Project 3 Example: Human-AI Collaboration Tester (HAICT) Exp. 7

ClinicalTrials.gov ID NCT05272189

Document Date: December 3, 2025

Introduction

Artificially intelligent (AI) systems are increasingly common and competent collaborators for humans performing a variety of everyday tasks. There are AI systems designed to help us park our cars, translate our voices into text messages, and filter out our spam email. AI collaborators have also been integrated into the workflows of expert humans performing socially important, and even potentially dangerous, jobs — e.g., assisting air traffic controllers in preventing midair collisions. Interestingly, some AI systems are designed to act as second observers to assist humans in making important *perceptual decisions:* Airport baggage screeners use AI support to decide whether there is a weapon in a bag; radiologists use AI support to decide whether a suspicious region on an X-ray is cancer; and law enforcement use AI support to identify suspects in surveillance video. What makes these examples different from AI sorting our email, or calculating flight paths, is that they all involve a human and a machine making a decision about the same perceptual stimulus (e.g., "Is that object in this suitcase a weapon or a dumbbell?").

If humans and AI are working together, even the most talented AI is only useful if its human collaborator can and will take advantage of its information. If the human does not trust the AI, she may ignore its advice (Beck et al., 2007). There are many different possible rules for combining the same human and AI information (e.g. Does the AI offer its information before, during, or after the human's initial decision?). These rules can produce different outcomes. We frame this situation in terms of signal detection theory (Hautus et al., 2021; Macmillan & Creelman, 2005). In that context, changing interaction rules can shift d', criterion, or both[1]. Moreover, the rules interact

[1] Note: In the psychophysical literature, d' is often called "sensitivity". However, in the medical literature, "sensitivity" refers to the true positive (TP) or hit rate. Accordingly, we will try to avoid using the term, "sensitivity" at all.

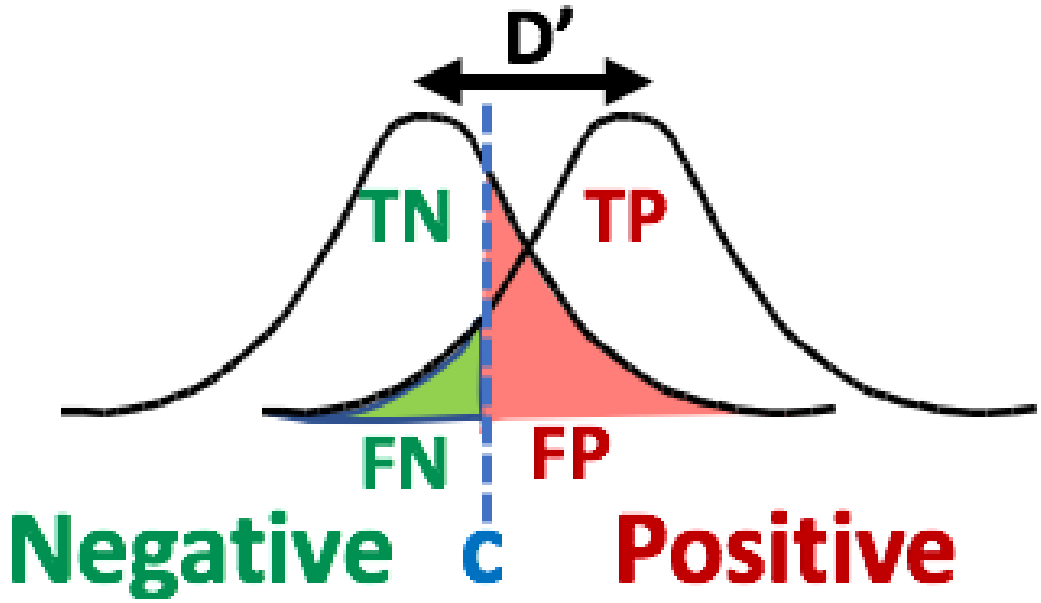
strongly with task variables (e.g. How prevalent are positive cases?). The interaction of task and rule can change the human user's attitude / trust in the AI (Hoff & Bashir, 2015). This, in turn, influences the combined results for human plus AI.

AI support for perceptual decision making is wide-spread. We will focus our discussion on medical images (Greenes, 2014) recognizing that the issues are similar in other domains from conversing with your car (Strayer et al., 2016) to automating airport screening [Hättenschwiler, 2017 #13395]. In medical image interpretation, much of this work falls under the heading of Computer-Aided Detection (CADe) and Computer-Aided Diagnosis (CADx) (Li & Nishikawa, 2015).

If experts were perfect at detecting and identifying clinically relevant findings in images, there would be no interest in CAD, but experts evaluating medical images make more mistakes that we would like, and CAD helps less than we would hope. Precise error rates in actual clinical practice are hard to establish though 20-30% appears to be a reasonable estimate in many radiologic domains (reviewed in Berlin, 2007). Some of these errors are, in a sense, desirable. For instance, a suspicious spot on a mammogram should be referred for further testing, even if it subsequently turns out to be a false positive. Of course, if AI could reduce these 'desirable' false positive errors, that would save unnecessary testing, expense, and worry. False negative / miss errors can have a higher cost in cases where early detection improves outcomes (Monticciolo et al., 2017).

Kundel and colleagues (1978) provided a useful 3-part taxonomy for false negative ("miss") errors, dividing them into search, recognition, and decision errors, based on eye movements recordings. Search errors are those where the eyes never fixate on the target at all. In recognition errors, the eyes briefly (<500 msec) fixate the target but then move on, as if the observer failed to recognize that something interesting had been fixated. Decision errors are those where the observer (and the observer's eyes) scrutinize the relevant stimulus location but fail to identify the target.

Figure 1: Two-alternative forced-choice decisions as signal detection problems.



In their simplest form, decision errors can arise when the observer is faced with a two-alternative forced-choice decision (e.g. Is this cancer or not?). In the terms of signal detection theory, as cartooned in Figure 1, truly negative and positive cases can be thought to produce overlapping distributions of internal response values within the observer. To make a decision, the observer must establish some 'criterion', 'c'; declaring all values above criterion to be positive and those below to be negative. The separation between the positive and negative distributions determines the difficulty of the decision and can be quantified by the parameter, D' (though the situation is

more complex and other metrics are more useful if, for example, the variance of the two distributions is not the same). When the distributions overlap, the observer *must* make errors: False positive (FP) or False alarm errors, shaded red, and False Negative (FN) or miss errors, shaded green. Shifting criterion can change the mix of errors but cannot eliminate them. A manipulation that increases D' reduces errors.

Two observers will typically perform better than either alone, especially if the 'noise' that degrades their performance is not identical. The relationship of two observers in SDT terms has been formalized by Sebastian and Geisler (2018). A CAD AI that offers its own assessment of the stimulus can be thought of as a second observer. The focus of our interest is in the interaction of that second observer with a human who is making the ultimate decision.

Search and recognition errors can arise when the perceptual task involves spatial uncertainty. If the task involves a search for possible targets (e.g. signs of breast cancer), it is possible to fail because you never looked at the target (search errors) or because, when you looked at the target, its importance was not registered (recognition errors). The use of AI to reduce these errors is important but not considered in this experiment.

As we are framing the Human-AI interaction in perceptual decision making, there is an error-prone human observer and an AI that might be useful. The resulting human-AI interaction can be formalized as a signal detection problem with two observers (Bartlett & McCarley, 2017; Bechar et al., 2009). For example, take a human and an AI, each with d' of 2.5 (comparable to performance in screening mammography). If the two observers are uncorrelated, the optimal joint d' would be about 3.5. As Bartlett and McCarley (2017) note, however, "In practice, unfortunately, people

often interact with automated aids in a suboptimal way". They may overweight or underweight the AI advice. The core problem is that there is no one optimal solution that covers all Human-AI interactions. To offer a trivial illustration, if the human is simply guessing and the AI is perfect, the obvious rule is to use the AI. If the situation is reversed, the obvious rule is to ignore the AI. The optimal use of the AI, therefore, depends on the details of the specific Human-AI interaction. Unfortunately, it is not practical to test a range of modes of interaction in real-world clinical situations. Expert observers are a limited resource and interventions (e.g. don't use AI on the next N patients) are often unethical or impractical in the real world. Our goal is to develop a "Human-AI Collaboration Test (HAICT) that can be used in the lab to identify candidate interventions that could be practically tried in real-world settings.

We can identify a non-exhaustive set of factors that will influence the AI-human interaction.

1) Absolute and relative skill of the human and AI: This defines a 2D space of conditions where the human and AI might both be very skilled or not or where one is more skilled than the other.
2) Signal & noise variability: The physical signal, presented to the human and/or the AI has some intrinsic variability. That variability could be different for signals (Do all masses look alike and are they all imaged with equal fidelity?) and for 'noise' (Do other possible masses all look alike or do they come in a wide range of forms?).
3) Correlation of AI and human responses; Are the human and AI responding in the same way to the same thing? The potential benefit of an AI declines as its correlation with the human increases. Thus, the situation is different if both parties are limited by the same physical signal/noise ratio as opposed to a condition where the human is looking at an X-ray while

the AI is basing its response on some entirely different measurement. Sebastian and Geisler (2018) have provided an SDT-based method for estimating the correlation between two observers.

4) Feedback: In many real-world settings, it is not available or available only incompletely, in aggregate, and/or after a significant delay. Consider screening mammography where some AI true positives will provide an immediate form of feedback when the radiologist clearly sees that CAD has marked something they found or missed. Feedback about other CAD responses would have to wait for subsequent follow-up (e.g. Was it a hit or a false alarm?). In the case of miss errors, by either the human or the AI, it is possible there is no feedback.
5) Costs and benefits of different outcomes: In mammography, for example, FP errors carry modest risk and some real psychological and financial costs. The costs of FN errors are much greater. A similar situation applies to airport security screening. The could also be operational costs if, for example, adding AI doubled the time or the cost of an exam.
6) Target Prevalence: Particularly in screening situations, target prevalence is an important variable. Breast cancer is rare in a screening population. Real threats are even rarer at the airport checkpoint. This obviously interacts with costs and benefits. Are 100 or 1000 FP errors more or less expensive than one FN error? Prevalence is known to influence human behavior (Levari et al., 2018; Wolfe et al., 2005).

This partial list describes a vast dataspace. It would be neither practical or fruitful to systematically cover it with experiments. One possible solution would be to model the effects. A valiant effort was made by Bechar et al (2009) who proposed to compute “An Objective Function to Evaluate Performance of Human-Robot Collaboration in Target Recognition Tasks” (Bechar et al., 2009).

Bechar et al created a linear model that incorporates many (but not all) of the variables described above and they could draw some interesting conclusions from the output. For instance, for their conditions, the utility of the AI increases as target prevalence increases; a potentially depressing conclusion for CAD systems deployed to help with the detection of rare targets. Their conclusions include the statement, "Since the number of parameters is large and, in addition, there are interactions between the parameters, it is difficult to predict the system performance".

How, then, can progress be made on deciding on the best use of an AI system? In this project, we propose a testbed that captures critical aspects of human-AI interaction in a signal detection paradigm. The idea is that a simple, 2-alternative forced-choice (2AFC) task can be tailored to have the properties of the real-world situation. Stimuli can be created to constrain the d' of the observer and the AI. The prevalence of targets can be set along with a reward structure for different types of correct and incorrect responses. Feedback can approximate the real-world situation. Having built a task with critical similarities to the real-world task, it is possible to test non-expert observers in relatively short experiment that manipulate factors like the timing of presentation of AI information or the criterion used by the AI. Such experiments will not eliminate the need to test hypotheses with real experts doing the real task. However, results from this Human-AI Collaboration Test (HAICT) can guide clinical testing to the more promising possibilities.

In the experiment described here, we take a single, hypothetical AI and use it as a *second reader* – presenting the AI opinion after the human has offered an initial reading.

HAICT methodology

HAICT treats the human-AI interaction as a signal detection problem with two observers. We generated a set of visual stimuli that both our human and "AI" observer must evaluate. This is analogous to a CADx situation where a human reader might use a CAD system to help determine if a mass is cancerous.

At the heart of the HAICT method is the use of artificial stimuli that can be precisely controlled in ways that real medical images cannot be. Thus, as shown in Figure 2, the stimuli were 20x20 square grids in which each square was a shade of red or green.

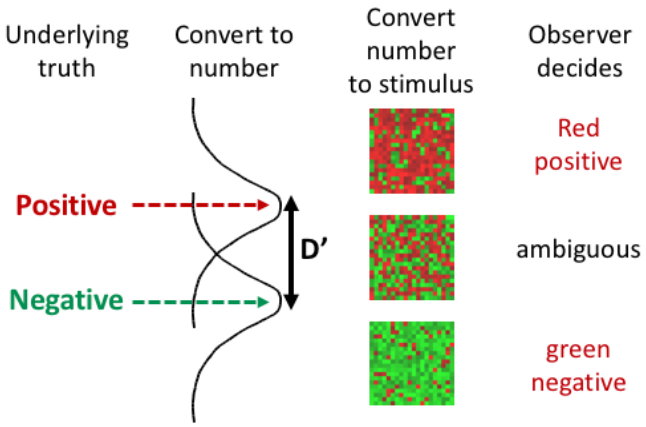


Figure 2: Construction of HAICT stimuli

Whether a case is "positive" or "negative" determines the numbers of red and green cells in the stimulus. For a negative trial, the number of red cells is derived from one normal distribution. For positive trials, the number is derived from another distribution with a higher mean. The difference in those means, converted to d' puts an upper limit on performance. That is, if the distributions are separated by d'=2.5, as they are in this case, even the most 'expert' observer cannot produce results with d' better than 2.5, except by chance. This method means that some positive cases will look negative and vice versa. It is important to explain this to participants, especially if the experiment involves feedback in order to avoid confusion and frustration.

Human and AI stimuli can be the same or they could be generated separately so that the maximum performance of the two observers can be independently varied. In this experiment, both the AI and human have a maximum performance of a d' of 2.5.

Specific Method

We tested 13 naïve observers (avg age 31, 8 female, 5 male). All had normal color vision as assessed by Ishihara color plates and acuity of 20/25 or better.

Two conditions were be tested:

1. Baseline - No AI input. Observer classifies each case as "bad" or "not" bad on their own.
2. Second Reader - The observer makes an initial decision about every case. The simulated AI silently classifies stimuli using a "conservative criterion" (c = 0.5). The logic for the conservative criterion is that the second reader is being used to cut down on false positive responses and so it is intended to raise questions about positive human responses that might be marginal. If the observer and AI disagree, then the AI informs the human observer. The observer is then given a chance to either change their response or go with their first opinion.

In each of these conditions, the prevalence of "positive cases" is varied in separate blocks. The percentage of targets could be 10%, 33%, 67% or 90%. Humans are known to become more conservative at low prevalence, missing more targets (Horowitz, 2017). Thus, there were 8 blocks per Observer (2 conditions X 4 levels of prevalence). Each Observer ran 200 trials in each block. Block order was randomized.

Stimuli were present in a darkened room on Macintosh desktop computers with Os at an approximate 60 cm viewing distance.

Data analysis

One observer was removed for poor performance. For the remaining 12 Os, we performed standard signal detection analysis. For each condition, we computed the probability of a "hit" or true positive response (correct positive responses / total positive trials) and "false alarms" or false positive response (false positive responses / total negative trials). These values are converted to "z-scores" (=inverse of the standard normal cumulative distribution), zHit & zFA. From those values, d' = zHit-zFA and criterion, 'c', = (zHit-zFA)/-2. Response time can also be looked at but is not of particular interest in this experiment.

References

Bartlett, M. L., & McCarley, J. S. (2017). Benchmarking Aided Decision Making in a Signal Detection Task. *Human Factors*, *59*(6), 881-900. https://doi.org/10.1177/0018720817700258

Bechar, A., Meyer, J., & Edan, Y. (2009). An Objective Function to Evaluate Performance of Human-Robot Collaboration in Target Recognition Tasks. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *39*(6), 611-620. https://doi.org/10.1109/TSMCC.2009.2020174

Beck, H. P., Dzinholet, M. T., & Pierce, L. G. (2007). Automation Usage Decisions: Controlling Intent and Appraisal Errors in a Target Detection Task. *Human Factors*, *49*(3), 429-437.

Berlin, L. (2007). Accuracy of Diagnostic Procedures: Has It Improved Over the Past Five Decades? *American Journal of Roentgenology*, *188*(5), 1173-1178. https://doi.org/10.2214/ajr.06.1270

Greenes, R. A. (2014). *Clinical Decision Support* ((Second Edition) ed.). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-398476-0.00001-4

Hautus, M. J., Macmillan, N. A., & Creelman, C. D. (2021). *Detection Theory*. Routledge.

Hoff, K., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, *57*(3), 407-434. https://doi.org/10.1177/0018720814547570

Horowitz, T. S. (2017). Prevalence in Visual Search: From the Clinic to the Lab and Back Again. *Japanese Psychological Research*, *59*(2), 65-108. https://doi.org/10.1111/jpr.12153

Kundel, H. L., Nodine, C. F., & Carmody, D. (1978). Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Invest Radiol*, *13*(3), 175-181. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=711391

Levari, D. E., Gilbert, D. T., Wilson, T. D., Sievers, B., Amodio, D. M., & Wheatley, T. (2018). Prevalence-induced concept change in human judgment. *Science*, *360*(6396), 1465-1467. https://doi.org/10.1126/science.aap8731

Li, Q., & Nishikawa, R. M. (2015). *Computer-aided detection and diagnosis in medical imaging*.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory*. Lawrence Erlbaum Assoc.

Monticciolo, D. L., Newell, M. S., Hendrick, R. E., Helvie, M. A., Moy, L., Monsees, B., Kopans, D. B., Eby, P. R., & Sickles, E. A. (2017). Breast Cancer Screening for Average-Risk Women: Recommendations From the ACR Commission on Breast Imaging. *Journal of the American College of Radiology*, *14*(9), 1137-1143. https://doi.org/http://dx.doi.org/10.1016/j.jacr.2017.06.001

Sebastian, S., & Geisler, W. S. (2018). Decision-variable correlation. *Journal of Vision*, *18*(4), 3-3. https://doi.org/10.1167/18.4.3

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J. R., & Hopman, R. J. (2016). Talking to Your Car Can Drive You to Distraction. *Cognitive Research: Principles and Implications*, *1*(16). https://doi.org/https://doi.org/10.1186/s41235-016-0018-3

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare targets are often missed in visual search. *Nature*, *435*(7041), 439-440. https://doi.org/10.1038/435439a