SAND2022-11370C

COGNITIVE IMPACTS OF COMPUTER VISION-BASED DECISION SUPPORT FOR INTERNATIONAL NUCLEAR SAFEGUARDS-RELEVANT VISUAL ANALYSIS TASKS

Z.N. GASTELUM
Sandia National Laboratories
Albuquerque, United States of America
Email: zgastel@sandia.gov

L.E. MATZEN
Sandia National Laboratories
Albuquerque, United States of America

K.M. DIVIS
Sandia National Laboratories
Albuquerque, United States of America

B. HOWELL
Sandia National Laboratories
Albuquerque, United States of America

Abstract

Computer vision-based decision support tools are being developed and evaluated to augment international safeguards practitioners for multiple types of visual analysis activities. These tools have significant potential to be integrated into human-machine teaming for efficient and effective processing of large quantities of safeguards data. Though some computer vision decision support tools may achieve high levels of performance, the performance of the human-machine system requires additional evaluation. Our team has conducted extensive human performance testing to understand the cognitive impacts of using computer vision decision support tools. Our safeguards-informed research has yielded three main conclusions regarding the use of decision support tools for visual tasks. First, even imperfect models can provide productive decision support for users. Second, there exists a balance between user speed and accuracy, resulting in users of high-performing models sometimes missing subtle but important errors. Third, user expertise impacts how the user responds to model errors. As a result, multiple decision support models may be needed to support users at different levels of expertise. The paper expands on the research behind these conclusions and provides directly actionable recommendations for safeguards implementation of computer vision-based decision support tools.

1. INTRODUCTION

Many international nuclear safeguards monitoring tasks involve visual inspection and assessments – either of physical objects in the field, or of visual data on screens from safeguards equipment, open-source information, and other data. Recent increases in performance of computer vision models that help humans to interpret visual data have led to evaluations of how these capabilities could be applied to safeguards tasks. Some examples of research in this area include:

- The collection and interpretation of open-source images to predict the operational status of nuclear facilities [1].
- The retrieval of relevant data using multi-modal search strings such as images, video, or audio [2].
- The monitoring of patterns of life and change detection via overhead imagery analysis [3].
- The identification and tracking of safeguards-relevant containers, and identification of deviations from patterns of life, from safeguards surveillance camera data [4].
- The localization of a user within a known facility based on surrounding contextual images [5].

These research efforts represent the potential for significant impact on inspector and analyst workload, supporting safeguards practitioners in sorting and prioritizing data for their expert reviews. It is critical to consider these research tools and capabilities as part of a broader system that includes users. Our team has developed and

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.

implemented a series of human performance experiments to assess how the outputs of these models impact users' cognition and decision-making during safeguards-relevant visual search tasks. We have specifically focused on the cognitive impacts to users when the models provide incorrect responses. Our team has assessed many aspects of decision support model errors, including visualizations of model outputs [6], error types [7], confidence levels [8], differences between domain-specific and domain-general tasks [9], expert and novice users [10], quality of machine learning explanations [11], and directions to users [11].

Here, we will examine three key findings that can inform the development and implementation of computer vision models to support international safeguards visual tasks. First, even models with relatively poor performance can provide productive decision support for users, but improvements in model performance generally produce larger benefits to users. Second, there are risks when users have models with very high performance. No model has perfect accuracy and users of very high-performing models are less likely to notice the instances where the model makes an error. Finally, user expertise impacts how the user responds to model errors. Some tasks may require alternative outputs or approaches to decision support for users with various levels of experience.

2. EXPERIMENTAL BACKGROUND/METHODOLGY

We will describe results of two experiments using visual search tasks in which users scan an image in search of a defined target. The visual search tasks were selected to be representative of multiple safeguards-relevant tasks such as open-source information analysis, satellite imagery analysis, or situational awareness in the field. The first visual search task – the T and L task – is used extensively within the corpus of research on human visual search (for example, [12] and [13]). It is a domain-general task that can be performed without any specific expertise, yet it relies on the same fundamental cognitive processes that are required for more specialized visual search tasks, such as those that are important in the safeguards domain.

In the T and L task, users search for a letter T that has a perfectly centered crossbar. The images contain distractor stimuli, referred to as "L"s, that have offset crossbars. In our implementation of the T and L task, the target and distractor stimuli were displayed in shades of grey on a cloudy grey background. Experiment participants were tasked with viewing these images and determining whether there was a target T in each one. In some conditions they had no decision support aid and had to find the targets through visual search. In other conditions, they saw decision support aids that purported to identify targets in the images. The appearance, placement, and accuracy of the decision support indicators was experimentally manipulated by our team according to the research design, see Fig. 1.

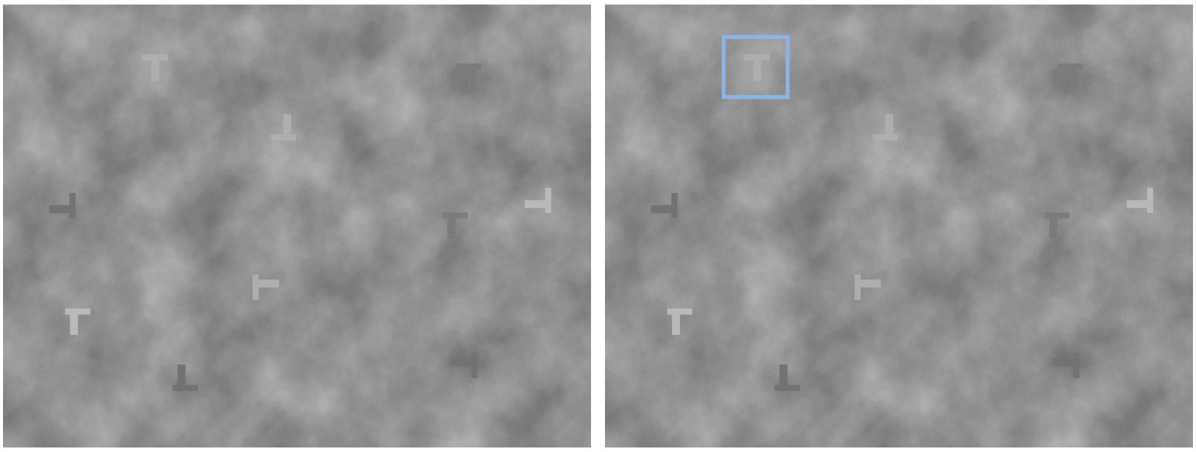

FIG. 1. Example T and L stimuli without decision support (left) and with simulated decision support (right)

We also used a more directly safeguards-focused search task, in which the users were tasked with viewing images and determining whether they contained hyperboloid-shaped cooling towers, see Fig. 2. In this case, all the images were open-source photographs of real structures. The targets were cooling towers and the distractors were other types of towers, such as windmills, smokestacks, and steam plumes for which the source was not visible. As with the T and L task, the location and accuracy of the decision support indicators was manipulated according to the experimental design. The ways in which the accuracy of the decision support tool were manipulated are described in Table 1.

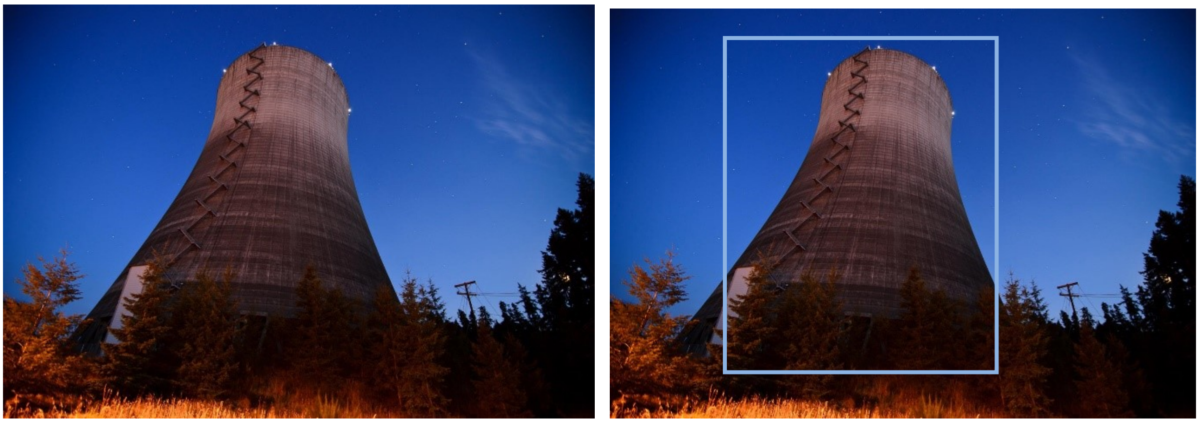

FIG. 2. Hyperboloid-shaped cooling tower without (left) and with (right) simulated machine learning responses. Image credit: Sharkhats, via Flickr. Image number 6115042441, 9/4/2011.

TABLE 1. DECISION SUPPORT MODEL OUTPUT MANIPULATIONS

Condition Name	Target Presence	Model Output
True Positive	Present	Model correctly identifies the target.
True Negative	Absent	No model output is shown, correctly indicating absence of a target.
False Positive	Absent	Model incorrectly identifies a distractor as a target.
False Negative	Present	Model misses the target, and thus no model output is shown.
False Positive + False Negative	Present	Model incorrectly identifies a distractor as a target when there is a target in a different location.

In both the T and L task and the cooling tower task, participants viewed the electronic images one at a time and selected a button to indicate whether there was a target in the image. They were tasked with correctly identifying the images that contained targets, regardless of the accuracy of the decision support output. The participants' accuracy and response time was recorded for each image. Most participants completed the task on Amazon Mechanical Turk. The online tasks included attention checks to ensure that the participants produced high-quality data. In addition to the online participants, we recruited a small number of nuclear fuel cycle experts from Sandia National Laboratories for a small case study comparing expert and non-expert performance (described in Section 5).

3. IMPERFECT MODELS PROVIDE DECISION SUPPORT

One of our key findings was that even relatively poorly performing computer vision models could provide positive impact on user performance. This finding has important implications for the development of decision support tools. On the T and L task, we found that users with no decision support had an average overall accuracy of 79%. When we added the simulated decision support outputs, we saw significant increase in user performance as shown in Fig. 3. We found that as the model performance increased, so did overall user performance. There was a benefit to performance even in cases where the decision support outputs were only 50% correct.

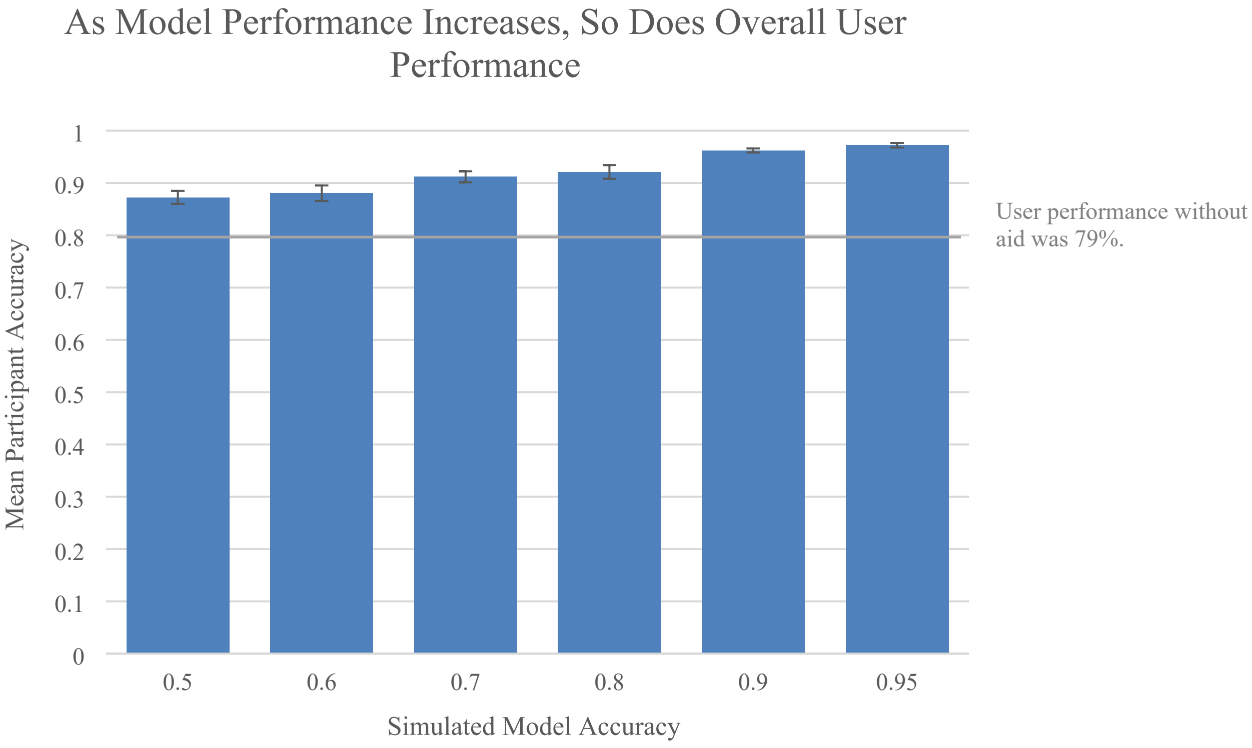

FIG. 3. User performance on the T and L visual search task over multiple simulated model accuracies

Based on these findings, we recommend that developers of decision support tools for international safeguards visual tasks focus on achieving reasonable performance metrics rather than strive for incremental increases once a moderate performance level is reached. Despite increasing user performance returns as model performance increases, the finding that even modestly performing models support user performance on this visual search task indicates that models will be able to positively impact user performance at earlier stages than we previously thought.

4. USERS OF HIGH PERFORMING MODELS SOMETIMES MISS SUBTLE BUT IMPORTANT ERRORS

Another finding from our experiments provides an important counterpoint to the finding discussed above. Although user performance generally improved as the accuracy of the decision support outputs increased, we found that participants were less likely to notice erroneous outputs. When the model outputs were highly accurate, participants were more likely to accept the model's decision and less likely to perform a careful visual search of the images to confirm that decision. This pattern was reflected in the participants' average response times, which were faster for more accurate models. It was also reflected by the participants' accuracy on trials where the model output was incorrect, as shown in Fig. 4. As the model's overall accuracy increased, the participants were less likely to identify instances where it had produced an output that was a false negative or a false positive.

This finding was replicated across two experiments using the T and L paradigm. In the first experiment, the decision support outputs had a mix of different types of errors, as described in Table 1. In the second experiment, all the model errors were of one type: either all false positives or all false negatives. In both cases, the participants' ability to recognize images where the decision support output was incorrect decreased as the overall accuracy of the decision support model increased. This decrease in user performance could be attributed to users overly trusting the models or becoming compliant with model output.

In the T and L task, where the targets can be somewhat difficult to find, participants tend to have lower accuracy for images that contain targets than for those that do not. In a baseline experiment with no decision support outputs, our participants responded correctly to 90% of the target absent items and only 72% of the target present items, on average. Similarly, when decision support outputs were provided, we observed that the participants performed worse on the false negative stimuli (which had a target that was missed by the decision support output) than the false positive stimuli (which had a non-target incorrectly identified as a target). When shown a false positive produced by the decision support tool, users could easily evaluate the item inside of the bounding box and come to their own conclusion about whether it was a target. In contrast, identifying a false negative required the users to conduct their own visual search of the image to see if *any* of the items might be a target. This was a much more time-consuming and effortful process, which is why the users were less likely to identify the false negatives. As the model outputs became more accurate overall, meaning that there were fewer false negatives for each participant, the participants were even less likely to search the images for missed targets. This made their performance on the false negatives decline even more.

Based on these findings, we recommend that computer vision model developers tune model thresholds to be more tolerant of false positives than false negatives. This essentially means lowering the threshold for a model to classify an item as a target, with the trade-off of having more false positives returned to the user. Because of the importance within the safeguards domain of limiting false negatives (missing a relevant target), we propose that the value of reducing potential false negatives outweighs the risks of including more false positives. False negatives pose a greater challenge to human users, who are less likely to notice them, even in very straightforward tasks like the T and L task. Furthermore, we expect that the workflows surrounding positive identifications (such as passing information to a more senior analyst or inspector) will counteract any potential performance decrease posed by the presence of additional false alarms from a model.

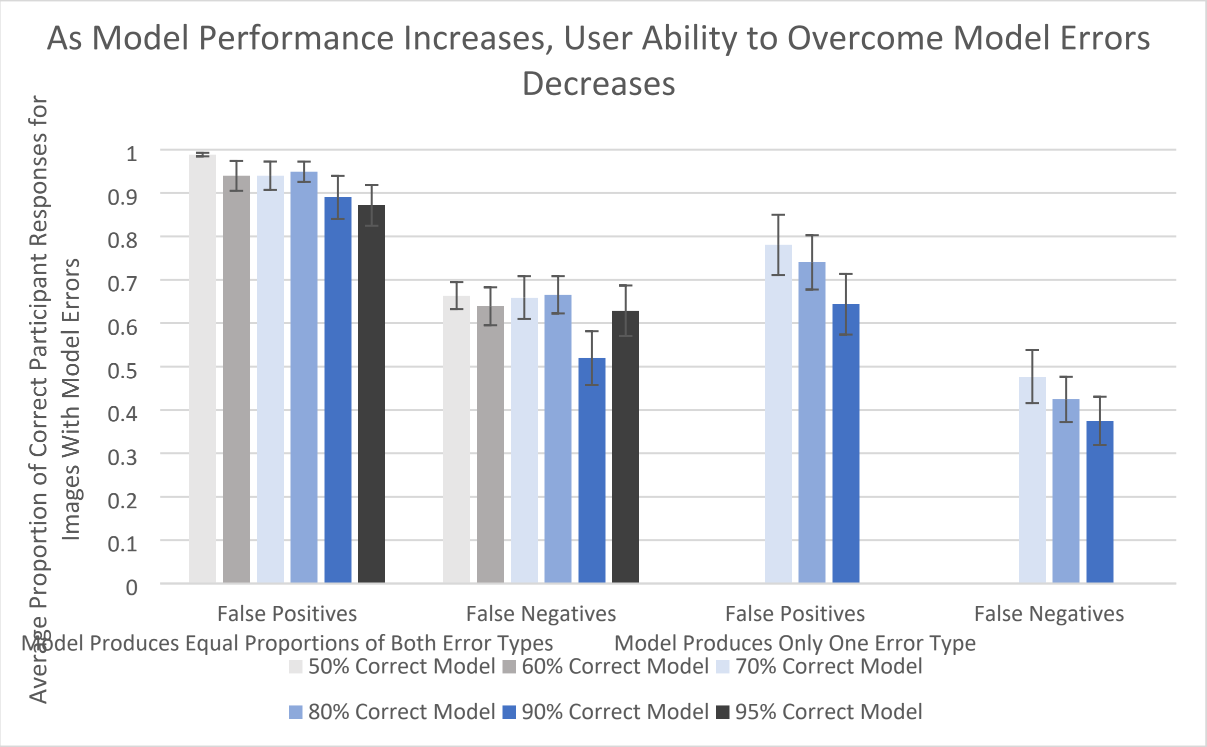

FIG. 4. User performance on the T and L visual search task when the model contained equal proportions of false positive and false negative results (left), and when the model contained only a single type of error (right). In both experiments, as model performance increased the user performance on error trials decreased.

5. USER EXPERTISE IMPACTS RESPONSE TO MODEL ERRORS

Finally, we wanted to generalize the results from the T and L study to a safeguards-relevant stimulus. For this test, we used the hyperboloid cooling tower data. Initially, we observed very different levels of performance between Amazon Turk participants completing the T and L task and the cooling tower task. We hypothesized that the difference was due to expertise because the hyperboloid cooling towers may be unfamiliar to some users. We recruited a small number of nuclear fuel cycle experts from Sandia National Laboratories to serve as an expert community and compared their performance to the Amazon Mechanical Turk users who we called novices, assuming few of them had extensive experience working with the nuclear fuel cycle. There were two versions of this task: one where no decision support outputs were provided, and one where we provided decision support outputs that were accurate 80% of the time. The results of this comparison for domain experts and novices are shown in Fig. 5.

In comparing the novice and expert performance on the cooling tower task, we observed that novice users benefitted the most from the true positive model outputs in terms of overall accuracy gains. However, the novices were also vulnerable to the decision support model's false positive and false negative responses. Novices were more likely to miscategorize an image when given an erroneous decision support output than when they were given no decision support output at all.

In contrast, the experts completing the same task were not negatively impacted by incorrect decision support outputs. Their performance for images with erroneous decision support aids was no worse than it was for images without decision support aids. However, like the novices, the domain experts benefited from seeing correct decision support outputs. We observed increased expert accuracy for both types of correct responses, true positive and true negative, even though the true negative model response does not display a visual output (recall from Table 1 that the true negative is depicted with no bounding box, indicating no target is present).

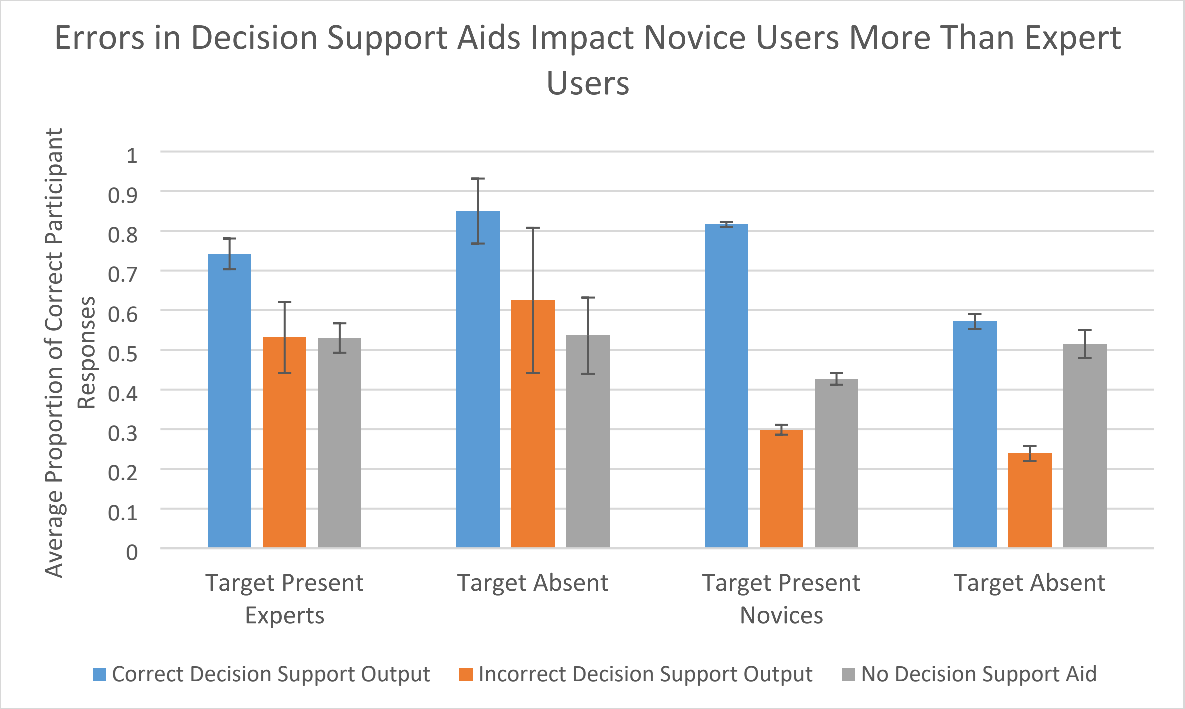

FIG. 5. User performance on the cooling tower visual search task, in which experts and novices performed identical tasks. While both groups were positively impacted by correct decision support output, novices were more negatively impacted by incorrect decision support than experts.

Based on these findings, we recommend that different models be developed and tailored for different types of users. We anticipate that safeguards analysts and inspectors–even those who are new to their positions–are experts in their field. Despite their expertise, we expect that at times safeguards practitioners will be making assessments on unfamiliar technologies or equipment. In these scenarios, safeguards experts may make decisions more like novices than experts. For novices, we observed that both false positives and false negatives had significant negative impact on user performance. However, even with both types of error negatively impacting performance, we make the case that false positives still be preferred over false negatives. While we do not advocate for models that provide so many false positives that users start ignoring model responses, analytical workflows that are informally and formally codified in safeguards will likely result in a real user requesting additional input from a peer, a technology expert, or other staff who can assist in dismissing the false positive (which also provides feedback to the original user, and is a powerful learning mechanism that will help them develop expertise). False negatives were similarly detrimental to user performance but may be harder to correct for operationally.

As we observed in the T and L task, our expert users were better able to respond to false positives than to false negatives. But in this task, we also observed that false negatives did not have a significant positive or negative impact on expert user performance, as accuracy stayed approximately the same as the no-aid condition.

6. DISCUSSION AND CONCLUSIONS

To summarize our findings, even moderately performing computer vision decision support can benefit user performance on a visual search task. Model developers should strive for good, but not necessarily near-perfect, model performance that prioritizes expedited model deployment and a good user interface. These factors are likely to provide a larger benefit to human-machine performance than incremental improvements to model performance.

Furthermore, even when model performance is very high, users should remain vigilant to their task. Even very good models have occasional errors, and users who become complacent and overly compliant with model recommendations are highly likely to overlook those errors.

It is also important to be mindful of the impact of different types of model errors. Our studies consistently found that people were less likely to notice a model error if it was a false negative than if it was a false positive. In these tasks, where a bounding box was used to indicate the presence and location of targets, false positives were cases where a bounding box was placed on an object that was not actually a target. In the domain-general T and L task, the participants simply had to look at the item inside of the bounding box and determine whether it was a T or an L, which was trivially easy (if they were paying attention and not just uncritically accepting the model's decisions). In the domain-specific task, identifying whether the tower inside of the bounding box was a hyperboloid cooling tower was similarly easy for domain experts. However, these false positives proved to be harder for novices, who may have been unsure of the correct answer and deferred to the decision support recommendation, incorrectly accepting the false positive results.

False negatives were quite difficult for people to identify in both tasks. In this case, there was a target in the image, but it was not marked with a bounding box. Rather than simply evaluating what was inside of the bounding box to determine whether the model was correct, participants would have to conduct their own visual search of the image to find any unmarked targets. This process requires considerable effort and may not succeed if the target is difficult to find. As a result, the participants in our studies failed to find many (often more than half) of the targets that not correctly marked by the decision support output.

Given these findings, it may be beneficial to tune models so that they are less likely to produce false negatives. This kind of tuning is likely to result in a higher rate of false positives, but these are easier for users to recognize and correct. While systems that produce too many false positives often suffer from user disengagement, visual search or target recognition tasks can benefit from a sensitive detection threshold that produces few false negatives.

Finally, for tasks in which safeguards practitioners have little experience, higher levels of model performance may be necessary to support effective decision making. We anticipate deployment of decision support models within analytical workflows in which novice analysts or expert analysts working tangential to their primary expertise would consult with others when presented with a potential positive (true or false) output. Within that context, we recommend that false positives still be favored over false negatives to avoid potential misses of relevant or valuable information. This is true even for models that are intended for novices for whom both types of errors were detrimental to performance.

Overall, our results indicate that computer vision decision support tools have considerable potential to enhance user performance both generally and in the international safeguards domain. In practice, these tools need to be thoughtfully implemented to consider deployment priorities that are likely domain-specific (such as the high consequences of false negatives in the international safeguards domain), user experience (experts or novices), and anticipated target prevalence (frequent targets versus rare targets), which will impact performance metrics. Finally, some highly specialized tasks might only be suitable for human experts. The best-performing systems will result from a combination of computer vision models implemented at tasks for which they excel with human users performing tasks at which they excel.

ACKNOWLEDGEMENTS

This research was funded by Sandia National Laboratories' Laboratory-Directed Research and Development (LDRD) program, Computing and Information Sciences investment area. 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. SAND-2022-

REFERENCES

[1] Z. N. Gastelum and T. M. Shead, "Inferring the Operational Status of Nuclear Facilities with Convolutional Neural Networks to Support International Safeguards Verification," *Journal of Nuclear Materials Management,* vol. XVLI, no. 3, pp. 37-48, 2018.

[2] Y. Feldman, M. Arno, C. Carrano, B. Ng and B. Chen, "Toward a Multimodal-Deep Learning Retrieval System for Monitoring Nuclear Proliferation Activities," *Journal of Nuclear Material Management,* vol. XLVI, no. 3, pp. 68-80, 2018.

[3] J. Rutkowski, M. J. Canty and A. A. Nielsen, "Site Monitoring with Sentinel-1 Dual Polarization SAR Imagery Using Google Earth Engine," *Journal of Nuclear Materials Management,* vol. XLVI, no. 3, pp. 48-59, 2018.

[4] Y. Cui, Z. N. Gastelum, R. Ren, M. R. Smith, Y. Lin, M. A. Thomas, S. Yoo and W. Stern, "Using deep machine learning to conduct object-based identification and motion detection on safeguards video surveillance," in *Proceedings of the IAEA Symposium on International Safeguards: Building Future Safeguards Capabilities*, Vienna, 2018.

[5] E. Wolfart, C. Sanchez-Belenguer and V. Sequeira, "Deep learning for nuclear safeguards," in *Proceedings of the INMM & ESARDA Joint Virtual Annual Meeting*, 2021.

[6] L. E. Matzen, M. C. Stites, B. C. Howell and Z. N. Gastelum, "Different Visualizations of Machine Learning Outputs Influence the Speed and Accuracy of User Evaluations," in *IEEE InfoVis x Vision Science Workshop*, Virtual, 2021.

[7] Z. N. Gastelum, L. E. Matzen, K. Divis, M. C. Stites and B. C. Howell, "Not All Errors are Created Equal: Examining Human-Algorithm System Performance for International Safeguards-Informed Visual Search Tasks," in *Proceedings of the Institute of Nuclear Materials Management & European Safeguards Research and Development Association Joint Virtual Meeting*, Virtual, 2021.

[8] A. P. Jones, M. C. Trumbo, L. E. Matzen, M. C. Stites, B. C. Howell, K. M. Divis and Z. N. Gastelum, "Evaluating the Impact of Algoirthm Confidence Ratings on Human Decision Making in Visual Search," in *Human-Computer Interaction International 2021*, Virtual, 2021.

[9] K. Divis, B. Howell, L. Matzen, M. Stites and Z. Gastelum, "The Cognitive Effects of Machine Learning Aid in Domain-Specific and Domain-General Tasks," in *Hawaii International Conference on System Science*, 2022.

[10] Z. N. Gastelum, B. C. Howell, K. Divis and L. E. Matzen, "When in Doubt, Trust AI? User Performance and Task Difficulty in a Safeguards-Relevant Visual Identification Activity," in *AI for Atoms*, Virtual, 2021.

[11] Z. N. Gastelum, L. E. Matzen, M. C. Stites, K. Divis and B. C. Howell, "Assessing Cognitive Impacts of Errors from Machine Learning and Deep Learning Models: Final Report," 2021.

[12] J. M. Wolfe, "VIsual Search," in *Attention*, University College London Press, 1998.

[13] J. M. Wolfe, K. R. Cave and S. L. Franzel, "Guided Search: An Alternative to the Feature Integration Model for Visual Search," *Journal of Experimental Psychology: Human Perception adn Performance,* vol. 15, pp. 419-433, 1989.