expertise, automation, and trust in X-ray screening of cabin baggage", *Frontiers in Psychology, 10:256*, 2019, pp. 1-11.

[8] N.A. Andriyanov, A.K. Volkov, A.K. Volkov, and A.A. Gladkikh, "Research of recognition accuracy of dangerous and safe x-ray baggage images using neural network transfer learning", *IOP Conference Series: Materials Science and Engineering*, *1061(1),* 2021, p. 012002.

[9] T. Rieger, L. Heilmann, and D. Manzey, D. "Visual search behavior and performance in luggage screening: effects of time pressure, automation aid, and target expectancy", *Cognitive Research: Principles and Implications*, *6*(1), 2021, pp. 1-12.

[10] S. Rice, "Examining single-and multiple-process theories of trust in automation", *The Journal of General Psychology, 136(3)*, 2009, pp. 303-322.

[11] Z.N. Gastelum, L.E. Matzen, M.C. Stites, A.P. Jones, M.C. Trumbo, B.C. Howell, and M. Higgins, "Evaluating the cognitive impacts of errors from analytical tools in the international nuclear safeguards domain", *Proceedings of the Institute of Nuclear Materials Management Annual Meeting*, July 2020.

[12] Z.N. Gastelum and T.M. Shead, "Inferring the operational status of nuclear facilities with convolutional neural networks to support international safeguards verification", *INMM-2018: Proceedings of the Institute of Nuclear Materials Management Annual Meeting*, *XLVI*(*3*), 2018, pp. 37-47.

[13] N.A. Macmillan & C.D. Creelman. *Detection theory: A user's guide*, Lawrence Erlbaum Associates, New Jersey, 2005.

[14] M.T. Khasawneh, S.R. Bowling, X. Jiang, A.K. Gramopadhye, and B.J. Melloy, "A model for predicting human trust in automated systems", *Proceedings of the 8th Annual International Conference on Industrial Engineering: Theory, Applications, and Practice*, Las Vegas, NV, 2003, pp. 216-222.

[15] S.M. Merritt and D.R. Ilgen, "Not all trust is created equal: Dispositional and history-based trust in human-automation interactions", *Human Factors, 50(2)*, 2008, pp. 194-210.

[16] R.G. Alexander and G.J. Zelinsky, "Visual similarity effects in categorical search", *Journal of Vision, 11(8):9*, 2011, pp. 1-15.

[17] N. Hättenschwiler, S. Merks, Y. Sterchi, and A. Schwaninger, "Traditional visual search vs. X-ray image inspection in students and professionals: Are the same visual-cognitive abilities needed?", *Frontiers in Psychology, 10:525*, 2019, pp. 1-17.

[18] J. Menjoulet, "The house and the nuclear plant", [Photograph], *Flickr*, 2018, Creative Commons License, https://www.flickr.com/photos/jmenj/15091518102

[19] E.A. Rogers, "Storage", *Flickr*, 2011, Creative Commons License, https://www.flickr.com/photos/reallyboring/6784162887

[20] D. Bates, M. Maechler, B. Bolker, and M. Magnor, "Fitting linear mixed-effects models using lme4", *Journal of Statistical Software, 67(1)*, 2015, 1-48.

[21] A.M. Treisman and G. Gelade, "A feature-integration theory of attention", *Cognitive Psychology, 12(1)*, 1980, pp. 97-136.

[22] A.T. Biggs, M.S. Cain, K. Clark, E.F. Darling, and S.R. Mitroff, "Assessing visual search performance differences between Transportation Security Administration Officers and nonprofessional visual searchers", *Visual Cognition*, *21*(3), 2013, pp. 330-352.

[23] S.M. Koller, C.G. Drury, & A. Schwaninger. "Change of search time and non-search time in X-ray baggage screening due to training", *Ergonomics, 52(6)*, 2009, pp. 644-656.

[24] J. M. Wolfe, "What do 1,000,000 trials tell us about visual search?", *Psychological Science, 9(1)*, 1998, pp. 33-39.

[25] J.M. Wolfe & M.J. Van Wert, "Varying target prevalence reveals two dissociable decision criteria in visual search", *Current Biology, 20(2)*, 2010, pp. 121-124.

[26] C.F. Nodine and H.L. Kundel, "Using eye movements to study visual search and to improve tumor detection", *Radiographics*, *7(6)*, 1987, pp. 1241-1250.

[27] J.D. Lee and N. Moray, "Trust, self-confidence, and operators' adaptation to automation", *International Journal of Human-Computer Studies*, *40*, 1994, pp. 153-184.