Cognitive Mechanisms for Calibrating Trust and Reliance on Automation

Leslie M. Blaha (leslie.blaha@us.af.mil)
711^{th} Human Performance Wing, Air Force Research Laboratory, Pittsburgh, PA 15213 USA

Christian Lebiere (cl@cmu.edu)
Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Corey K. Fallon (corey.fallon@pnnl.gov) and Brett A. Jefferson (brett.jefferson@pnnl.gov)
Pacific Northwest National Laboratory, Richland, WA 99352 USA

Abstract

Trust calibration for a human-autonomy team is the process by which a human adjusts their understanding of the automation's capabilities; trust calibration is needed to engender appropriate reliance on automation. Herein, we develop an Instance-based Learning ACT-R model of decisions to obtain and rely on an automated assistant for visual search in a UAV interface. We demonstrate that model matches well the human predictive power statistics measuring reliance calibration; we obtain from the model an internal estimate of automation reliability that mirrors human subjective ratings. Our model is a promising beginning toward a computational process model for trust and reliance for human-machine teaming.

Keywords: Cognitive architectures; Trust in automation; Human-machine teaming

Introduction

Trust calibration is the process team members go through to adjust their attitudes or expectancy of a favorable response from other teammates or of a positive outcome of a team effort (Lee & See, 2004). Research in both all-human teams and human-automation teams indicates that trust, and its behavioral proxy reliance, fluctuate over time. For human-automation teams, this calibration-related fluctuation reflects the human's process of learning when to rely on the automation. Fallon, Murphy, Zimmerman, and Mueller (2010) describe this as a sensemaking process wherein the human learns the conditions under which automation performs well, and how to properly interpret indicators provided by the machine system, to promote "appropriate use" (see also, Lee & See, 2004). Without calibration, the team may suffer from automation misuse or disuse by the human teammates (Lee & See, 2004; Parasuraman & Riley, 1997).

Lee and See (2004) formulated a conceptual model of appropriate trust formation. Within this model, trust calibration is part of a closed-loop process wherein people use task goals, context, and their own beliefs (including current trust level) to form an intent about using automation and then take Reliance Actions. The subsequent behavior of the automation and impact on the world (witnessed directly or via display) feed back into the human's Information Assimilation and Belief Formation processes, which then feed the Trust Evolution process, the Intent Formations, and Reliance Actions. Lee and See argued that information available about the automation and the results of Reliance Actions are critical to the trust formation process (as do Chen & Barnes, 2014; Fallon et al., 2010; Lyons et al., 2016, and many others). Merritt and Ilgen (2008) refer to trust emerging from interactions and experience with a system as history-based trust; they contrast history-based trust with other forms of trust, such as a person's general tendency to trust (dispositional trust; Jessup, Schneider, Alarcon, Ryan, & Capiola, 2019; Kramer, 1999; Merritt & Ilgen, 2008). It follows from this perspective that appropriate calibration can be defined as the degree of correspondence between a person's trust in automation and the automation's capabilities (see also, de Visser et al., 2020; Lee & Moray, 1994; Muir, 1987).

In this work, we develop a computational cognitive model of human decisions to rely on automation using Instance-based Learning Theory (IBLT; Gonzalez, Lerch, & Lebiere, 2003). Using an IBL model, we can explicitly model decisions about automation reliance and observe how those decisions are informed by task performance, transparency information, and the automation's behavior over time. In this way, we implement a computational processes mirroring elements of Lee and See's (2004) conceptual model.

Importantly, we do not explicitly incorporate a trust mechanism in the IBL model; rather, we maintain the perspective that trust is an attitude, separate from the cognitive decision making mechanisms, and that reliance is the behavioral indicator of trust. We seek to understand if and how trust calibration emerges from the task experience and decision making processes that are explicitly defined within the IBL model. In the remainder of this paper, we will describe an experiment on trust calibration measuring both intention formation and reliance action decisions. Then we will describe the IBL model and demonstrate its performance on this two-stage task. We will show that model reliance decisions mirror the human behavior, and we can extract an internal model bias that parallels human subjective judgments of automation reliability. We conclude that we have a strong candidate computational process model for trust calibration through experience with automation.

The Human-Automation Teaming Task

We leverage empirical data collected by Fallon, Blaha, Jefferson, and Franklin (2019) using the COgnitive Behavioral AnaLytics Testbed (COBALT). COBALT is an experimental interface developed by Fallon and Blaha to enable the study of trust, reliance on automation, task performance, and cognitive

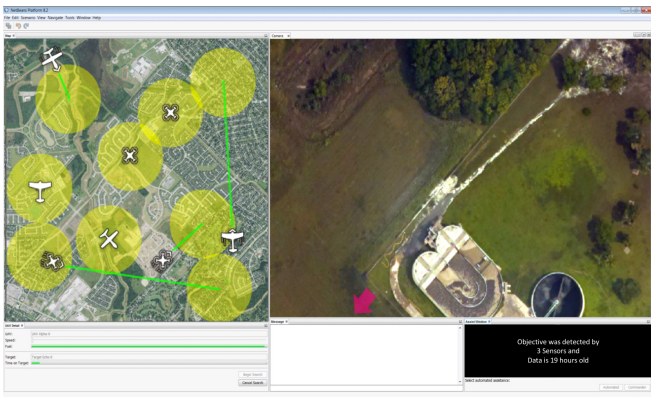

Figure 1: A mid-trial screenshot of the COgnitive Behavioral AnaLytics Testbed (COBALT) task environment. This image shows an AUTOASSIST cue (pink arrow) selected for search, with an age + number text transparency cue (lower right). Readers are referred to Fallon et al. (2019) for more details about the interface.

workload while manipulating task characteristics, automation transparency, and interface design choices. COBALT is comprised of modular windows in which participants interact with automation to search for objectives in an aerial imagery. Figure 1 provides a screenshot of COBALT.

Two-stage Trial Structure Each trial of the task involves two stages: a decision stage and a search stage, as diagrammed in Figure 2. In the decision stage, participants must decide whether they would like the AUTOASSIST or COMMANDER to aid their visual search. Participants cannot perform the search without selecting an aid. Participants are provided transparency cues to help them decide if the AUTOASSIST will be a reliable choice.

The search stage begins as soon as the assist type is selected. Participants are tasked with searching for a predetermined objective randomly placed in the image. Participants are provided with search guidance in the form of an arrow overlaid on the search window. When reliable, this cue points directly to the objective; when unreliable, it points to some other random location. Participants can choose to follow the assist cue or search unguided.

Using the terminology of Lee and See's trust calibration model, the assist selection stage is an example of an automation reliance Intention Formation; participants indicate intention about using automation when they select the AUTOASSIST search aid. The search stage is an example of a Reliance Action. Participants following the AUTOASSIST cue are relying on automation; participants searching unguided are not.

Assist Types The assist types varied in their reliability and timing. The COMMANDER option provided a 100% reliable cue, always pointing to the objective. However, there was a 5 sec. delay between selecting the COMMANDER button and the COMMANDER assist arrow appearing on screen to aid the participant. While waiting for the COMMANDER assist, participants can search unaided for or wait for the COMMANDER cue to appear.

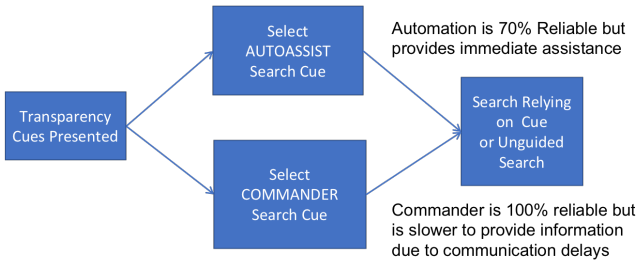

Figure 2: Diagram of the trial stages. A trial starts with presentation of transparency cues. Participants make a 2AFC search assist cue selection. Then they complete the visual search by either relying on the selected assist or searching unguided.

AUTOASSIST simulates automation to provide a search cue. Unlike the COMMANDER, it is available immediately at the start of the search stage. However, AUTOASSIST is only 70% reliable, meaning it correctly pointed to the objective on 70% of trials and to a random location on the others.

Automation Transparency Cues Transparency information was provided on every trial to aid participants in their assist type decisions. Participants could use the cues to learn when the AUTOASSIST would be unreliable. The two types of transparency cues were: the age of the data and number of sensors available to the automation. The number of sensors ranged from 1 to 3, and AUTOASSIST was unreliable if there was only 1 sensor. The data age varied from 1 to 36 hours old, and AUTOASSIST was unreliable if the data was over 24 hours old.

Fallon et al. (2019) used four transparency cue conditions. In the age-only condition, a statement about the age of the sensor data was given; no information about the number of sensors was provided. In the number-only condition, a statement about the number of sensors was given; no information about the age of the data was provided. In the age + number text condition, a statement included both the age and number information. And in the age + number graphic condition, the combination of age and number information was presented in a visual representation leveraging a circle-packing graphic.

Feedback An important component of modeling learning from experience is accounting for the feedback received about the outcome of one's decisions. We model two types of feedback received by participants during the search stage. The first was direct observation of success or failure of the AUTOASSIST. On trials when the AUTOASSIST was unreliable, it would visibly fail by disappearing from the screen at the moment of failure.

The second source of feedback was the total time to execute each search; participants were not given explicit timing information, but experienced how long each search took and

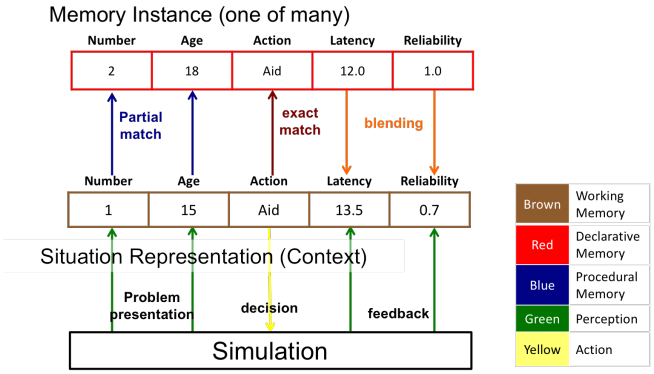

Figure 3: Diagram of the IBL model for the COBALT task decisions. The colors indicate the ACT-R mechanisms. See the text for a description of the instance representation. The upper half shows one trial as an instance in working memory (middle row), which is matched to similar prior instances in declarative memory (top row) through the matching and blending functions (blue, red and orange arrows). The lower portion of the diagram shows the perception and action mechanisms interacting with the task interface.

if multiple assist cues had to be selected to locate the objective (participants could request the COMMANDER after the AUTOASSIST failed, for example). The fastest search trials occurred when a participant selected a reliable AUTOASSIST and followed it directly to the objective. The slowest search trials occurred when a participant selected an unreliable AUTOASSIST, followed it and observed its failure, and then attempted some combination of unguided search, calling and waiting for the COMMANDER assist, and then relying on the COMMANDER assist to locate the objective. In this way, the negative feedback from long search times with unreliable automation might influence experience differently than the positive reinforcement and short search times associated with relying on reliable automation.

Human Data Sixteen participants completed the four transparency conditions (Age, Number, Age + Number Text, and Graphic) of the COBALT task, for a within-subjects manipulation. A single condition contained 13 blocks of 11 trials for a total of 572 trials per participant. Transparency condition orders were randomized between participants.

Cognitive Model of Automation Reliance

We developed a model of the automation reliance decisions in COBALT task with Instance-base Learning Theory (IBLT; Gonzalez & Dutt, 2011; Gonzalez et al., 2003) implemented in the ACT-R cognitive architecture (Anderson et al., 2004). IBL a methodology for modeling problem solving and decision making that relies on previous experiences rather than pre-defined strategies. Those experiences are stored in the declarative memory of the cognitive architecture, whose mechanisms support adaptive storage and associative retrieval. Experiences are stored in memory as a combination of situation features, decision taken, and observed outcome.

Memory instance availability is controlled by activation:

$$A_i = \log\left(\sum_{j=1}^{N} t_j^{-d}\right) \quad (1)$$

where i is the memory and d is the decay parameter controlling the power law of recency; the summation over all references to that memory provides the power law of practice. Given a particular situation, relevant memories are retrieved by computing their match score that combines their activation with their degree of relevance:

$$M_i = A_i + \sum_{j=1}^{l} MP \times Sim(d_j, v_{ij}) \quad (2)$$

where j is a feature in the situation representation, d_j is the corresponding value in the current situation, v_{ij} is the corresponding value in memory i, and Sim is the similarity between those two values. Rather than retrieving a single memory, a consensus outcome is generated using the memory blending mechanism satisfying this constraint:

$$V = \underset{V_j}{\text{argmin}} \sum_{i=1}^{k} P_i \times Sim(V_j, v_{ij})^2 \quad (3)$$

where V is the consensus value among the set V_j of possible values, and P_i is the probability weight of memory i, reflecting its match score M_i through a Boltzmann softmax distribution.

Our IBL model adopts a straightforward representation of the problem. Examples are shown in Figure 3, where the middle row is a current trial instance, and the top row is one similar instance from declarative memory. The situation features are the age and/or number cues; the decision is whether to rely on the COMMANDER or AUTOASSIST (labeled *aid* in Figure 3), and the outcome is whether the AUTOASSIST was reliable (*Reliability*) and time to complete the visual search (*Latency*). To make a decision, the model generates an expected outcome for each assist type by performing blended retrievals for the specific situation feature(s) available (age, number, both) and each assist type, extracting an expected value for total search time. The model selects the assist type with the lowest expected search time. It then generates an expectation for the reliability of the automation in a similar manner, using a blended retrieval over situation feature(s) and selected assist type. The model then executes the option, and stores a new instance combining that situation's feature(s), the option chosen, and the outcomes experienced in terms of reliability and search latency. Finally, at the end of each condition, the model generated a general expectation of reliability through a blending retrieval with no features specified.

IBL models need either a back-up strategy (such as random exploration) to get started, or some initial instances to bootstrap the process. We chose the latter route, creating three instances to represent as broad a range of outcomes as possible. Those instances could have resulted from a short practice phase, or fairly straightforward reflection upon the instructions; both instructions and a few practice trials were given to

Table 1: Signal Detection Theory Mapping of Automation Reliance Intention Formations

	Actual Reliability of AUTOASSIST	
	Reliable	Unreliable
AUTOASSIST Selected	True Positive	False Positive
COMMANDER Selected	False Negative	True Negative

COBALT participants. The first instance featured the most reliable cues (3 sensors and 1-hour-old data), a decision to rely on AUTOASSIST, and outcomes of reliable AUTOASSIST and fastest search time (directed search time of 3 seconds). The second instance featured the least reliable cues (1 sensor and 36-hours-old data), a decision to rely on AUTOASSIST, and outcomes of unreliable AUTOASSIST, and the slowest search time (random search time of 15 seconds). The third instance featured average cues (2 sensors and medium age), a decision to rely on COMMANDER, and outcomes of reliability and an intermediate search time (wait then direct search for a total time of 8 seconds). We use ACT-R default parameters: decay $d = 0.5$; activation noise $s = 0.25$; mismatch penalty factor $MP = 1.0$; linear similarities over $[0, -1.0]$.

Results

We focus on three aspects of the data collected by Fallon and colleagues: decision stage intention formation choices, search stage automation reliance actions, and the subjective ratings of the AUTOASSIST's reliability. We consider together the human and model data. Our goal is to evaluate if the model captures well the human behaviors and if the IBL model's internal representation reflects trust calibration.

Predictive Power Metrics We quantify reliance calibration with predictive power analysis, based on a signal detection theory (SDT) characterization of automation use decisions (Feinstein, 1975). SDT quantifies the decision rates about the two assist types balanced with the ground truth of the AUTOASSIST's reliability. For the decision stage, we define a true positive as a decision to request AUTOASSIST when reliable and a false positive as a decision to request AUTOASSIST when it is unreliable. Table 1 defines all four SDT categories for the decision stage, reflecting intention formation accuracy, and Table 2 gives the definitions in the search stage for reliance actions accuracy.

We define positive predictive power:

$$\text{PPP} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}. \quad (4)$$

PPP gives the proportion of trials a participant appropriately chose AUTOASSIST out of all trials on which the participant selected the AUTOASSIST option. We define negative predictive power:

$$\text{NPP} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Negative}}. \quad (5)$$

NPP gives the rate at which the participant appropriately did *not* select the AUTOASSIST (selected COMMANDER in the decisions stage or did not follow an unreliable search cue) when it would have been unreliable, out of all the trials on which the participant did not select AUTOASSIST.

We selected PPP and NPP as our metrics for appropriate reliance because they reflect the decision maker's ability to correctly select the automation when it will be reliable and not select the automation when it will be unreliable, respectively, while accounting for the prevalence of reliable and unreliable trials in the experiment. Accounting for the base rate of reliability is a core part of the definition of trust/reliance calibration. We note the more common SDT metrics d' and β (decision criterion) have been used in many studies to examine human judgements about the reliability of alarms or automation recommendations. These metrics emphasize the participants' abilities to discriminate signal cues from noise or non-signals. Application in the present study would measure participants' abilities to discriminate the transparency cues indicating the AUTOASSIST's reliability; the emphasis is on how participants internally represent the transparency cues. Understanding this internal representation is important for selecting effective transparency cues, but our present interests are more about quantifying decision makers' automation reliance, informed by those cues. PPP and NPP better serve this goal. Additionally, there is evidence that PPP and NPP better reflect the time-varying nature of decision-making processes without changing their statistical properties (Repperger, Warm, Havig, Vidulich, & Finomore, 2009).

Assist Selection Decisions Predictive Power Figure 4 (top) gives the predictive power for both the humans and models in the assist selection decision stage. The bars give the means, and the points are the individual decision makers. For the human decision makers, PPP and NPP are fairly high. PPP (right) is similar across all transparency cue conditions; NPP (left) shows a bit more variability, with the highest NPP observed in the Number-only condition. NPP for humans is comparable to their PPP. Between both metrics, we can infer that generally people chose the appropriate assist more often than the inappropriate one.

The model closely reproduced the average level of PPP in all conditions and NPP in the text and graphic (two-cue) conditions. The model underestimates NPP in the Age-only and Number-only conditions, meaning the model makes a higher number of false negative decisions than humans. This discrepancy might result from transfer between conditions, as the model currently makes the assumption that no transfer occurs across conditions because of distinct representations of situation features. It is possible that participants relied on information between conditions, improving performance on single cues, relative to the model lacking between-condition learning. Recent increases in representation flexibility in the

Table 2: Signal Detection Theory Mapping of Reliance on AUTOASSIST Search Cues

	Actual Reliability of AUTOASSIST	
	Reliable	Unreliable
AUTOASSIST Followed	True Positive	False Positive
AUTOASSIST Not Followed	False Negative	True Negative

ACT-R architecture enables us to explore alternative assumptions in future work.

Reliance on Search Assist Predictive Power The second way we quantify reliance on automation is to further use Equations 4 and 5 in the search stage to examine the proportion of trials wherein people followed the AUTOASSIST cue when it was or was not reliable. Table 2 summarizes the SDT definitions for AUTOASSIST search reliance actions. Here, we consider only the subset of trials on which participants selected AUTOASSIST in the decision stage, because there was no automation reliance action when COMMANDER was selected. PPP with the Table 2 mapping is the proportion of trials on which a participant followed a reliable AUTOASSIST out of all trials on which participants followed the AUTOASSIST; NPP is to the proportion of trials on which the participant did not follow the unreliable AUTOASSIST cue out of all trials on which they did not follow the AUTOASSIST.

Figure 4 (bottom) shows the distributions of PPP (right) and NPP (left) for the search stage of the COBALT trials. The means for PPP are higher than NPP, within each measure, the human means are similar across all conditions. The distributions for NPP have a larger variance, in addition to the lower means. The low (approximately .25) NPP means that the participants are taking more false negative reliance actions than true negatives. Compared to the decision stage (Figure 4 upper), the search NPP means are much lower; PPP distributions and means are similar in the two task stages.

In the search stage, the model generates expectations of the AUTOASSIST's reliability, which we translated into predictive power measures. The model qualitatively reproduced both PPP and NPP behaviors. We observe that the larger variability for NPP might result from individual differences in strategy, which we plan to explore in future work.

Perceived Reliability of the Automation An exciting result that emerges from the model is an estimate of the probability of overall automation reliability that appears to parallel the human subjective ratings.

At the completion of each condition, participants were asked to estimate the AUTOASSIST's correctness for that condition. Ratings were given as a value between 0 and 100%. The ground truth automation reliability was always 70%. Figure 5 gives the means and individual ratings from

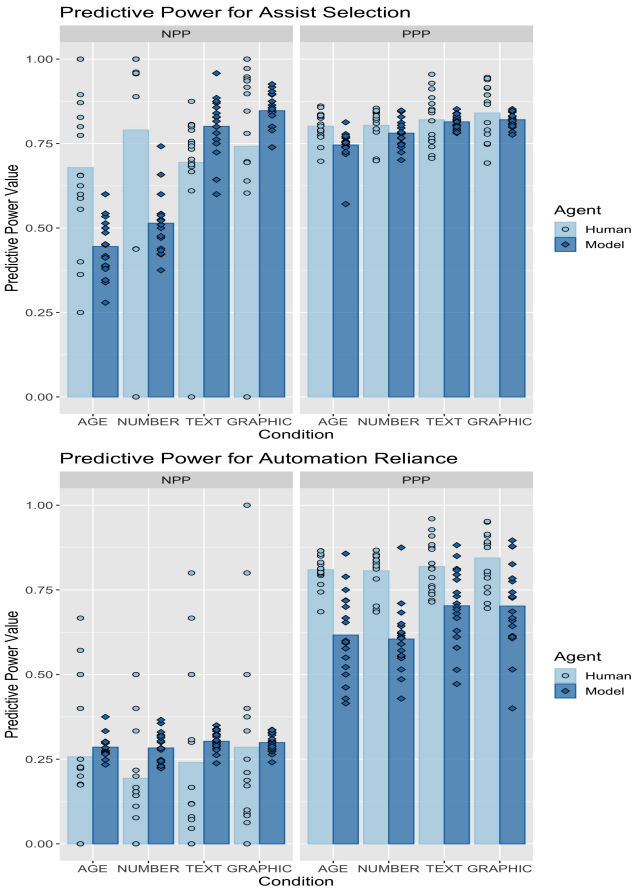

Figure 4: Distributions of predictive power values for all conditions for the decision stage (top) and search stage (bottom) of the COBALT trials.

both humans and IBL models; a horizontal line indicates the actual automation reliability. As shown in Figure 5, on average, both people and models over-estimated the AUTOASSIST's reliability. Over-estimations were larger in the text and graphic conditions than in the single cue conditions.

Previous efforts established the ability of IBL models to reproduce human cognitive biases resulting from the interaction of cognitive mechanisms and task statistics (Lebiere et al., 2013). This predictive basis for judgments of (over)trust raises the potential of using cognitive models to support human-machine teaming in ways that automatically compensate for human biases. Importantly, as conceptualized by definitions of trust calibration, internal estimates of reliability were shaped through reliance experiences.

Relationship to Conceptual Trust Calibration Model

Our IBL model's performance provides empirical support for the closed-loop dynamic calibration process of Lee and See's (2004) model. However, our process model does not yet integrate the moderating factors outlined in their conceptual model. Despite only formalizing the cognitive mechanisms in Lee and See's (2004) feedback loop, our approach was still

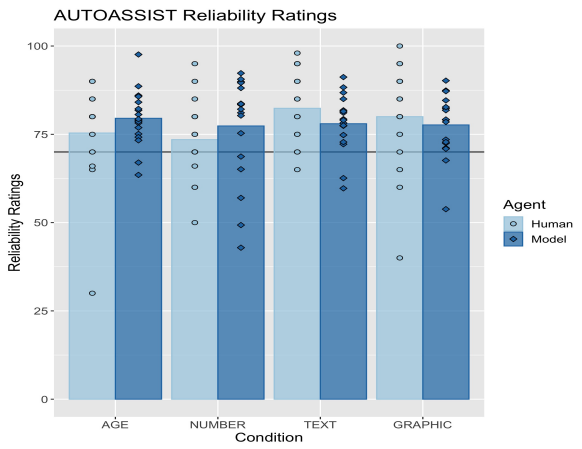

Figure 5: Mean perceived reliability ratings (bars) and individual reliability ratings (points) each decision maker. The light blue is the human subjective ratings data; the dark blue are reflect blended values in the IBL model. The horizontal line at 70 indicates the ground truth automation reliability.

able to fairly closely mimic human responses. IBL model performance suggests that the individual, organizational, cultural and environmental context played a less important role in influencing trust calibration within this controlled task environment. In some ways, these findings are not surprising because we attempted to control for (and did not manipulate) many of these variables. What is less clear from our findings is whether our model's ability to simulate trust calibration would generalize to other less constrained environments where individual, organizational, cultural and environmental context might be more influential. If they do, such findings would suggest that the feedback loop in the bottom portion of Lee See's model is the most powerful driver of trust calibration. Perhaps the experience gained from interacting with the automation has such a powerful impact on trust and reliance calibration that simply modeling this cycle is sufficient to replicate human trust dynamics. The impact of organization, culture environment and individual differences must be explored; the IBL model should allow for a systematic investigation into the impact of these variables on trust calibration.

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060.

Chen, J. Y., & Barnes, M. J. (2014). Human–agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems*, *44*, 13–29.

de Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, *12*, 459–478.

Fallon, C. K., Blaha, L. M., Jefferson, B., & Franklin, L. (2019). A capacity coefficient method for characterizing the impacts of automation transparency on workload efficiency. In *Proceedings of the Human Factors and Ergonomics Society* (pp. 827–832).

Fallon, C. K., Murphy, A. K., Zimmerman, L., & Mueller, S. T. (2010). The calibration of trust in an automated system: A sensemaking process. In *2010 International Symposium on Collaborative Technologies and Systems* (pp. 390–395).

Feinstein, A. R. (1975). XXXI. On the sensitivity, specificity, and discrimination of diagnostic tests. *Clinical Pharmacology & Therapeutics*, *17*(1), 104–116.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review*, *118*(4), 523–551.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591–635.

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The measurement of the propensity to trust automation. In *International conference on human-computer interaction* (pp. 476–489).

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology*, *50*(1), 569–598.

Lebiere, C., Pirolli, P., Thomson, R., Paik, J., Rutledge-Taylor, M., Staszewski, J., & Anderson, J. R. (2013). A functional model of sensemaking in a neurocognitive architecture. *Computational Intelligence and Neuroscience*, *2013*, 921695.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*(1), 153–184.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*, 50–80.

Lyons, J. B., Koltai, K. S., Ho, N. T., Johnson, W. B., Smith, D. E., & Shively, R. J. (2016). Engineering trust in complex automated systems. *Ergonomics in Design*, *24*(1), 13–17.

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

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, *27*(5-6), 527–539.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253.

Repperger, D., Warm, J., Havig, P., Vidulich, M., & Finomore, V. (2009). Modifying sensitivity/specificity for sensors using positive and negative predictive power measures. In *IEEE 2009 National Aerospace & Electronics Conference (NAECON)* (pp. 190–194).