Self-Awareness in Machines Boosts Human Trust

Dana Warmsley, Krishna Choudhary, Jocelyn Rego, Emma Viani, Praveen Pilly

HRL Laboratories, Malibu, CA 90265, USA
dmwarmsley@hrl.com, kchoudhary@hrl.com,
jrrego@hrl.com,emma.i.viani@gmail.com,pkpilly@hrl.com

Abstract. Low trust in autonomous systems remains a significant barrier to adoption and performance. To effectively increase trust in these systems, machines must perform actions to calibrate human trust based on an accurate assessment of both their capability and human trust in real time. Existing efforts demonstrate the value of trust calibration in improving team performance, but overlook the importance of machine self-assessment capabilities in the trust calibration process. In our work, we develop a closed-loop trust calibration system for a human-machine collaboration task to classify images and demonstrate about 40% improvement in human trust and 5% improvement in team performance with trained machine self-assessment compared to the baseline, despite the same performance level between them. Our trust calibration system applies to any semi-autonomous application requiring human-machine collaboration.

Keywords: Machine Self-Assessment · Trust in AI · Autonomous Systems.

1 Introduction

Low trust in autonomous systems remains a significant barrier to adoption and performance, especially in high-stakes, safety-critical missions. A clear signal of this problem is the high frequency of human takeover events when the system's behavior does not match human expectations, or the human is insufficiently confident in its situational understanding. In this work, we develop a closed-loop trust calibration system that improves trust via reliable machine self-assessment to proactively align human trust with machine capability in real time.

The contributions of this work are three-fold: (1) Developed a closed-loop trust calibration system that leverages real-time trust prediction, machine self-assessment, and a dynamic reasoning model that determines the best machine action for trust calibration. (2) Empirically compared cumulative trust of human subjects in machines with learned self-assessment to those that lack it. (3) Developed a paradigm to rigorously assess the effects of trust modeling and self-awareness in machines on human trust for operationally relevant contexts.

2 Prior Work

Early work in trust calibration focused on transparency, which involved consistently offering uncertainty information, confidence estimates, or system reliability to encourage appropriate trust in the machine [9][15]. More recently, efforts shifted to *adaptive* trust calibration, where the system either selectively determines when to provide information to the user to calibrate trust (e.g., when a human exhibits over- or under-reliance), or uses information about the user to adapt its behavior. Adaptive trust calibration efforts are of wide interest since having the human continually monitor information cues can increase workload [8] [1], and adaptation allows for personalization to the individual. Here, we highlight some recently developed adaptive trust calibration systems relevant to his paper. See [13] for a recent survey on trust calibration.

[10] developed a framework for offering trust calibration cues when over- and under-trust were detected. Over-trust occurs when the human incorrectly believes the machine will perform the task better, and under-trust happens when the human falsely believes the machine will perform worse. They found that adaptively offering cues (visual, audio, verbal, etc.) improved trust calibration and team performance. [5] presented Pred-RC to adaptively select when to provide reliance calibration cues, where reliance is considered an observable trust-related behavior. If Pred-RC determined that the probability of reliance is higher with the cue than without and the probability of machine success is high, the cue is shown to encourage reliance. Pred-RC reduced the number of cues needed while avoiding over/under-reliance on the machine.

[2] learned a Partially Observable Markov Decision Process model that used inferred trust values to determine what robot actions would maximize team performance. In a table-clearing task, the robot learned to build human trust by clearing low-risk objects (high-risk objects) when trust was low (high). [1] developed a POMDP that modeled the effects of automation reliability, transparency, scene complexity, gaze behaviors, and reliance on human trust and workload dynamics in Level 2 driving scenarios. The model was leveraged to use current human trust and workload levels to calculate the optimal level of system transparency necessary to calibrate trust in real time.

We model our own experiments after [7], which investigated trust calibration, compliance, and transparency in an autonomous image classifier. They tested whether showing the classifier's confidence values would increase trust in it. They found that trust was largely based on system performance (accuracy) and did not increase as a result of presenting system confidence information to the human. We hypothesize that they did not see an overall increase in trust because they used class probabilities as proxies for system confidence, which has shown to be a poor method for self-assessment [6]. Accurate machine self-assessment is critical since cues to calibrate trust can actually worsen calibration if they are not reliable [16].

In this work, we developed a closed-loop trust calibration system that adaptively asks for human assistance during the image classification task based on both predicted human trust and self-assessed machine capability. We place spe-

cial emphasis on accurate machine self-assessment in encouraging appropriate trust in and reliance on automation, and show in experiments that improved self-assessment capabilities increase overall trust in the machine, reduce over- and under-reliance behaviors, and increase team performance.

3 Closed-Loop Trust Calibration System

In what follows, we describe the three major components of our closed-loop trust calibration system, developed for human-machine teaming in the image classification domain. The first component is a machine self-assessment module that estimates the image classifier's confidence in its label, independent of class probabilities. The second is a real-time human trust prediction model. The third is a dynamic reasoning component that, given the classifier's confidence and the human's trust level, determines whether or not to ask the human for assistance. These components were trained and evaluated using data from two rounds of experiments in which humans worked with semi-autonomous image classifiers to classify 50 images and rated their trust after assessing machine performance for each image (details in Section 4). We present component-specific results in the following sections, and team performance-related results later in the paper.

3.1 Machine Self-Assessment

Neural networks trained as image classifiers typically have a final layer of neurons, where each neuron corresponds to a class in the dataset. The neuron with the highest probability (after softmax operation) is chosen as the image label. A widely used baseline for confidence in that label is its corresponding probability. In practice, this probability is not constrained to correlate with the accuracy of the predicted label, leading to highly $overconfident$ errors. Indeed, softmax probabilities are known to be non-calibrated, sensitive to adversarial attacks, and inadequate for detecting out-of-distribution examples [6] [4].

[4] introduced a new confidence metric based on the *True Class Probability (TCP)*, the probability of the correct class, regardless of whether that class was chosen as the predicted label by the classifier. As the TCP is not known at test time, they introduce a method to learn the confidence by implementing a separate neural network (ConfidNet) trained to estimate the TCP during training.[12] provided an alternate method to train the confidence neural network to output "correctness" instead of TCP. That is, the neural network is trained to output a value of 1 if the label is correct and 0 otherwise. We follow this method for machine self-assessment in our work. In Figure 1, we show a comparison of learned self-assessment to the baseline use of probability on a subset of images from the STL-10 dataset [3]. As expected, it outputs predominantly low values for incorrect labels, and predominantly high values for correct labels. Moving forward, we use the terms "Unaware Classifier" for the image classifier that uses class probability as a confidence score and "Aware Classifier" for the image classifier that uses learned self-assessment [12], since the network trained on top of the image classifier is aware of the latter's capability to classify images.

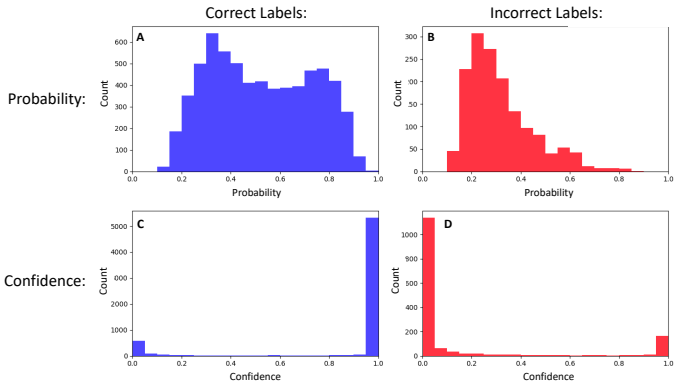


Fig. 1: The classifier produces a probability for the label (top row), which may be correct (A) or incorrect (B). These distributions have significant overlap, so probability is a poor indicator of correctness. In comparison, learned self-assessment (second row), produces values close to 1 for correct labels (C) and close to 0 for incorrect labels (D).

3.2 Trust Modeling and Prediction

In real-world applications, humans will not regularly provide feedback for the machine to assess the need for trust calibration. Trust calibration systems must predict human trust from potentially sparse information. Early approaches utilized rule-based and statistical models, with recent research shifting towards Long Short-Term Memory (LSTM) networks to better capture temporal dependencies and enhance predictive accuracy [11]. As such, our system employs an LSTM for human trust prediction based on real-time temporal data. The LSTM prediction model was trained on data collected from the first round of experiments, using the predictive features of ground truth accuracy of the image classifier (since humans reviewed machine performance in each trial and could intervene if needed), the classifier confidence in its label, and the compliance of the participant (whether a participant chose to assist the classifier when assistance is (not) requested). We used Mean Squared Error (MSE) to evaluate model performance on a validation set, for which we obtained an MSE of 1.67. Our model was then employed to predict human trust in the second round of experiments, for which it obtained an MSE of 3.3. Example prediction results for a single participant can be seen in Figure 2. Note that for our purposes, we primarily needed the model to predict general trends of trust, and not the precise trust scores as such.

3.3 Dynamic Reasoning Model

The Dynamic Reasoning Model determines when to ask for assistance based on predicted human trust and machine confidence. In round one of experiments, the model used a dynamic threshold. Machine confidence values below the threshold resulted in the machine asking for assistance. After initializing this threshold at 50% for the first trial of the experiment, the threshold was adjusted according to the compliance of human actions with the machine's request for assistance.

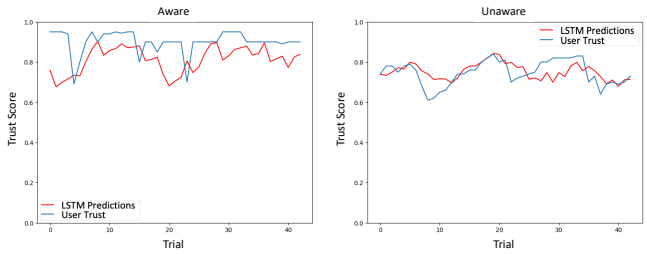


Fig. 2: Real-time results of LSTM model on a single participant. There are only 43 trials in the plots because the model requires seven previous time steps to make predictions.

If the machine asked for assistance but assistance was not given, the threshold was decreased by 5 points, thus triggering requests for assistance at relatively lower confidence values in the following trials. If the machine did not ask for help but its human partner intervened anyway, the threshold was increased by 5 points, resulting in requests for assistance at relatively higher confidence values in future trials. By dynamically changing the threshold for the machine to request assistance in this manner, we calibrate human trust based on machine capability in real time.

In round two of experiments, we replaced the threshold rule with a model trained on data from round one. We hypothesized that a model with knowledge of trust ratings, machine confidences, and instances of human intervention would be able to determine when assistance is both needed by the machine and is likely to be given by its human partner. We trained a 3-layer feedforward neural network to predict whether a human would assist the machine in an image classification trial, given the human's trust level and the machine's reported confidence in the current trial. When tested on a held-out set of examples from round one of experiments, our model reached an accuracy of 83.94% for the Aware classifier, and 82.91% for the Unaware classifier.

4 Experiments

During experiments, human participants were asked to team with an autonomous image classifier to complete an image classification task, with the goal of maximizing team performance while minimizing their own effort under time constraints. Each participant engaged in two sessions - one with the Unaware classifier (probabilities as confidence scores) and one with the Aware classifier (using learned self-assessment). We hypothesized that improved self-assessment capabilities would lead to improved overall trust and team performance since humans are more likely to trust and appropriately rely on a machine that knows when it can and cannot successfully complete a task. We offer the following hypotheses:
H1: We will observe increased overall trust in the Aware classifier, despite equal machine performance (classification accuracy).
H2: Teaming with the Aware classifier will result in a larger reduction in human

over- and under-reliance on the machine, since improved self-assessment means the machine is better able to ask for assistance when needed.
H3: Teaming with the Aware classifier will result in better team performance (classification accuracy). Reduction in over- and under-reliance behaviors reduces both machine and human error.

4.1 Experimental Paradigm

Image Classification Task: During the image classification task, a Graphical User Interface (GUI) built using the PsychoPy Python library served as the point of interaction between participants and the image classifiers. In a single session of the main task, participants were presented with 50 images consecutively. At the start of each image slide (Figure 3A), participants were shown the image, the name of the classifier (R2D2 or Wall-E), the current team score, a count down clock, the classifier's request or refusal for assistance ("I Need Assistance" or "I Do Not Need Assistance"), and "Assist" and "Do Not Assist" buttons. They were given five seconds to choose whether to assist, after which the machine submitted its own label as the team label. If the participant chose to assist, they were prompted to enter a label into a text box to stand as the team label.

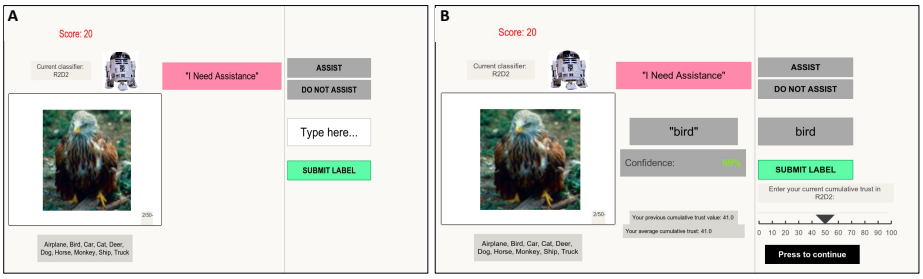


Fig. 3: The human-classifier interface during an experimental trial. Participants are instructed to choose to comply with a statement that assistance is or is not needed.

After submitting the team label (Figure 3B), the GUI displayed an updated team performance score (which is penalized when there is misalignment between machine correctness and human intervention), the classifier's label for the image, and the classifier's confidence in that label. The color of the confidence score ranges from red to green on a sliding scale, determined by the level of confidence (0-100 scale). After viewing this information, the subject was asked to report their cumulative trust in the classifier based on their overall experience with the classifier thus far on a 0 - 100 scale. Previous and average cumulative trust scores were displayed to aid the participants in keeping track of their trust development and to encourage them to view their current trust rating as cumulative.
Procedure: In an adaptation session, subjects first completed a pre-experiment survey and read through instructional slides describing the task and GUI. Instructions informed participants that they would team with a machine partner to classify images and that the machine would ask for assistance when it thinks

that its label is wrong. It was made clear that they did not have to comply with machine requests for help and that they had the option to assist even when the machine did not ask for help. Participants were instructed to report overall trust in the machine at the end of each trial, after viewing the machine's label and confidence score for that trial's image. Participants then completed a demo to ensure they understood the task and how to interact with the GUI. They then performed the core image classification task over two main sessions for the two classifiers, with a post-experiment survey to assess their overall trust in the classifier and gain insight into their impression of the classifier's performance.

Pre- and Post-experiment Surveys: The pre-experiment survey included a demographic survey on age, race, gender, country of birth, education, prior experience with image classifiers, and prior experience with semi-autonomous systems. They completed the mini-IPIP scale to assess Big Five personality traits and a well-established propensity to trust automation survey [14] before engaging with the image classifier. The post-experiment survey (Figure 5) was presented to participants after each of the Aware and Unaware sessions to gauge their overall experience with each classifier. We used a validated trust in automation survey ([14]), further replacing the term "decision aid" with "classifier" for specificity. Participants rated each entry on a scale of 1-10.

Images: We used the STL-10 dataset, which consists of 10 classes of objects (airplane, bird, car, cat, deer, dog, horse, monkey, ship, truck). To create uniformity in our experiments, we created two groups of 50 hand-picked images that were high contrast, had clear, singular objects, and were not used to train the self-assessment model. We made the selection such that the class distribution and the accuracy within each class were preserved. For example, the classifier we used in the experiments (EfficientNet-B0) accurately classified airplanes only 50% of the time, while it accurately classified ships 99% of the time. We ensured this asymmetry was reflected in the image groups that the human subjects viewed.

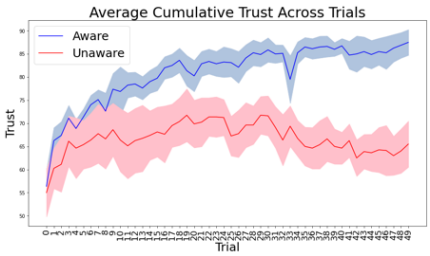
Experimental Design: We performed two rounds of experiments with 8 subjects each. In both rounds, we used a 2 x 2 x 2 counterbalanced within-subjects design where subjects were exposed to the Aware and Unaware classifiers (both with 80% accuracy), the name of the classifier (Wall-E or R2D2), and the set of 50 images presented (Group 1 or Group 2). The two rounds differed in terms of the Dynamic Reasoning model deciding when to ask for assistance (section 3.3).

5 Experiment Results - Round One

Paired t-tests were used to determine if there were significant differences ($p < 0.05$) in reported trust, over- and under- reliance on the machine, and team performance between the Aware and Unaware classifiers.

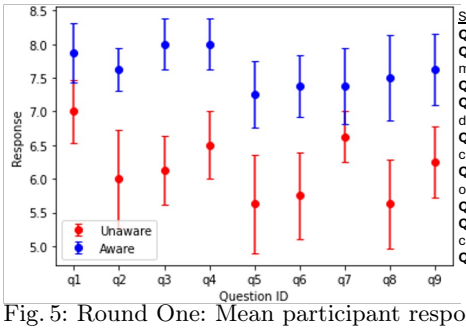
Cumulative Trust: Subjects reported 34% higher trust in the Aware classifier, as compared to Unaware (Figure 4, left). As shown on the right of Figure 4, this result is significant ($p = 0.0002$) and has a large effect size (Cohen's d = 1.93). These results support *H1*.

Over- and Under-reliance: Under-reliance occurs when the participant assists the machine even though it was capable of correctly labeling the image (a proactive human takeover). Over-reliance occurs when the participant does not assist even though the machine cannot correctly label the image. Sessions with the Aware classifier resulted in fewer proactive takeovers (4.1 vs. 9.37 takeovers) and fewer instances of over-reliance (2.37 vs 4.75 machine errors without human assistance) on average as compared to Unaware. These results were statistically significant with large effect size (Figure 4, right), and support *H2*.
Team Performance: Team performance for participants that worked with the Aware classifier (95% classification accuracy) surpassed those working with the Unaware classifier (90% classification accuracy). These results support *H3*.



	Classifier	Mean (±MSE)	Cohen's d Effect Size	p-value
Cumulative Trust	Unaware	65.5 (±4.97)	1.93	0.00
	Aware	**87.5 (±2.76)**		
Proactive Human Takeovers (Under-reliance)	Unaware	9.37 (±1.1)	1.54	0.01
	Aware	**4.1 (±.39)**		
Machine Error (Over-reliance)	Unaware	4.75 (±.64)	1.69	0.00
	Aware	**2.37 (±.26)**		
Team Performance (Accuracy)	Unaware	.90 (±.01)	1.74	0.00
	Aware	**.95 (±.005)**		

Fig. 4: Round One: Closed-loop trust calibration system with improved self-assessment results in increased human trust, reduced over- and under-reliance, and improved team performance. Improved self-awareness leads to improved trust in the machine by 34%.



Survey Questions
Q1: To what extent does the classifier perform its function properly
Q2: To what extent can the classifier's behavior be predicted from moment to moment?
Q3: To what extent can you count on the classifier to do its job?
Q4: To what extent does the classifier perform the task it was designed to do in the system?
Q5: To what extent does the classifier respond similarly to similar circumstances at different points in time?
Q6: What is your degree of faith that the classifier will be able to cope with other system states in the future?
Q7: What is your degree of trust in the classifier to respond accurately?
Q8: What is your degree of trust in the classifier's ability to assess its own capability?
Q9: What is your overall degree of trust in the classifier?

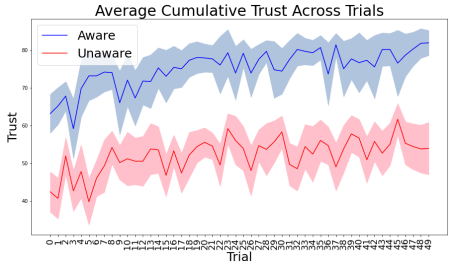
Fig. 5: Round One: Mean participant response to the post-experiment survey. Participants had an overall higher trust in and preference for the Aware classifier.

Figure 5 shows the mean participant response to questions in the post-experiment survey. The purpose of surveying participants after each session was to (1) gauge their overall trust in the machine once the task was completed, (2) validate the self-reported trust values observed during the experiment, and (3) understand how different aspects of the classifiers affected their trust. Overall, participants perceived the Aware classifier as higher performing than the Unaware classifier. Question 9, in particular, validates the results in Figure 4, indicating that participants did indeed have higher trust in the Aware classi-

fier. Question 8 supports our hypothesis that the difference in self-assessment capabilities largely drove this difference in perceived trust.

6 Experiment Results - Round Two

Round two involved using real-time prediction of trust instead of relying on compliance of human intervention, and used a neural network to determine when to ask for assistance instead of a threshold-based rule. Results of this round of experiments indicate that the first round's results hold even when the system does not rely directly on reported trust in each trial (Figure 6). As such, we are confident that our system will be applicable to more realistic situations in which the machine operates with sparse human feedback.



	Classifier	Mean	Cohen's d	p-value
Cumulative Trust	Unaware	53.87 (±6.9)	1.82	0.001
	Aware	**81.88 (±3.31)**		
Proactive Human Takeovers (Under-reliance)	Unaware	5.87 (±1.04)	1.27	0.001
	Aware	**2.75 (±.64)**		
Machine Error (Over-reliance)	Unaware	7.13 (±.61)	1.03	0.001
	Aware	**4.87 (±.89)**		
Team Performance (Accuracy) **Machine accuracy = 80%	Unaware	.85 (±.01)	1.01	0.002
	Aware	**.90 (±.02)**		

Fig. 6: Round Two: We obtain similar, statistically significant results of increased human trust (by 52%), reduced over- and under-reliance, and improved team performance.

7 Conclusion

In this work, we developed a closed-loop trust calibration system for human-machine collaboration in the image classification task that included a real-time human trust prediction model, a machine self-assessment model, and a dynamic reasoning model that determined when the machine should ask for human assistance to calibrate trust. We performed human subject experiments to highlight the importance of accurate self-assessment for trust calibration. Specifically, we showed that improved self-assessment capabilities result in increased overall trust in the autonomous image classifier, reduced over- and under-reliance behaviors on the part of the human, and improved overall team performance in the classification task. In future work, we would like to extend our experiments such that (1) we require multi-tasking on the part of the human and (2) the human may not be the expert and may also be uncertain about their ability to accomplish the task. We expect these extensions will increase the applicability of our trust calibration system to more complex, real-world scenarios.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Kumar Akash, Neera Jain, and Teruhisa Misu. Toward adaptive trust calibration for level 2 driving automation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 538–547, 2020.
2. Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018.
3. Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 215–223, 2011.
4. Charles Corbiere, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Perez. Addressing failure prediction by learning model confidence. *Advances in Neural Information Processing Systems*, 32, 2019.
5. Yosuke Fukuchi and Seiji Yamada. Selectively providing reliance calibration cues with reliance prediction. *arXiv preprint arXiv:2302.09995*, 2023.
6. Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330, 2017.
7. Martin Ingram, Reuben Moreton, Benjamin Gancz, and Frank Pollick. Calibrating trust toward an autonomous image classifier. *Technology, Mind, and Behavior*, 2021.
8. Alexander Kunze, Stephen J Summerskill, Russell Marshall, and Ashleigh J Filtness. Automation transparency: Implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics*, 62(3):345–360, 2019.
9. Joseph E Mercado, Michael A Rupp, Jessie YC Chen, Michael J Barnes, Daniel Barber, and Katelyn Procci. Intelligent agent transparency in human–agent teaming for multi-UxV management. *Human Factors*, 58(3):401–415, 2016.
10. Kazuo Okamura and Seiji Yamada. Adaptive trust calibration for human-ai collaboration. *PLOS ONE*, 15(2):e0229132, 2020.
11. Oluwatobi Olabiyi, Eric Martinson, Vijay Chintalapudi, and Rui Guo. Driver action prediction using deep (bidirectional) recurrent neural network, 2017.
12. Kiyofumi Miyoshi Tsz Yan So Sivananda Rajananda Webb, Taylor W. and Hakwan Lau. Performance-optimized neural networks as an explanatory framework for decision confidence. *bioRxiv preprint bioRxiv:2021.09.28.462081*, 2021.
13. Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2023.
14. Jessie Yang, Christopher Schemanske, and Christine Searle. Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *Human Factors*, 65(5):862–878, 2023.
15. Jessie Yang, Vaibhav V Unhelkar, Kevin Li, and Julie A Shah. Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 408–416, 2017.
16. Michelle Yeh and Christopher D Wickens. Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors*, 43(3):355–365, 2001.