NOT ALL ERRORS ARE CREATED EQUAL: EXAMINING HUMAN-ALGORITHM SYSTEM PERFORMANCE FOR INTERNATIONAL SAFEGUARDS-INFORMED VISUAL SEARCH TASKS

Zoe N. Gastelum, Laura E. Matzen, Kristin Divis, Mallory C. Stites, Breannan C. Howell
Sandia National Laboratories
Albuquerque, NM, USA

Abstract: The International Atomic Energy Agency has expressed interest in deep learning models to support information processing for multiple safeguards verification activities, including surveillance data review and open source information monitoring. Modestly performing deep learning models have been shown to increase performance of human-algorithm systems, and in some domains deep learning models have exceeded performance levels of humans working alone. Yet even the best performing humans and algorithms make errors. Sandia National Laboratories is currently investigating a breadth of variables that impact human-algorithm system performance, focusing on model errors and user trust. In this paper, we will present results from two experimental tracks examining how error types and error frequencies from simulated deep learning models for a safeguards-informed object detection task impact user performance.

TESTING HUMAN PERFORMANCE & VISUAL SEARCH FOR SAFEGUARDS

Machine learning algorithms that support visual tasks (e.g., autonomous driving) are being increasingly adopted for everyday uses. As a result, there has been growing interest in the safeguards community regarding machine learning implementation to support a number of visual search and identification safeguards tasks. Example research areas in machine learning applications to visual search for safeguards include:

- Detecting and tracking safeguards-relevant objects in surveillance camera imagery [1];
- Detecting and categorizing activity from overhead imagery [2]; and
- Retrieving and assessing relevance of images in open source data streams [3; 4].

The outcomes of visual safeguards verification activities are highly consequential, such as requests for information from a state, re-visitation to a site for clarification, or complementary access visits. Given the potential negative ramifications from an error during a visual safeguards verification activity – on IAEA staff as well as on state parties – it is imperative to understand not just the performance of machine learning algorithms that can support visual search and identification, but also the interaction between those algorithms and their human users. Specifically, our research seeks to understand how users are impacted when machine learning algorithms provide incorrect results.

Over the course of two years, our research team at Sandia National Laboratories has studied multiple aspects of human interaction with errors from machine learning [see 5]. In this paper, we present the results of two tracks of experimental research. First, we describe the impact of machine learning model accuracy on user performance, in a domain-general task that is relevant but not directly tied to international safeguards verification. Then, we examine the impact of machine learning error type on user performance, across domain-general and safeguards-relevant tasks. We conclude with recommendations to calibrate human-algorithm systems to the domain-specific task requirements such as prioritization of acceptable machine learning errors and describe the importance of domain-specific testing which we posit may be additionally impacted by expertise.

APPROACH

In prior cognitive studies of international safeguards verification activities [see 6; 7; 8], we have utilized some of Sandia's vast human performance testing capabilities to measure physiological performance (e.g., eye tracking) and field-studies measuring performance in realistic deployment environments. Due to the ongoing global COVID-19 pandemic, we were unable to conduct any in-person human performance testing. We refocused our human performance testing to remote data collection using Amazon's Mechanical Turk, a crowdsourcing platform which we used to track user accuracy and response times in our experimental tasks.

In our research, we asked our participants to decide for each stimulus if a target was present or absent. Some participants were shown stimuli with no simulated machine learning output; others were shown stimuli with the simulated machine learning output. We varied the accuracy of the machine learning output, and varied the machine learning output accordingly between five response types:[1] Hit, Correct Rejection (CR), False Alarm (FA), Miss, and FA+Miss. A rubric of the machine learning response types is in Table 1.

CONDITION NAME	TARGET PRESENT	MODEL OUTPUT CORRECT
Hit	Yes	Yes – Model correctly identifies the target.
Correct Rejection (CR)	No	Yes – No model output is shown to indicate absence of a target.
False Alarm (FA)	No	No – Model incorrectly identifies a distractor as a target.
Miss	Yes	No – Model incorrectly misses a target. No model output is shown to indicate (incorrectly) the absence of a target.
FA+Miss	Yes	No – Model incorrectly identifies a distractor as a target when there is a target in a different location.

Table 1: Simulated machine learning output types.

We conducted two parallel tracks of research. One was domain-general, engaging cognitive processes that are important in international safeguards as well as other domains. The other was domain-specific, using a task that was directly safeguards-relevant.

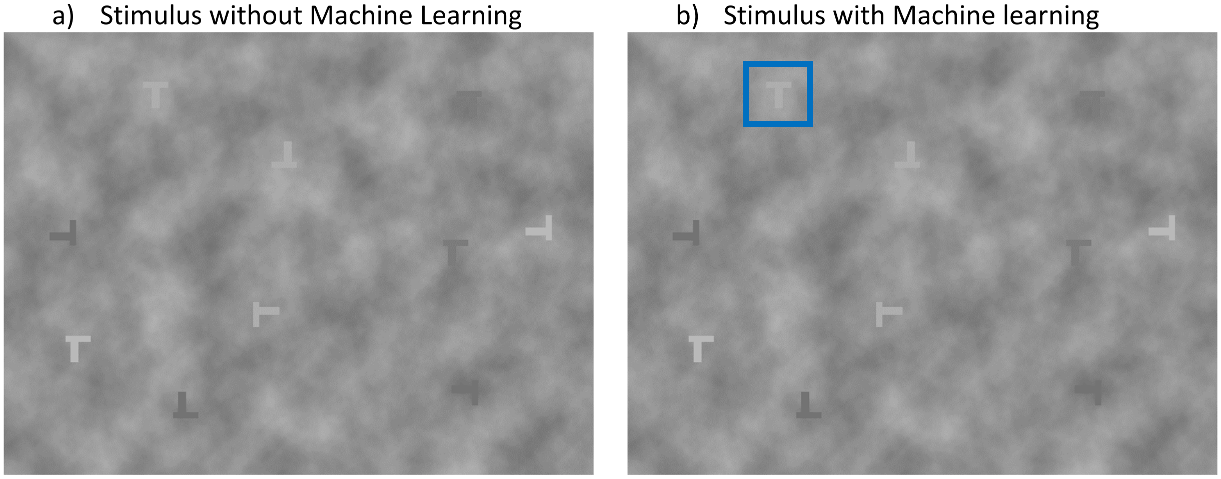
Domain-General Experimentation

First, we wanted to understand the impact of machine learning errors on human performance in a domain-general setting. We selected the "T and L task," which is widely used in the cognitive science literature to study the cognitive processes involved in visual search [9]. In the task, users are presented with a stimulus containing a cloudy background in which 10 potential targets were placed. The target object was the letter "T" with a perfectly aligned crossbar, of which there was no more than one per stimulus. The remainder of the distractor objects were imperfectly aligned cross bars, termed Ls. The Ts

[1] In this work, we use the cognitive science terms for response type rather than the machine learning terms. The response type names in the machine learning community are (respectively): true positive, true negative, false positive, false negative. The false positive + false negative designation is not typically used in machine learning, based on confounding performance metrics between machine learning model types.

and Ls were arranged in four different orientations – with the crossbar pointing up, down, right, or left, in four shades of gray. The participants were tasked with determining if the target – a perfect T – was present in the image. An example T/L stimulus is provided in Figure 1a.

As described above, we wanted to understand the impact of machine learning algorithms on this visual search task. We simulated realistic output from machine learning algorithms, rather than using real machine learning, so that we could experimentally manipulate the accuracy of the output, the types of errors that would be presented to users, and the order in which we presented stimuli to users. A sample of the simulated machine learning output is shown in Figure 1b.



a) Stimulus without Machine Learning

b) Stimulus with Machine learning

Figure 1: a) Example Ts and Ls. T is in the top left quadrant) b) Stimulus with the simulated machine learning output.

Safeguards-Specific Experimentation

We also wanted to understand how these visual search tasks translated to safeguards-relevant tasks. Overhead imagery review, surveillance camera data review, and open source imagery review are safeguards-relevant tasks that are cognitively similar tasks to the T/L visual search. Because overhead imagery visual search has been well-documented within the cognitive science community [see 10], and based on data that we had available, we opted to apply the visual search task to an open source imagery collection task. As with the previous task, participants were asked if a target was present or absent for a given stimulus. In some cases, they were provided experimentally simulated machine learning responses of the same types as the T/L experiment. For our safeguards task, the target was a hyperboloid-shaped cooling tower like those frequently associated with nuclear power plants (also common for coal plants). An example image of a nuclear cooling tower is in Figure 2, with and without machine learning output. This sample image is more immediately obvious than those used in most of our research – it is shown here as a demonstration only. Due to the poor visibility of the blue bounding boxes in most of our images which feature blue or gray skies, we updated the bounding box color to red for the cooling tower experiments.

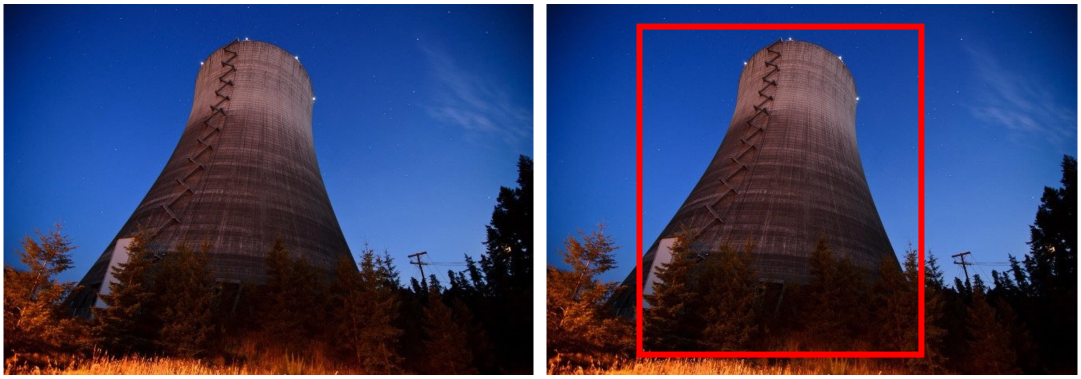


Figure 2: Hyperboloid-shaped cooling tower without (left) and with (right) simulated machine learning responses. Image credit: Sharkhats, via Flickr. Image number 6115042441, 9/4/2011.

BASELINE VISUAL SEARCH PERFORMANCE

To understand the impact of machine learning, we first needed to understand general human performance on our two types of task without any machine learning support. We ran a baseline study for each of the tasks – T/L and cooling towers – to determine our baseline performance rates.

Participant Performance on T/L Task without Machine Learning Assistance

For our T/L baseline task, participants were presented with the T/L stimuli with no simulated machine learning responses. Each participant saw 120 images, where 72 images (60%) contained a target T while the other 48 images (40%) contained only Ls. There were 10 additional "catch trials"[2] in which participants were directly instructed to press either the "target absent" or the "target present" response button. A total of 37 participants provided usable data that passed our quality assurance standards (greater than 80% accuracy for catch trials and greater than 60% accuracy overall).

For the T/L task, we found that participants performed better on stimuli in which the target was absent (mean accuracy 90% correct) compared to those in which the target was present (mean accuracy 72% correct), as shown in Figure 3. That is to say, participants were more likely to miss a target than they were to incorrectly identify a distractor L as being a T.

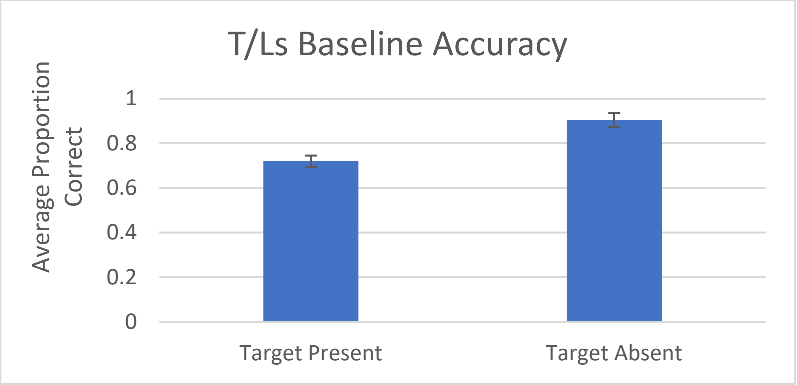


Figure 3: Participant accuracy on target absent and target present images without simulated machine learning support.

[2] A catch trial is a task within the experiment in which the participants are presented with specific response instructions rather than a stimulus. Catch trials are used as a quality control measure to filter out bots and participants who are not paying attention to the task.

Participant Performance on Cooling Tower Task without Machine Learning Assistance

For the cooling tower baseline task, participants viewed 240 images. Similar to the pattern in the T/L task, participants performed better on target absent stimuli (mean accuracy 81% correct) than they did on target present stimuli (73% correct). See Figure 4 for participant performance. The performance for the cooling tower stimuli was lower overall than the performance for the T/L stimuli. This difference is likely because the cooling tower targets could be more difficult to discern than the Ts in the T/L task. There was a great deal of variability in the size of the cooling tower targets and the visual clutter in the images, whereas the T/L stimuli were more tightly controlled and always had targets of the same size and backgrounds with the same degree of visual clutter. In addition, the participants were not experienced in the safeguards domain, so they may have had difficulty with distinguishing cooling towers from other types of towers in some cases.

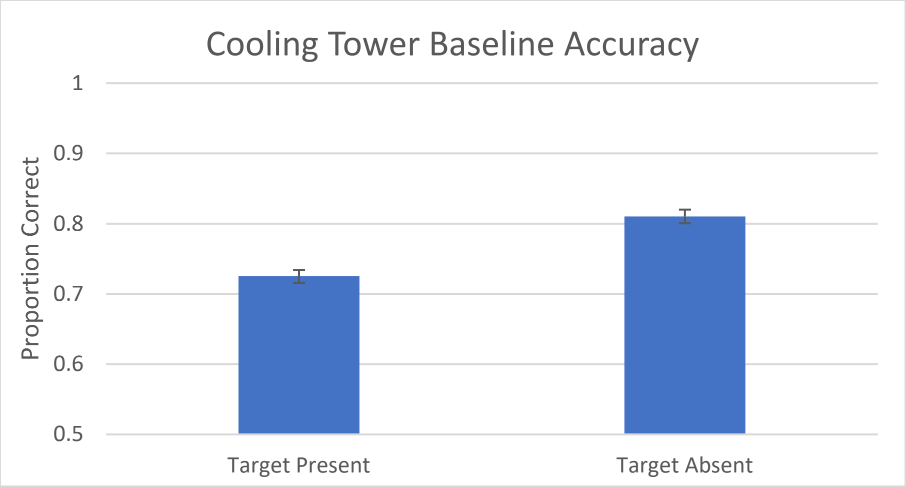


Figure 4: Average participant accuracy on baseline cooling tower stimuli.

EFFECTS OF MODEL ACCURACY AND ERROR TYPE IN THE DOMAIN-GENERAL TASK

After establishing the baseline levels of performance for these tasks without the assistance of a machine learning model output, we conducted experiments in which we provided participants with simulated machine learning outputs to aid them in these visual search tasks. Participants were given the same tasks but were told that the images had been evaluated using a machine learning algorithm that placed boxes around potential targets. The participants were told that the bounding boxes were intended to assist them with locating and identifying targets in the images, but they were warned that the bounding boxes might not always be accurate. We manipulated the accuracy of the model outputs to test the impact of model errors on the participants' performance on these tasks.

Model Accuracy

For the domain-general T/L task, we tested six levels of model accuracy: 50%, 60%, 70%, 80%, 90%, and 95% correct. Participants recruited through Amazon's Mechanical Turk were allowed only to complete one accuracy condition. There were 120 stimuli and 10 catch trials presented in random order. Completion time averaged just under 14 minutes. A total of 36 participants completed each version of the task. After removing participants whose data did not meet our quality assurance standards, we had between 34 and 36 participants in each of the model accuracy conditions.

We found that as the model accuracy improved, our participant accuracy also improved, albeit not at the same rate as the model performance. See Figure 5. Even when the model outputs were correct only 50% of the time, participants had high accuracy for their own performance. This indicates that the participants were not blindly trusting the model but were also evaluating the images to ensure that the model's output was correct.

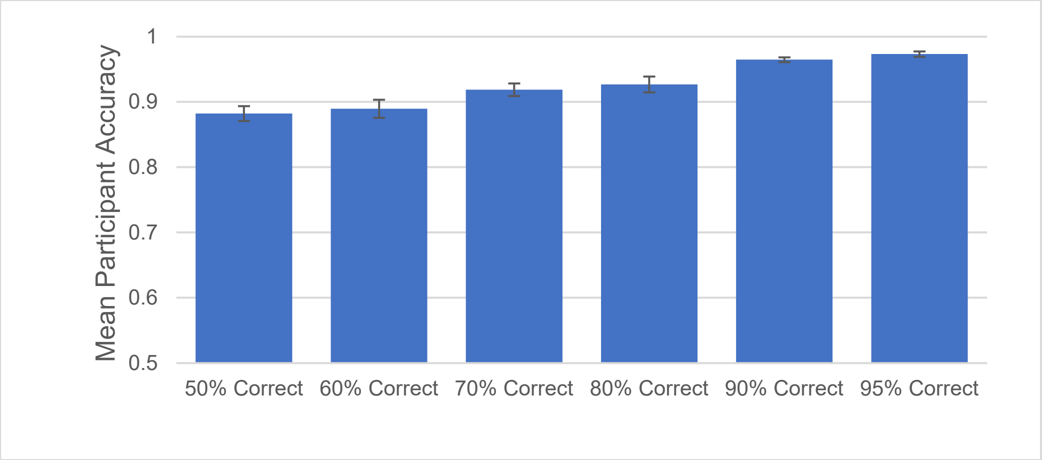


Figure 5: Mean overall accuracy for machine learning model accuracy rates. The error bars indicate the standard error of the mean.

Impact of Different Error Types

While model accuracy and human accuracy are both important, as we note in the title of this paper not all errors are the same. In the machine learning community, the difference in importance of error types is reflected in the performance metrics used to describe most machine learning classification models:

- *Precision* – the proportion of results returned that are relevant (Hits divided by the sum of Hits and FAs).
- *Recall* – the proportion of total relevant population that was correctly returned (Hits divided by the sum of Hits and Misses).

These performance metrics allow developers to hone their models for specifically desired model performance traits, which may not be reflected in a general measure of accuracy. This is especially relevant as machine learning models are adopted in higher consequence domains, such as international safeguards. A model that produces a Miss in international safeguards visual tasks could result in analysts or inspectors missing a safeguards-relevant event or signature, while a FA might be seen as a tolerable nuisance.

Beyond the impact of a model's error rate, we hypothesized that different *types* of model errors would impact human performance in different ways. To assess this, we analyzed target present and target absent trials, separating the cases where the model was correct from the cases where it was incorrect. Figure 6 shows the impact of correct and incorrect model outputs on target present trials for each level of overall model accuracy. When the model outputs were correct (Hits), participants performed much better than they did on the baseline condition, where no bounding boxes were provided. When the model outputs were incorrect (Miss and FA+Miss), the participants performed worse than they did when no bounding boxes were provided. The difference in accuracy between the Hit and Miss trials remained fairly stable, regardless of the overall accuracy of the model outputs. The Miss and FA+Miss conditions were the conditions that led to the poorest human performance across all levels of overall model accuracy. In these conditions, there is a target in the image, but the bounding box is either absent (Miss) or placed incorrectly (FA+Miss). The images used in this study were identical to the images used in the baseline study, but when participants had the expectation of seeing a bounding box around targets, their performance decreased whenever that bounding box was incorrect. These results are especially concerning given the high importance that the international safeguards community places on missed classifications. As mentioned previously, due to the potential significance of a missed event or signature, there is relatively high tolerance for FA results, and extreme aversion to Misses.

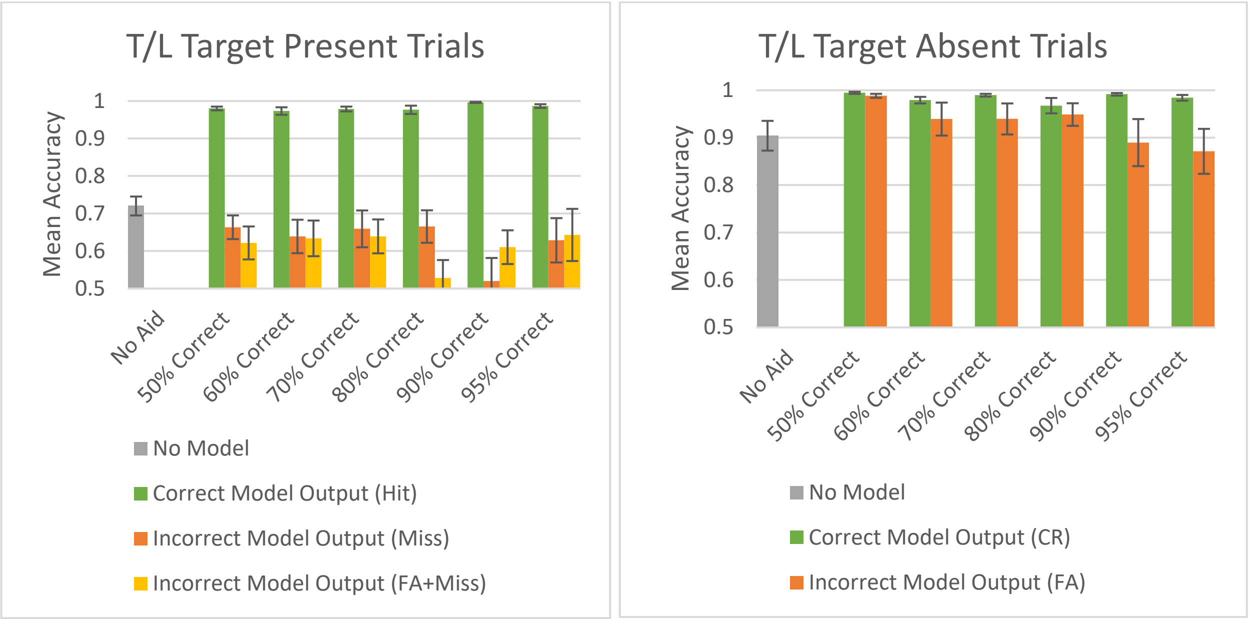


Figure 6: Participant performance across error types and error rates for target present (left) and target absent (right) images.

The right side of Figure 6 shows the impact of correct and incorrect model outputs on target absent trials for each level of overall model accuracy. When the model outputs were correct (CRs), participants once again performed better than they did on the baseline condition, where no bounding boxes were provided. Then the model outputs were incorrect (FAs), the participants still performed better than they had in the No Aid condition, but their accuracy *decreased* as the model became more accurate overall. This indicates that as model accuracy increased to 90% and above, participants were "trusting" the bounding boxes even when they should not have. In the FA trials, the bounding box is placed around an L and there was no target present in the image. For the highly accurate model conditions, it appears that the participants were not carefully evaluating the contents of the bounding box.

At the conclusion of the error rate experiment, we asked participants to guess how accurate the model outputs were. They estimated what percentage of the model's outputs had been correct on a scale that ranged from 20% - 100% (with intervals of 10%). In general, most participants estimated the performance of the model that they were shown relatively well. However, a notable pattern emerged when comparing participants who over- or underestimated the model's accuracy. When the model provided incorrect responses (Miss, FA, or Miss+FA), participants who had overestimated the accuracy of the machine learning model significantly under-performed compared to participants who had correctly estimated or over-estimated the performance of the model.

EFFECTS OF ERROR TYPE IN THE SAFEGUARDS-SPECIFIC TASK

While participants' performance on the domain-general task is telling, we recognized the need to understand performance within the specific context of a safeguards-relevant task. For the cooling tower stimuli, a group of 233 participants was tested on a set of images for which the simulated machine learning outputs were correct 80% of the time. This research is described in detail in [11].

Although we did not manipulate the overall error rate for the cooling tower experiment, it is instructive to compare the participants' overall accuracy when performing this task with and without machine learning assistance. Recall that when participants performed the task without machine learning outputs, they averaged 81% correct for the target absent images and 73% correct for the target present images. When the bounding boxes were added, they averaged 73% correct for target absent images and 84% for target present images. Similarly to the T/L task, we see a significant improvement in performance with the addition of simulated machine learning output when the images contain a target. While both the T/L and cooling tower tasks had increased performance on target present images when the simulated machine learning output was present, the cooling tower task had a considerable drop in performance on target absent images with the addition of the simulated machine learning output that was inconsistent with performance on the T/L task.

Once again, our primary goal was to investigate the impact of different types of errors on participant performance. These results are shown in Figure 7. The participants were near ceiling for Hits, just as they were in the T/L task. For the Miss and Miss+FA trials, participants had substantially worse performance than they did in the baseline condition, when they did not see any bounding boxes at all. These performance decreases represent a significant risk for international safeguards, where "Misses" by either a human or a model could result in the non-detection of safeguards-significant activities.

For the target absent images with correct model outputs (i.e., no bounding box), participants performed equally well as in the baseline condition. When there was a false alarm, the participants had extremely poor performance, averaging only 36% correct. The participants' performance was lower in both of these conditions than it was in the equivalent conditions in the T/L task. This may reflect the participants' uncertainty about the stimuli, which were not as clear-cut as the T/L stimuli. Given this additional variability and subjectivity, the participants tended to accept the recommendations of the machine learning model. Importantly, their performance on the FA images was much worse than in the baseline condition, demonstrating that the machine learning outputs had a substantial influence on the participants' decisions.

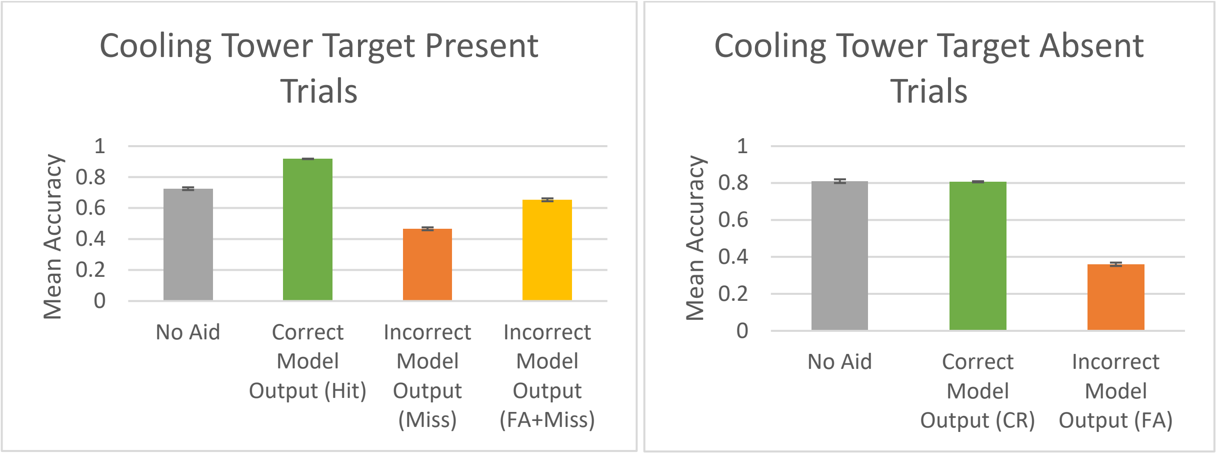


Figure 7: Average participant accuracy by machine learning response type for target present (left) and target absent (right) stimuli.

DISCUSSION

In this research, we found that participant performance differs in response to different model accuracies, and in response to different types of model errors. We observed that models with very high performance can result in over-compliance with the model – users accept model recommendations that are incorrect, for images they would have been able to categorize correctly without machine learning aid. While we did not test very high performance models using the cooling tower dataset, we hypothesize that the impact of the FA and Miss recommendations might be further exacerbated with high accuracy models in a similar fashion to what we observed with the T/L task.

For the T/L task, participants performed very well when rejecting FAs produced by the model, while FAs represented the worst performance category for participants seeing the cooling tower stimuli. We think this may be due to participant inexperience with cooling towers and are conducting additional research to determine the impact of expertise in this task.

We note that the balance between model FAs and model Misses can usually be tuned by the model developers. For organizations with higher tolerances of FAs, such as international safeguards, models might be tuned to minimize levels of Misses but allow for higher FA rates. However, any such tuning would need to be done within the context of the human-algorithm system: even if safeguards practitioners are more tolerant of FAs, that also represented the lowest performance category for cooling tower participants. FAs from the model were causing FAs from the model users (as opposed to the T/L task where participants more accurately dismissed the FAs). The needs of the organization, and the interaction of the users with the model, should influence how machine learning models are implemented.

We have described multiple times the differences in participant performance in general – and specific impacts of different types of machine learning errors – that differ between the domain-general and the safeguards-relevant stimuli. If these performance differences were based on expertise, we might consider how to differ implementation of ML for novice and experienced users. Regardless of the findings on the role of expertise, we have observed a broader need for additional domain testing when implementing models in new areas.

ACKNOWLEDGEMENTS

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and 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. SAND2021-9466 C.

This work was funded by Sandia's Laboratory-Directed Research & Development program, under the Computational and Information Sciences investment area. Thanks to Timothy Shead, Michael R. Smith, and Kurt Larson for their support to this work.

REFERENCES

[1] Cui, Y., Ren, Y., Smartt, H., Thomas, M., Smith, M.R., Yoo, S., Stern, W., Lin, Y., Gastelum, Z., "Using deep machine learning to conduct object-based identification and motion detection on safeguards video surveillance." IAEA Symposium on International Safeguards, 5-8 November 2018, Vienna, AUSTRIA.

[2] Warner, T., Keskinen, A., Rutkowski, J., Duckworth, S., "Exploitation of high-frequency acquisition of imagery from satellite constellations within a semi-automated change detection framework for IAEA safeguards purposes." IAEA Symposium on International Safeguards, 5-8 November 2018, Vienna, AUSTRIA.

[3] Feldman, Y., Arno, M., Carrano, C., Ng, B., Chen, B., "Toward a multimodal-deep learning retrieval system for monitoring nuclear Proliferation Activities." Journal of Nuclear Materials Management, Vol. 46, No. 3 (2018).

[4] Gastelum, Z., Shead, T., "Inferring the operational status of nuclear facilities with convolutional neural networks to support international safeguards verification." Journal of Nuclear Materials Management, Vol. 46, No. 3 (2018).

[5] Gastelum, Z., Matzen, L.E., Stites, M.C., Jones, A., Trumbo, M., Howell, B.C., Higgins, M., "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.

[6] Gastelum, Z.N., Matzen, L.E., Stites, M.C., Smartt, H.A., "Human Performance Testing on Observation Capture Methods for International Nuclear Safeguards Inspections: Transferring Knowledge from the Field to Headquarters and Back." Proceedings of the Institute of Nuclear Materials Management Annual Meeting, July 2019.

[7] Gastelum, Z.N., Stites, M.C., Matzen, L.E., "The role of maps in site knowledge and wayfinding: A human performance evaluation for international nuclear safeguards inspections." European safeguards Research & Development Association Symposium, May 2019.

[8] Gastelum, Z.N., Matzen, L.E., Stites, M.C., Smartt, H.A., "Cognitive science evaluation of safeguards inspector list comparison activities using human performance testing." Proceedings of the Institute of Nuclear Materials Management Annual Meeting, July 2018.

[9] Wolfe, J. M., "Guided search 2.0 a revised model of visual search." Psychonomic bulletin & review, Vol. 1, No. 2 (1994).

[10] McNamara, Laura A., and Laura M. Klein. "Context-sensitive design and human interaction principles for usable, useful, and adoptable radars." In Radar Sensor Technology, vol. 9829, p. 982906. International Society for Optics and Photonics, 2016.

[11] Divis, K.M., Howell, B.C., Matzen, L.E., Stites, M.C., Gastelum, Z.N., "The cognitive effects of machine learning aid in domain-specific and domain-general tasks." Hawaii International Conference on System Sciences (submitted).