AMERICAN PSYCHOLOGICAL ASSOCIATION

Technology, Mind, and Behavior

ISSN: 2689-0208

https://doi.org/10.1037/tmb0000032

Calibrating Trust Toward an Autonomous Image Classifier

Martin Ingram[1, 2], Reuben Moreton[2], Benjamin Gancz[2], and Frank Pollick[1]

[1] School of Psychology, University of Glasgow

[2] Qumodo, London

Successful adoption of autonomous systems requires appropriate trust from human users, with trust calibrated to reflect true system performance. Autonomous image classifiers are one such example and can be used in a variety of settings to independently identify the contents of image data. We investigated users' trust when collaborating with an autonomous image classifier system (AICS) that we created using the AlexNet model (Krizhevsky et al., *Advances in neural information processing systems*, 2012). Participants collaborated with the classifier during an image classification task in which the classifier provided labels that either correctly or incorrectly described the contents of images. This task was complicated by the quality of the images processed by the human-classifier team: 50% of the trials featured images that were cropped and blurred, thereby partially obscuring their contents. Across 160 single-image trials, we examined trust toward the classifier, while we also looked at how participants complied with the classifier by accepting or rejecting the labels it provided. Furthermore, we investigated whether trust toward the classifier could be improved by increasing the transparency of the classifier's interface, by displaying system confidence information (SCI) in three different ways, which were compared to a control interface without confidence information. Results showed that trust toward the classifier was primarily based on system performance, yet this also was influenced by the quality of the images and individual differences among participants. While participants typically preferred classifier interfaces that presented confidence information, it did not appear to improve participants' trust toward the classifier.

Keywords: trust, autonomous systems, image classifier, human–machine collaboration

The success of new technologies is dependent on whether they are accepted by the end user. Our understanding of how users accept new technologies has developed over time, the initial Technology Acceptance Model (TAM) put forward by Davis et al. (1989) was heavily centred on the perceived usefulness and perceived ease of use of the system, as the primary determinants for technology acceptance. More recently, extensive work by Venkatesh et al. (2012) has sought to develop upon earlier iterations of TAM by integrating further, more diverse determinants of acceptance, such as system price, the user's habits, and even the hedonistic pleasure gained from using the system (Venkatesh, 2015). This suggests that innovation alone is not enough for new technologies to be successful, and that there is a myriad of psychological, social, and environmental factors that inform the ultimate acceptance of technology.

While the successful adoption of new technologies is tied to users' acceptance of them, the users also need to learn to use the technology correctly. Just because someone accepts a new technology, it does not automatically follow that they will use it appropriately. This is particularly the case with autonomous systems, which are technologies that use Artificial Intelligence (AI) to undertake tasks with a degree of independence from their user. As these autonomous systems become more advanced, their capacity for complex tasks also increases, yet with this the opportunity for errors increases too (Parasuraman et al., 2000). As such the success of autonomous systems also relies upon appropriate trust from their human user, to ensure these systems are used correctly. Ideally, operators' trust will be calibrated to reflect the actual performance capabilities of the autonomous system, ensuring they do not distrust a functional system (too little trust), or mistrust a dysfunctional system (too much trust; Muir, 1987; Parasuraman & Riley, 1997). In this study, we sought to understand how humans calibrate their trust toward an autonomous image classifier system (AICS).

Action Editor: Danielle S. McNamara was the action editor for this article.

ORCID iDs: Martin Ingram https://orcid.org/0000-0001-8030-7085; Reuben Moreton https://orcid.org/0000-0001-6369-6031; Frank Pollick https://orcid.org/0000-0002-7212-4622.

Disclosure and Acknowledgments: This work was supported by the Economic and Social Research Council and the Scottish Graduate School of Social Science: Reference No. ES/P000681/1.

Conflict of Interest: There are no known conflicts of interest associated with this research. This research was funded by the Economic and Social Research Council (ESRC), and has been supported by Qumodo, a research and development company who are represented as authors on this manuscript.

Data Availability: An anonymized version of this data set is available through the U.K. Data Service ReShare repository here: https://dx.doi.org/10.5255/UKDA-SN-854151. The U.K. Data Service is funded by the Economic and Social Research Council (ESRC) who provided funding for this project.

Disclaimer: Interactive content is included in the online version of this article.

Contact Information: Correspondence concerning this article should be addressed to Frank Pollick, School of Psychology, University of Glasgow, 62 Hillhead Street, Glasgow, G12 8QB, United Kingdom. Email: Frank.Pollick@glasgow.ac.uk

Autonomous Image Classifier Systems

AICS are technologies that can independently classify the contents of image-based data, using advances in deep learning and convolutional neural network research (Chan et al., 2015; Howard, 2013). A major advantage of AICS is that they can process large quantities of data quickly and independently, thereby reducing demand on human users. For example, in the U.K., London's Metropolitan police force is interested in using AICS to help process digital forensic evidence, to reduce their officers' workload and limit their exposure to graphic content (Murphy, 2017). Moreover, AICS can be trained to distinguish specific, highly complicated patterns and features: An AICS was recently able to identify breast cancer with an accuracy comparable to human experts (McKinney et al., 2020). AICS can also be used in lower stakes settings, for example, the popular app "PlantNet" can provide users with classifications for images of plants and flowers that they encounter (Goëau et al., 2014). Even though these applications are impressive, the performance of AICS can reflect the expertise and potential biases of the engineers who design the systems, as well as the quality of the data set used to train their algorithms (Danks & London, 2017; Rudin, 2019). Thus, AICS are vulnerable to errors and will require appropriate trust from human operators. This is particularly important, given the potential application of AICS in a wide variety of settings, where AICS may be responsible for supporting high stakes decisions. Thus, we sought to examine how users calibrated their trust toward an AICS, and how this trust translated into compliance with the system's decisions, when completing an image classification task. By doing so, we provide an insight into trust specifically toward AICS, which we hope will benefit the design and deployment of AICS in real-world settings, while also providing further insights for the wider trust-in-automation literature.

Understanding Trust Toward Automation

Across the literature, trust-in-automation has been studied in a wide variety of human–machine teams, and arguably has most commonly been studied with autonomous vehicles (Jing et al., 2020). When considering how technology is used in different human–machine teams, Larson and DeChurch (2020) make a distinction between technology and agents. Technology is something that is used by teams to achieve their goals, much like a tool, while agents fill a distinct role within the team which goes beyond mere augmentation, and inherently improves the team's performance as a result (Larson & DeChurch, 2020). For agents, they also draw a distinction between robots, which are agents with embodied physical characteristics, and AI which are disembodied agents that perform tasks that traditionally require human intelligence, such as visual identification and decision-making. Trust has previously been studied with both robot-based agents (Desai et al., 2013; Selkowitz et al., 2017), and with AI-based agents, such as automated software repair systems (Ryan et al., 2019), virtual cognitive agents (de Visser et al., 2016; Hertz & Wiese, 2019), and decision support systems (Sauer et al., 2016; Yu et al., 2019; Zhang et al., 2020). Regarding AICS, these systems most closely align with the examples of AI-based agents. It should however be noted that within our experimental design, we afforded the AICS limited agency, as human users supervised each classification decision, with the authority to overrule each one. Whereas in real-world applications, AICS may be employed as agents with greater autonomy when working within teams.

We interpreted trust toward an AICS through the lens of Hoff and Bashir (2015) model of trust toward automation, which separates trust into three broad layers. Dispositional Trust relates to stable human-centric factors, such as culture, age, and personality traits, which inform users' general disposition toward technology. This would reflect the users' attitudes toward the AICS, and more broadly technology in general. Situational Trust relates to fluctuating human-centric factors, such as mood and attention, as well as environmental factors, such as task difficulty, workload, and organizational setting, which can all vary over time. We believe that when using an AICS, a significant factor for operators' trust would be the quality of images being processed, which could increase the difficulty of system classifications, particularly if the operator feels they could easily classify the images themselves. Finally, Learned Trust is split into two separate sublayers: Initial Learned Trust that reflects the user's historical experience of similar systems, and the reputation of the current system, while Dynamic Learned Trust reflects their ongoing experiences of working with the system. When working with an AICS, Learned Trust will likely be informed by the users' ability to interpret the system's decision-making, particularly if the image is difficult to classify. Additionally, in industrial applications, operators may have previous experiences with other AICS, which may inform their trust toward newly introduced systems. Hoff and Bashir (2015) suggest these three layers of trust combine to ultimately inform how users rely upon the autonomous systems during collaboration, which would be crucial for appropriate use of AICS. Therefore, when investigating trust toward an AICS, we created experimental manipulations that were consistent with Hoff and Bashir (2015) model and contextualized our hypotheses and subsequent findings within their theoretical framework.

System Performance

Hoff and Bashir (2015) demonstrate the complex relationship between human, mechanical, and environmental factors that combine to inform trust toward autonomous systems. However, their model stipulates that when interacting with automation, system performance is the central modulator of trust toward automation. In this vein, Yu et al. (2019) reported close relationships between perceived system accuracy, trust, and reliance upon an automated fault detection system, and demonstrated that users will modulate their trust and reliance in response to system performance. Thus, when collaborating with an AICS, we anticipated system performance, defined as the classifier's ability to correctly label the contents of images, will have the biggest influence on trust: *(H1a)* System performance, whether the classifier's label correctly describes images, will have the strongest influence on trust toward the classifier.

Image Clarity

While system performance should be the main driver of trust toward the AICS, the classifier's performance itself is likely to be dependent upon the quality of images being processed. Hoff and Bashir (2015) Situational Trust encompasses factors which make tasks more difficult to accomplish, and we believe image quality would be a particularly important factor within the context of AICS use. When images have lower clarity, through factors such as occlusion and blurring, the contents of the image may be harder for human users to identify. Moreover, when an AICS processes lower clarity images, the system's performance is also likely to be harder to evaluate, given the increased uncertainty of the contents of the images, which may itself impact upon trust toward the classifier. Thus, when working with an AICS the quality of the images processed could be considered as an environmental factor, given the operator may have limited control over image clarity. A similar issue was explored in a study by Merritt et al. (2013) involving trust toward an automated baggage scanner where trust toward the scanner was affected by the difficulty of the task. Specifically, trust was lowest in blocks where the scanner's performance was considered as "obviously poor," and highest when "obviously good," given the presence of weapons was made relatively obvious to participants. However, in the more difficult, ambiguous block, where the contents of luggage were cluttered, trust was found to be lower than the "obviously good" block, yet higher than the "obviously bad" block, illustrating the effect of task difficulty. Similar findings were reported in another study that involved an automated letter detection aid, in which participants were more likely to accept the system's advice in trials with higher difficulty (Schwark et al., 2010). This suggests that the difficulty of the task facing human–machine teams may influence how human users interpret and use automated system advice. Of course, the influence of task difficulty is likely to vary between autonomous systems, as different systems will be employed in different occupational settings, with varying consequences associated with system errors. Nonetheless, we anticipated that the relationship between system performance and trust toward the AICS would be modulated by the quality of the image being processed: *(H1b)* Image Clarity will significantly interact with system performance when predicting trust toward the classifier. With unclear trials, trust will be lower when the classifier is correct, and higher when the classifier is incorrect, illustrating participants' uncertainty about the classifier's performance.

Individual Differences

Trust toward an AICS could also be influenced by the operator's cognitive understanding of the system and task, which can be prone to biases intrinsic to each individual (Israelsen & Ahmed, 2019). Some examples of these biases include: Automation Bias, where automation performance is perceived as inherently superior to human performance (Goddard et al., 2011); and Perfect Automation Schema, where individuals may believe that automation is almost always perfectly reliable (Dzindolet et al., 2002). These biases reflect differences in trust stemming from the experiences of individual human users. Hoff and Bashir (2015) characterize biases toward trusting machines as a form of Dispositional Trust, which are relatively stable over time, and reflect users' tendencies independently of context. In order to understand how human-centric factors influenced trust toward the AICS, we considered each participant's score in the Propensity to Trust Machines Questionnaire (PTMQ; Merrit, 2011), as a form of Dispositional Trust. PTMQ scores can be used to characterize each user's predisposition toward trusting technology, in which higher scores represent higher self-reported tendencies to trust new technologies. The use of PTMQ was highlighted in the study by Merritt et al. (2013), which showed individuals with higher PTMQ scores had higher trust toward the automated baggage scanner when it processed luggage with cluttered contents, during the ambiguous performance block. This suggests that users with higher PTMQ scores were less likely to have their trust influenced by the difficulty of the task, even though the uncertainty of task success would make it harder to evaluate system performance more accurately. Thus, users' existing tendencies toward trusting machines may influence trust, even when environmental factors complicate their evaluations: *(H1c)* Participants with higher Propensity to Trust Machines scores will trust the classifier more when processing unclear images, where performance may be more difficult to evaluate.

Improving Trust Through Transparency

Trust toward autonomous systems may also be improved when system decision-making is made more transparent (Tomsett et al., 2020). For example, drivers reported greater trust toward a driving aid within an autonomous vehicle simulator when provided with explanatory feedback messages (Koo et al., 2015). The Situation awareness-based Agent Transparency (SAT) model proposes that autonomous system transparency can be improved by providing users with more detailed information that is relevant to system performance (Chen et al., 2014). Within the lens of the SAT model, human users may calibrate their trust more appropriately if the system provides more detailed information about its current task (Chen et al., 2014). Using the SAT model, Selkowitz et al. (2017) report increased trust toward an autonomous robotic squad member as it provided users with more detailed situational information, such as system motivations and predicted task outcomes. However, this trend was not apparent in the condition with the most information, implying there may be a limit to how much information is beneficial to users' trust (Selkowitz et al., 2017). Hoff and Bashir (2015) suggest that these design features which increase transparency can help users to understand the system's purpose and process when carrying out tasks, thereby improving the user's Learned Trust. Thus, we sought to understand if we could improve trust toward an AICS by making its decisions more transparent through displays of system confidence information (SCI).

SCI is a representation of system certainty when carrying out tasks and can benefit trust toward autonomous systems (Zhang et al., 2020). For example, SCI cues helped users to appropriately align their trust toward a navigational robot; lowering trust when confidence was low to accommodate poorer performance, and elevating trust when confidence was high (Desai et al., 2013). Similarly, Verame et al. (2016) report individuals were more likely to accept the decisions of an autonomous document reader when it displayed "very high" confidence, compared to when displaying "medium" or "low" confidence. This suggests SCI may improve system transparency, and in turn users' trust and strategies for collaboration. However, there are a variety of ways that SCI can be represented within the interface of autonomous systems. Previous examples include confidence discretized into high/medium/low categories, represented with icons (Desai et al., 2013) or with text (Verame et al., 2016); as numerical probabilities (9/10 = high confidence; Zhang et al., 2020); or visually through the color and opacity of icons (Selkowitz et al., 2017). Within the context of the SAT model, it is possible that more detailed forms of SCI would make AICS decision-making more transparent, and therefore be most likely to improve users' trust toward the system.

Regarding systems specifically designed to classify image-based or text-based content, Ribeiro et al. (2016) suggest SCI could be displayed through a bar graph to illustrate the probabilities of the most likely options for each decision. Arguably Ribeiro et al.'s (2016) suggestion presents SCI in a more transparent format than the previous examples above, as it provides the user with the system's confidence for the final decision relative to the confidence for other likely classification options. However, there is conflicting evidence surrounding the utility of bar graphs when conveying information, as evidence suggests they can be difficult to comprehend (Chaphalkar & Wu, 2020), and can lead to biases in readers' thinking (Godau et al., 2016). Contrarily, bar graphs have been considered useful when illustrating results with borderline differences, and reportedly require less time to interpret than raw data tables alone (Brewer et al., 2012). Therefore, we created three separate experimental interfaces that illustrated SCI in different formats and compared them against an interface without SCI (Control Interface). We adopted Ribeiro et al.'s (2016) recommendation of using a bar graph to illustrate SCI (Graphical Interface), we also displayed SCI using text-based percentages (0%–100%; Numerical Interface), and lastly used color cues to represent SCI discretized into high/medium/low categories (Iconography Interface). Thus, we explored the benefits of displaying SCI within the AICS interface: *(H2a)* Relative to the control interface, the confidence information presented within the experimental interfaces will improve overall trust toward the classifier.

We also sought to understand whether SCI would be more useful when the task difficulty increased, specifically when the classifier processes unclear images: *(H2b)* When processing unclear images, trust will be higher toward the experimental interfaces because they provide users with more information.

Lastly, we explored whether the addition of SCI in the experimental interfaces would increase users' workload, measured through subjective task load, and the amount of time participants spent in each trial: *(H2c)* When working with the experimental interfaces, participants' task load will be higher given interfaces with SCI present more information per trial.

Methods

Participants

74 participants (37 F, 36 M, 1 Non-Binary), primarily university students (Mean Age = 26.2, Min = 19, Max = 55), were recruited through the University of Glasgow's School of Psychology subject pool. All participants were compensated at a rate of £6 per hour for their time. 51% of participants considered themselves native English speakers. Ethical approval was obtained from the University of Glasgow, College of Science and Engineering ethics committee.

Design

We used a 2 × 2 × 4 within-subjects design where participants saw two levels of Classifier Performance (Correct, Incorrect) combined with two levels of Image Clarity (Clear, Unclear), within each of the four Interface-specific blocks (Control, Graphical, Iconography, Numerical). In each single-image trial ($n = 160$) the classifier's label would either correctly or incorrectly match the image displayed, which was purposely made easy or difficult to evaluate due to the clarity of the image. The ordering of blocks was randomized, as was the ordering of trials within each block ($n = 40$). The average participant took 17 s to complete each trial, and 12 min to complete each block.

Materials

Image Classifier

Participants interacted with an AICS based on the AlexNet image classifier model (Krizhevsky et al., 2012), which used MATLAB's Deep Learning and Image Processing Toolboxes (MATLAB ver. R2017a). AlexNet is a pretrained convolutional neural network, trained to classify objects within a 227 × 227-pixel net. To process each image, the file must first be resized to fit these dimensions, after which AlexNet is able to read the image. AlexNet can output a range of classifications and probabilities to illustrate its interpretation of images.

Classifier Performance

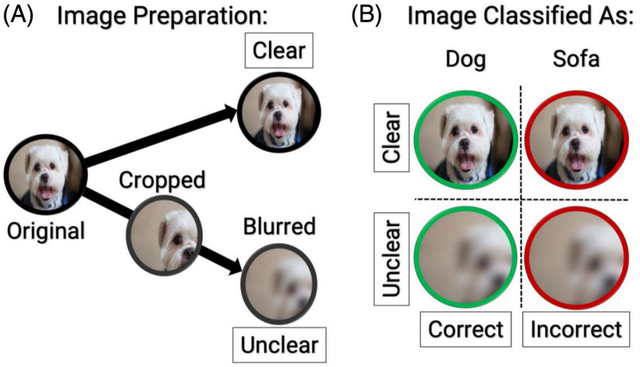
Participants viewed a series of 160 images selected from The Open Images Data set V4 (OIDV4; Kuznetsova et al., 2020). These images featured categories such as household objects, nature scenes, food items, vehicles, and animals. These were used to create four sets of 40 single-image trials, with each having 20 correct and 20 incorrect trials. The classifier's performance was considered as correct when AlexNet provided labels that appropriately matched the image's original label in OIDV4, otherwise performance was considered incorrect. Classifier performance was intrinsically linked to each image; performance only varied between images.

Image Clarity

The contents within 50% of images were made unclear to make the classifier's performance harder to evaluate. These images were first cropped, to partially show their contents, and then overlaid with a Gaussian blur when displayed to participants (See Figure 1).

Figure 1

Preparation of Images



Note. Each trial featured a single image. Classifier performance was based on AlexNet's classification for the image (B), while Image Clarity was based on the quality of the image (A). Clear trials featured images with unobscured contents, while unclear trials featured cropped images that were overlaid with a Gaussian blur when presented to participants.

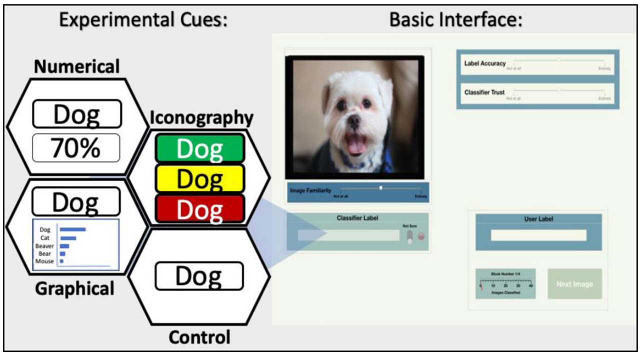
Across all 160 trials, participants saw 40 trials of each combination of Classifier Performance and Image Clarity: Correct-Clear, Correct-Unclear, Incorrect-Clear, Incorrect-Unclear, which were evenly distributed and mixed across four sets of images. These sets were organized to ensure they contained the same quantity of categories (animals, vehicles, objects, etc.), while the average classifier confidence was made similar in each set of images (Min = .5, Max = 53.6). Each set of images was randomly matched to an interface for each participant. Data associated with one image was corrupted during data collection, and therefore unusable (74 trials removed from initial 11,840 observations).

Image Classification Task

Participants used a mouse and keyboard to interact with the classifier's Graphical User Interface (GUI), built within the MATLAB app designer, (MATLAB ver. R2017a) (See Figure 2). The classifier's label for each image appeared in a box underneath the image, while participants could overwrite the classifier with their own label for each image. If participants did not understand the classifier's label, they could specify this with a small button beside the label. Additionally, if participants believed the classifier's label was wrong, yet were unable to provide a better correction themselves, they wrote "No" or "Don't Know" in their own user label box.

Participants rated the classifier's performance on a visual analog scale within the GUI, using three different interactive sliders corresponding with: (a) How familiar they were with the contents in each image, (b) How accurately they believed the classifier's label described the image, (c) Their trust toward the classifier. They were instructed that ratings of label accuracy should reflect the classifier's performance in each individual single image trial, while ratings of trust should represent their continuous interaction with the classifier throughout the experiment. All sliders went from 0%–100%, represented with visual anchor points of "Not at all" and "Entirely." Data was collected from each slider after each trial and would reset to the midpoint (50%) between trials. Each slider would change color (blue) to cue participants toward the rating they needed to provide next, guiding the participant throughout each trial. Compliance with the classifier was defined as trials where the participant did not overwrite the classifier's label. Participants moved between trials by using the "Next Image" button, which only became active after all three sliders had been used.

Figure 2
Interface Differences



Note. All four classifier GUIs contained the same basic elements. Cues of SCI were only added to the lower left-hand side of the interface, to ensure visual similarity.

Interface Design

All four interfaces contained the same basic features but varied in the SCI they displayed (See Figure 2). The Control interface provided no SCI. The Iconography interface provided the simplest form of SCI, discretized as low, medium, or high confidence, represented by the classifier's label changing color to be red, yellow, or green, respectively. The Numerical condition was more precise, presenting SCI as a text-based numerical percentage, ranging from 0%–100% representing low–high confidence. The Graphical condition was the most complex representation of SCI, illustrated as a horizontal bar graph visualizing the distribution of the classifier's five most probable labels for each image.

Questionnaires

NASA-TLX. After each task block, participants reported their subjective task load when working with each GUI, on a low–high scale (0%–100%; Hart & Staveland, 1988).

Propensity to Trust Machines Questionnaire. A series of six questions where participants rated on a 7-point Likert scale how likely they are to trust machines (Merritt et al., 2013). Half of the participants completed the PTMQ before the experiment started, and the rest after completing the experiment.

Debriefing Questionnaire. Participants answered seven short questions detailing their thoughts about the classifier (Appendix), which they completed following the last block of the experiment. They could also expand on each answer by writing a short paragraph, to explain these thoughts in further detail.

Procedure

All participants read an information sheet explaining the nature of the experiment, before giving written consent. Before the experiment began, they were taught to use the basic elements within the GUI. All participants were briefly informed how AlexNet could provide labels for each image. They were told that in certain blocks AlexNet would also display different forms of SCI, to help support its labeling decisions. They were given further specific instructions about each type of SCI prior to the relevant blocks. In each trial, the participant first rated how familiar they were with the image. The classifier then provided the label for each image, to ensure participants' familiarity was not informed by the classifier's label. Participants then rated the accuracy of the classifier's label, and their trust toward the classifier. Lastly, participants decided to keep or replace the classifier's label for the image, before moving to the next trial. Following completion of the experiment and questionnaires, all participants were given a debriefing form, which explained the study in further detail.

Analysis

ANOVA

Our data were not normally distributed, therefore we had to depart from canonical tests and instead opted for a nonparametric alternative: The Aligned Ranks Transform ANOVA (ART-ANOVA; Wobbrock et al., 2011). This test allowed for examination of multiple factors and their interactions within our repeated measures design. Our primary dependent variable of interest was: (a) participants' trust toward the classifier (Trust). In addition to this, we wanted to explore how trust reflected participants' behavior, and examined (b) how participants decided to accept/reject the classifier's labels for images (Compliance). To assess whether our stimuli selection was balanced (c) we also looked at participants' familiarity with the images presented (Familiarity). Lastly, we considered (d) the average time taken for trials in each combination of conditions, as an objective measure of task load (Trial Time). Consequently, four ART-ANOVA models were conducted, all containing the same three main factors and their interactions: Classifier Performance, Image Clarity, and Interface, using the "ARTool" package in R version 4.0.2 (Kay & Wobbrock, 2020; R Core Team, 2020). Additionally, a Kruskal–Wallis test was also conducted to examine the effect of interface on subjective task-load scores (NASA-TLX). Effect sizes were calculated for each main effect using partial eta squared. Pairwise comparisons for significant main effects were carried out using contrasts from the "emmeans" package, with Bonferroni corrections applied to account for multiple comparisons (Lenth, 2020).

Additional Analyses

Nonparametric Kendall's tau correlations were used to examine the relationships between participants' PTMQ scores and their average trust toward the classifier, as well as their average compliance with the classifier, which we compared across each combination of Classifier Performance and Image Clarity.

Visualizations

Static and interactive visualizations were created using the "ggplot2" and "plotly" R packages (Sievert, 2020; Wickham, 2016).

Data Availability

An anonymized version of this data set is available through the U.K. Data Service ReShare repository here: https://dx.doi.org/10.5255/UKDA-SN-854151. The U.K. Data Service is funded by the Economic and Social Research Council (ESRC) who provided funding for this project.

Results

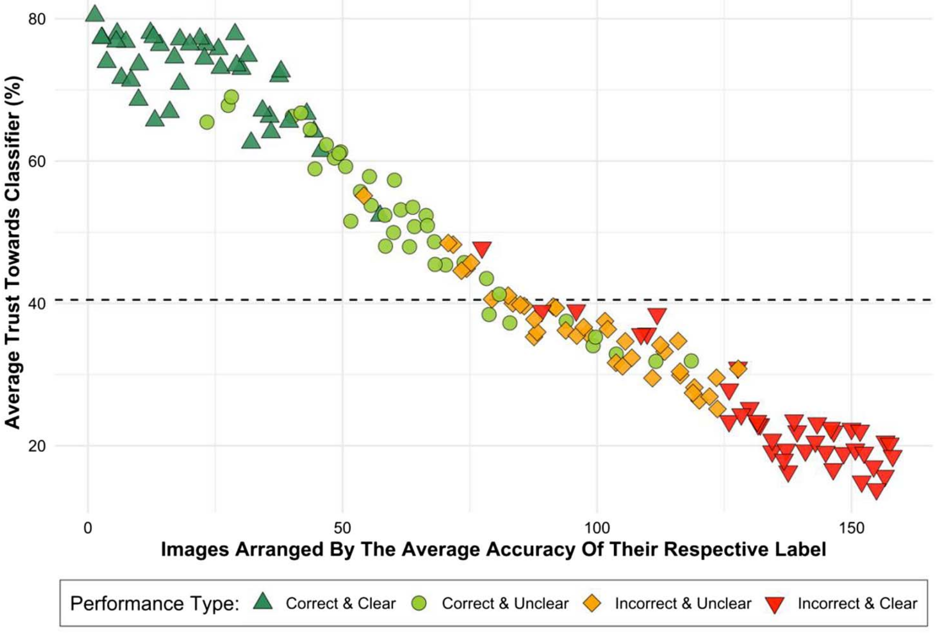
Classifier Performance and Image Clarity

Trust

Overall, trust was highest in trials where the classifier was correct, and lowest in trials where the classifier was incorrect (Figure 3, and Table 1). However, this relationship was complicated by the clarity of the image. Participants' trust tended to be closest to the grand mean ($M = 45.77$) when processing unclear images, and furthest when processing clear images. For example, if the classifier's label was correct yet the image was unclear (Correct-Unclear: $M = 51.25$, $SD = 16.02$), trust tended to be lower toward the classifier, compared to when the images were clear (Correct-Clear: $M = 72.07$,

Figure 3

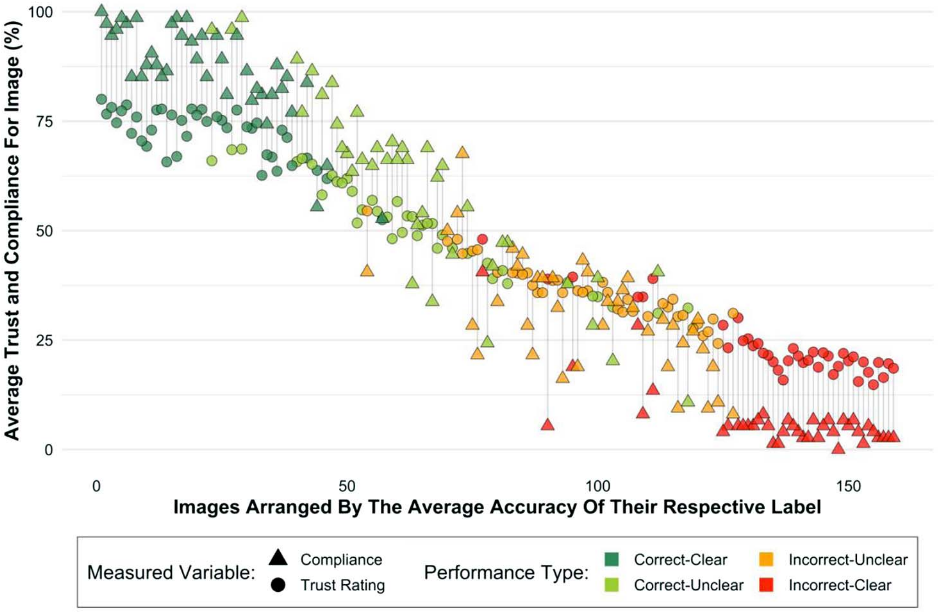
Trust Scores for Each Image Used in the Experiment, Arranged by Accuracy



Note. Stimuli arranged by participants' average accuracy rating of classifier's label for the image. Dashed line represents grand median trust.

Figure 4

Difference Between Trust and Compliance for Each Image Used in the Experiment, Arranged by Accuracy



Note. Stimuli arranged by participants' average accuracy rating of classifier's label for the image.

$SD = 22.77$). Inversely, when the classifier was incorrect trust was higher for unclear images (Incorrect-Unclear: $M = 36.12$, $SD = 15.50$), and lower for clear images (Incorrect-Clear: $M = 23.62$, $SD = 16.75$). ART-ANOVA for Trust revealed a significant interaction between Classifier Performance and Image Clarity $F(1, 73) = 205.27$, $p < .001$, $\eta p^2 = 0.74$, and significant main effects for both Classifier Performance $F(1, 73) = 226.49$, $p < .001$, $\eta p^2 = 0.76$, and Image Clarity $F(1, 73) = 24.8$, $p < .001$, $\eta p^2 = 0.25$. This supports H1a: The classifier's performance was the main driver of trust toward the classifier. This also supports H1b: Image Clarity significantly interacted with system performance when influencing trust toward the classifier.

Compliance

A similar pattern emerged when examining how participants accepted and rejected the classifier's labels (Figure 4, Table 1). ART-ANOVA for Compliance revealed a significant interaction between Classifier Performance and Image Clarity $F(1, 73) = 544.24$, $p < .001$, $\eta p^2 = 0.88$, and a main effect for Classifier Performance $F(1, 73) = 1275.09$, $p < .001$, $\eta p^2 = 0.95$. However, there was no significant main effect for Image Clarity $F(1, 73) = 1.21$, $p = .27$, $\eta p^2 = 0.02$.

Familiarity

In general, participants were more familiar with the images in the Correct-Clear and Incorrect-Clear combinations, and less familiar with the images in the Correct-Unclear and Incorrect-Unclear combinations, as we expected (Table 1). While there was no difference in familiarity between the Correct-Clear ($M = 92.89$, $SD = 8.58$) and Incorrect-Clear ($M = 92.31$, $SD = 9.15$) stimuli, there was however a difference between the stimuli in the Correct-Unclear ($M = 41.30$, $SD = 14.39$) and the Incorrect-Unclear combinations ($M = 29.89$, $SD = 12.82$). Therefore, we cannot rule out the possibility that some of the differences in Trust and Compliance were related to differences in Image Familiarity in the unclear images. ART-ANOVA for Image Familiarity revealed a significant

Table 1

Descriptive Statistics for Trust Score, Label Accuracy, Image Familiarity, Compliance, and Time as a Function of Performance Type

	Trust (%)		Label accuracy (%)		Classifier compliance (%)		Familiarity of images (%)		Time per trials (Seconds)	
Performance Type	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Correct & clear	72.07	22.77	92.19	7.13	86.89	17.77	92.89	8.58	13.10	5.25
Correct & unclear	51.25	16.02	59.30	12.35	60.20	24.13	41.30	14.39	17.66	7.02
Incorrect & unclear	36.12	15.50	34.46	13.76	31.29	25.08	29.89	12.82	20.26	8.63
Incorrect & clear	23.62	16.75	10.32	9.07	6.62	11.35	92.31	9.15	19.98	7.13

interaction between Classifier Performance and Image Clarity $F(1, 73) = 175.22$, $p < .001$, $\eta p^2 = 0.71$, and main effects for both Classifier Performance $F(1, 73) = 226.94$, $p < .001$, $\eta p^2 = 0.76$, and Image Clarity $F(1, 73) = 798.24$, $p < .001$, $\eta p^2 = 0.92$.

Propensity to Trust Machines

Trust

Participants' total scores in the PTMQ were distributed as follows: $M = 28.25$, $SD = 7.03$, Range = 11.25–39.30. PTMQ scores predicted higher trust toward the classifier in three of the four different combinations of Classifier Performance and Image Clarity (Figure 5). While these relationships are relatively weak, they suggest that individual differences may inform trust toward an AICS, particularly when processing unclear images, where system performance may be harder to evaluate. Specifically, participants with higher PTMQ scores were more likely to trust the classifier during Incorrect-Clear trials: $r_\tau = 0.09$, $p < .05$, Incorrect-Unclear trials: $r_\tau = 0.12$, $p < .01$, and during Correct-Unclear trials $r_\tau = 0.14$, $p < .001$, yet interestingly this relationship was not present during Correct-Clear trials $r_\tau = 0.06$, $p = .101$. Nonetheless, this supports H1c: Participants with higher PTMQ scores tended to trust the classifier more when processing unclear images, where performance may be more difficult to evaluate.

Compliance

PTMQ scores predicted higher compliance with the classifier in only two of the four different combinations of Classifier Performance and Image Clarity (Figure 6). Specifically, participants with higher PTMQ scores were more likely to accept the classifier's label only when the classifier was correct, during Correct-Clear trials: $r_\tau = 0.15$, $p < .001$, and Correct-Unclear trials: $r_\tau = 0.1$, $p < .05$. PTMQ scores did not predict greater compliance during Incorrect-Unclear trials: $r_\tau = 0.01$, $p = .83$, and Incorrect-Clear trials: $r_\tau = -0.02$, $p = .62$.

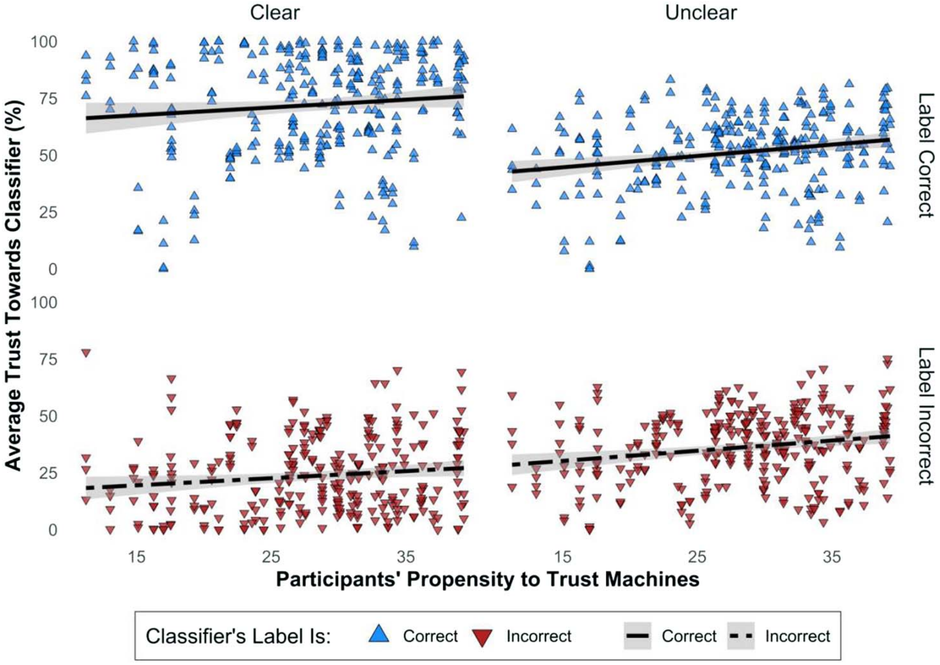
Interface Differences

Trust

Across the classifier's different interfaces, trust was highest toward the Numerical interface ($M = 47.45$, $SD = 25.27$), and lowest toward the Control interface ($M = 44.74$, $SD = 26.05$; Figure 7, Table 2). Despite this, trust toward the classifier was not significantly increased when participants worked with the experimental interfaces. While they did not improve trust, most participants reported an explicit preference for working with the interfaces that displayed SCI, suggesting they still found them beneficial on some level (Table 2). ART-ANOVA for Trust revealed no significant main effect of Interface $F(3, 219) = 1.66$, $p = .18$, $\eta p^2 = 0.02$. Thus, H2a was not supported: SCI did not improve overall trust toward the classifier. Moreover, there was no interaction between Interface and Classifier Performance, nor was there between Interface and Image Clarity. Thus, when the classifier's performance was difficult to evaluate, SCI did not improve participants' trust toward the classifier. This means that H2b was also not supported: confidence information did not improve trust when processing unclear images.

Figure 5

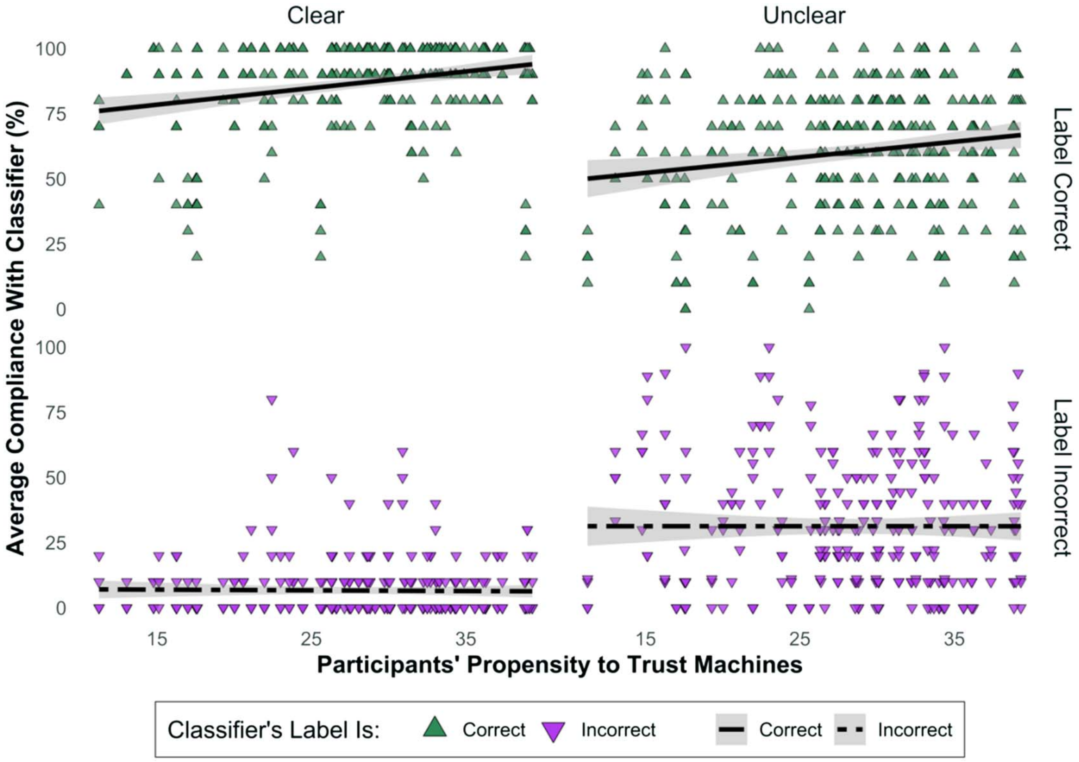
Correlations Between Participants' PTMQ Scores and Average Trust Toward the Classifier



Note. Correlations calculated for each type of performance in each block.

Figure 6

Correlations Between Participants' PTMQ and Average Compliance With the Classifier



Note. Correlations calculated for each type of performance in each block.

Compliance

Participants were most likely to accept the classifier's label when working with the Graphical interface ($M = 47.34$, $SD = 35.84$), and least likely when working with the Iconography interface ($M = 45.27$, $SD = 36.17$; Figure 7, Table 2). Despite there being no difference in trust between experimental interfaces, there were small significant differences in compliance with the classifier, suggesting some participants may have been more likely to accept the label provided by the classifier when it provided them with confidence information. ART-ANOVA for Compliance revealed a small main effect of Interface $F(3, 219) = 3.26$, $p < .05$, $\eta p^2 = 0.04$. However, pairwise comparisons suggest the differences between interfaces were nonsignificant: with the most notable being between the Graphical and Iconography interfaces ($p = .098$) and between the Graphical and Numerical Interfaces ($p = .134$). There were no interactions involving the interface factor.

Task Load

NASA-TLX

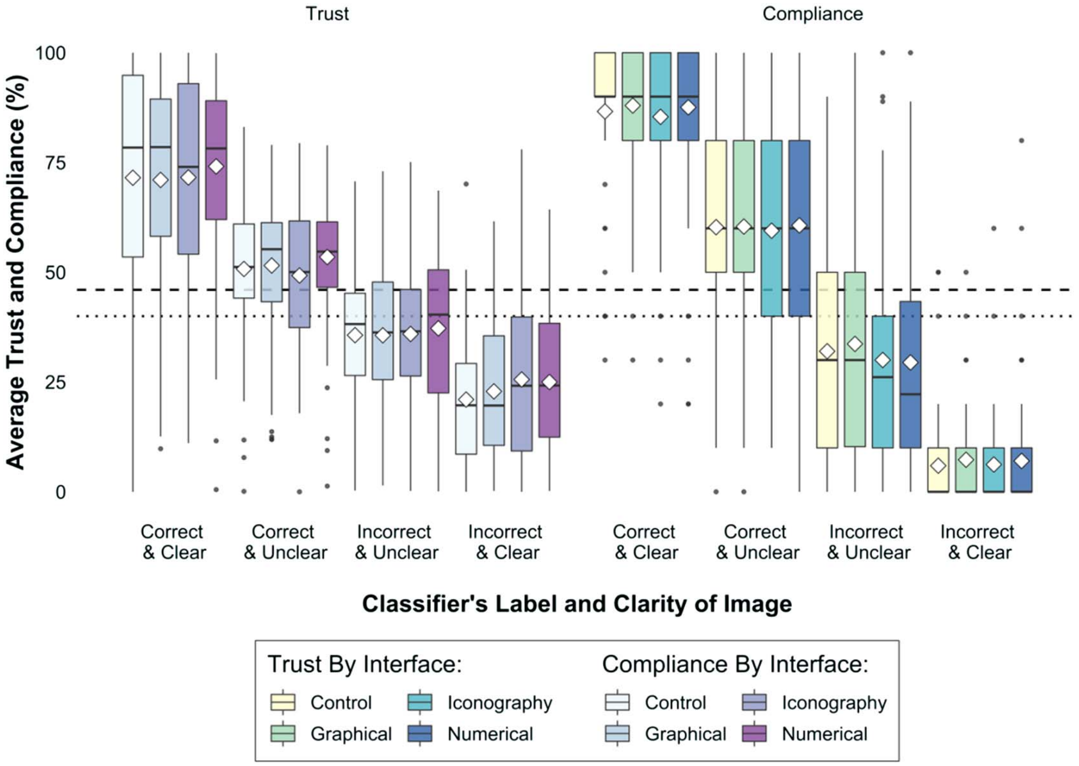
A Kruskal–Wallis test on participant's subjective task load scores revealed no differences between the experimental and control interfaces $H(3) = 0.401$, $p = .94$, (See Table 2). This suggests that the extra information presented by the classifier's experimental interfaces did not increase participants' subjective workload.

Trial Time

On average, participants spent the most time (seconds) per trial when working with the Graphical interface ($M = 19.24$, $SD = 7.21$), and the least time with the Control interface ($M = 16.43$, $SD = 6.83$). This suggests that participants did not necessarily ignore the extra information presented within the experimental interfaces, particularly when working with the Graphical interface. There were also significant differences in Trial Time relating to Classifier Performance and Image Clarity, however, these were less interesting. This was because participants were expected to take longer in trials when the classifier was incorrect, given they had to overwrite the classifier's label, and in trials with unclear images, given the classifier's performance is harder to interpret. ART-ANOVA on Trial Time revealed a main effect of Interface $F(3, 219) = 11.47$, $p < .001$, $\eta p^2 = 0.14$, which suggests that participants took longer to complete trials when presented with the classifier's SCI. Pairwise comparisons illustrated most of the significant differences were attributable to the Graphical interface, in comparison to the Control ($p < .001$), Iconography ($p < .001$), and Numerical interfaces ($p < .05$). There were also significant main effects for Classifier Performance $F(1, 73) = 206.14$, $p < .001$, $\eta p^2 = 0.74$, and Image Clarity $F(1, 73) = 41.77$, $p < .001$, $\eta p^2 = 0.36$, as well as an interaction between Classifier Performance and Image Clarity $F(1, 73) = 91.95$, $p < .001$, $\eta p^2 = 0.55$.

Figure 7

Participants' Trust and Compliance With the Classifier When Witnessing Each Performance Type With Each Interface



Note. Dashed line represents overall median trust toward the classifier, and dotted line represents median compliance with the classifier. White diamonds represent individual means for each combination. Black dots represent outliers.

Thus, H2c was not completely supported outright, as SCI presented in the experimental interfaces did not increase subjective participants' task load scores. However, there were significant differences attributable to experimental interfaces when considering the average time spent per trial as an objective measure of task load, with the Graphical interface generally being the most time-consuming.

Discussion

This study sought to understand how individuals calibrated their trust toward an AICS when completing an image classification task. Trust toward the classifier was primarily based on the accuracy of the system's description of images. Trust tended to be highest when the classifier's label was correct, and lowest when incorrect. However, the clarity of the image being processed also influenced trust, such that if the contents of the image were clear then participants were more extreme with their trust, yet with unclear images their trust regressed toward the mean. Moreover, there was also evidence of individual differences among participants. The participants with a positive bias toward machines, as indicated by higher scores on the PTMQ, tended to trust the classifier slightly more when processing unclear images. Thus, this study provides an insight into how human users place trust in a system designed to make classifications on image-based data, and expands upon this by also exploring how environmental and interpersonal factors contribute to users' trust

Table 2

Descriptive Statistics for Trust, Accuracy, Compliance, Familiarity, Time, TLX, Aesthetics, and Overall Preference as a Function of Interface

	Trust the classifier (%)		Label accuracy (%)		Classifier compliance (%)		Familiarity of images (%)		Time per trials (Seconds)		Task load (NASA-TLX)		Aesthetic rating (1–7)		Favorite interface
Interface	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*(%)*
Control	44.74	26.05	49.28	32.19	46.21	36.52	64.35	30.73	16.43	6.83	224.20	77.83	4.51	1.43	7
Graphical	45.27	25.44	49.98	31.56	47.34	35.84	63.76	31.40	19.24	7.21	230.68	78.47	4.54	1.39	50
Iconography	45.60	25.25	47.75	32.53	45.27	36.17	63.81	31.44	17.14	8.02	231.43	85.08	4.62	1.40	17
Numerical	47.45	25.27	49.27	32.77	46.18	37.16	64.49	30.59	18.18	8.23	225.57	80.12	4.53	1.48	26

toward the system. Additionally, we further built upon this by investigating whether trust toward the classifier could be improved by increasing system transparency through different displays of SCI, yet found little support with the formats we used. The implications of these findings are discussed below.

Trust Toward an AICS

In line with previous research, system performance was the primary driver of trust toward the AICS (Hoff & Bashir, 2015; Yu et al., 2019). This is unsurprising, given autonomous systems are typically designed to handle a specific set of tasks, and therefore task errors represent a violation of their fundamental purpose. However, evaluations of AICS performance seemed to extend beyond simple correct versus incorrect judgments, as trust toward the classifier varied within correct and incorrect trials. This possibly reflects the nuance in the image classification task, where the classifier must go into more detail than the simpler yes/no type judgments provided by other autonomous systems (Merritt et al., 2013; Yu et al., 2019). At the same time, we should also consider that the classification of images is a relatively familiar task that the human user can often complete by themselves. By contrast in Selkowitz et al.'s (2017) study, the robotic squad member provided users with various forms of navigational and situational data and is therefore arguably a more complicated task for the user to undertake. Undoubtedly, the greatest benefits of AICS will arise in applied settings when users are tasked with processing large quantities of data, instead of individual images. Nonetheless, our results provide an interesting insight into how individuals perceive the decisions of AICS systems. For example, when the classifier incorrectly labeled one image of a rowboat as a speedboat participants' average compliance was low, yet trust remained relatively high, despite the error (Figure 4). This illustrates how participants were able to accommodate errors when there is categorical overlap between classifications, and may itself be worth further, more rigorous investigation in future studies.

Evaluations of classifier performance were also informed by how difficult the image was to classify: If the contents of the image were clear, trust was generally higher when correct, and lower when incorrect, compared to when processing images with unclear contents. This appears in line with previous research where trust toward an autonomous baggage scanner was also influenced by the difficulty of the task (Merritt et al., 2013). By building on this, our study illustrates how task difficulty, considered as a component of Situational Trust within Hoff and Bashir (2015) model, can also influence trust toward AICS. Moreover, participants' compliance with the classifier was also informed by the difficulty of the task. Compliance was typically highest in trials where the classifier was clearly correct, and lowest in trials where it was clearly incorrect. However, this compliance was less uniform in trials with unclear images, suggesting participants were more likely to replace the classifier's labels in difficult trials. Similar to this, changes in task difficulty have been shown to influence how medical practitioners use Clinical Decision Support Systems (CDSS). Goddard et al. (2014) report that practitioners were more likely to switch decisions when working with a CDSS in scenarios requiring difficult prescriptions. While this uncertainty may appear detrimental to the operator, Lyell et al. (2018) report that using CDSS helped lower users' cognitive load when dealing with more difficult prescriptions. Therefore, when working with autonomous systems to overcome difficult tasks, the advice of the system may still be beneficial even if the system's decision is ultimately replaced or overruled by the operator. While it is worth noting that the increase in difficulty in the previous studies differs from the methods used in the current study, our findings provide further illustration of how changes in the difficulty of the task may influence how operators use autonomous systems.

Additionally, there were individual differences between participants' trust toward the classifier, which may be attributable to their PTMQ scores. Specifically, individuals with higher PTMQ scores tended to have slightly higher trust toward the classifier, particularly during trials with unclear images. Interestingly, higher PTMQ scores also correlated with higher compliance with the classifier, but only in trials where the classifier was correct. This may suggest that while the individuals with higher PTMQ scores tended to trust the classifier more, they remained critical of its performance and their positive bias did not correspond with higher acceptance of incorrect labels. Within Hoff and Bashir (2015) model, these individual differences are indicative of Dispositional Trust specific to each operator. The importance of individual differences is also illustrated within models of technology acceptance, which recognize the influence of moderating factors such as the age, gender, and experiences of the operator (Venkatesh et al., 2012). Here, we used convenience sampling in our participant recruitment, and therefore primarily focussed on PTMQ scores as a measure of individual differences. Nonetheless, this echoes previous findings where individuals with higher PTMQ scores and greater Automation Bias tended to place more trust in autonomous technologies (Goddard et al., 2014; Merritt et al., 2013). Therefore, our findings support previous literature suggesting that individual differences can influence trust and attitudes toward autonomous technology. In particular, we demonstrate that biases toward technology could make individuals more likely to trust an AICS when working with it, yet crucially these biases do not automatically translate into making the individual more likely to accept erroneous decisions from the system.

Improving Trust

Both the SAT model (Chen et al., 2014) and Hoff and Bashir (2015) model suggest that users are more likely to trust autonomous systems with more transparent interfaces. However, we found little support for SCI improving trust toward the AICS, despite previous evidence suggesting confidence information can benefit trust toward autonomous systems (Desai et al., 2013; Zhang et al., 2020). For example, there was no apparent benefit to trust during the more difficult trials with unclear images, despite SCI providing greater information about the classifier's decision. It is possible that the formats we used to convey SCI were not optimal, and that participants were unable to effectively extract the information. This possibility is consistent with previous evidence suggesting that individuals may have difficulty understanding information presented in formats such as bar graphs (Chaphalkar & Wu, 2020; Godau et al., 2016). Likewise, as discussed above, the image classification task itself may have been relatively easy for participants to complete by themselves, meaning that the classifier's decisions, and by extension SCI, may have been of limited use to participants. Additionally, any potential benefits from SCI may have been lost due to the low overall reliability of the classifier

Table 3
Descriptive Statistics for Responses to Questions 1–6 From Debriefing Questionnaire

	Response (1–7)	
Question	*M*	*SD*
How helpful did you think the classifier was? <Not at all/A great help>	4.63	0.87
How predictable was the classifier's behavior? <Predictable/Unpredictable>	4.64	1.38
How specific did you think the classifier's labels were? <Too specific/Too general>	3.44	1.12
If you had to describe it to someone, how you would characterize the classifier? <Teammate/Tool>	5.48	1.50
If you had to classify another set of images, would you want to work with the classifier again? <With classifier/Alone>	2.97	1.37
If you had to classify another set of images, which type of collaborator would you prefer? <Computer/Human>	4.54	1.51

within our experiment, which stemmed from our experimental design. Hoff and Bashir (2015) consider system reliability as a subcomponent of system performance, and while design features such as SCI can improve system transparency, any benefits to trust may be lost due to system reliability being more influential than transparency. This could be supported by participants' responses during debriefing, where they rated the classifier as more a tool than a teammate, and often found it unpredictable (Table 3). Future studies may benefit from employing high and low reliability conditions, in order to explore this further.

Despite this low reliability, participants still found the classifier helpful (Table 3). Moreover, they overwhelmingly preferred working with the classifier's SCI interfaces (Table 2) and did not appear to feel encumbered by the extra information, which still suggests SCI is potentially beneficial. Furthermore, participants spent the most time per trial with the experimental interfaces, particularly the Graphical interface (Table 2). While this does not automatically mean that SCI improved participants' comprehension of the classifier's decisions, it does suggest some processing of this confidence information. While we were primarily interested in trust toward an AICS, it would be beneficial to examine whether SCI can improve users' understanding of these systems. Alongside developing appropriate trust toward autonomous systems, there is also a growing interest in promoting the explainability of autonomous systems, particularly given the "black box" nature of contemporary machine learning approaches (Abdul et al., 2018). In future studies, it would be useful to examine whether displays of SCI can improve the explainability of AICS decisions. This may be particularly well-suited to cases when a classifier assigns the same classification to two distinctly different objects that share similar image features, such as texture and shape.

Beyond Confidence Information

Ultimately, trust toward the AICS could be limited by the way that AICS systems use deep learning techniques when learning to classify images, which can make their decision-making inherently difficult to explain (Gilpin et al., 2019). As a result, these systems may lack the explainability of other autonomous systems, which may make them fundamentally difficult to trust completely (Rudin, 2019). A recent article by Chen et al. (2019) suggested that the decisions of AICS can be made easier to interpret by highlighting important features within sections of an image, through visual cues such as bounding boxes, in order to support the classification for the full image. By doing so, Chen et al. (2019) argue that AICS can mimic the reasoning process of humans when classifying images, where the system can illustrate to the user that the classification is based upon shared features with a prototypical image of the classification, essentially: "this image looks like that image." Thus, the "black box" nature of AICS systems might mean that providing SCI alone could be inappropriate for improving trust, and instead users' trust could ultimately benefit from efforts that make AICS decisions more easily interpreted. Our lack of empirical support for SCI improving trust toward the AICS may be disappointing for potential designers, however, these findings still raise important considerations. Designing interfaces for autonomous systems is a complex process, and based on our evidence, simply providing a single indicator of system decision-making such as SCI, may not be the best way to improve users' trust, at least in the case of AICS. While displays of SCI have shown promise in previous studies, (Verame et al., 2016; Zhang et al., 2020), it may not be a "magic bullet" for improving trust toward all automation. Nonetheless, while SCI did not explicitly improve participants' trust toward the AICS, most participants still preferred interfaces that displayed SCI, which suggests it might be beneficial to some degree. Thus, this study motivates further research into developing novel methods for conveying the decision-making of systems like AICS.

Limitations

This study involved interaction with an AICS in a relatively low stakes task, where participants worked with the classifier to label neutral stimuli. Applied uses of AICS may also include higher stakes tasks, such as identifying patients with diseases (McKinney et al., 2020). In such cases, trust toward an AICS may be even more susceptible to system errors, given the more serious consequences of false alarms and missed cases. By contrast, in our experiment, there were no consequences associated with system errors. Regardless, participants still modulated their trust in response to system successes and errors, while they tended to comply with the classifier only when it was correct, suggesting they took the task seriously despite these low stakes. Future studies may wish to build upon these findings by introducing greater consequences for task errors.

Conclusion

During a human–computer image classification task, trust toward an AICS was primarily based on the classifier's ability to label images. Additionally, image clarity significantly interacted with AICS performance, and further informed participants' ratings of trust and compliance, illustrating the role of task difficulty in their evaluations. Furthermore, some of the variance in trust toward the AICS appeared to have been attributable to individual differences among participants, as those with higher propensity to trust machine scores tended to have slightly higher trust toward the classifier. Lastly, while most participants preferred interfaces that displayed SCI, it did not appear to improve their trust toward the classifier, despite previous studies suggesting confidence information can improve trust.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018, April). *Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda* [Conference session]. CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - ACM Montreal, QC, Canada. https://doi.org/10.1145/3173574.3174156

Brewer, N. T., Gilkey, M. B., Lillie, S. E., Hesse, B. W., & Sheridan, S. L. (2012). Tables or bar graphs? Presenting test results in electronic medical records. *Medical Decision Making*, *32*(4), 545–553. https://doi.org/10.1177/0272989X12441395

Chan, T. H., Jia, K., Gao, S., Lu, J., Zeng, Z., & Ma, Y. (2015). PCANet: A simple deep learning baseline for image classification? *IEEE Transactions on Image Processing*, *24*(12), 5017–5032. https://doi.org/10.1109/TIP.2015.2475625

Chaphalkar, R., & Wu, K. (2020). Students' reasoning about variability in graphs during an introductory statistics course. *International Electronic Journal of Mathematics Education*, *15*(2), Article em0580. https://doi.org/10.29333/iejme/7602

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). *This looks like that: deep learning for interpretable image recognition* [Conference session]. NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems - ACM. https://arxiv.org/abs/1806.10574

Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). *Situation awareness-based agent transparency* (No. ARL-TR-6905). Army Research Lab Aberdeen Proving Ground Md Human Research and Engineering Directorate.

Danks, D., & London, A. J. (2017, August). *Algorithmic bias in autonomous systems* [Conference session]. IJCAI: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia. https://doi.org/10.5555/3171837.3171944

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, *35*(8), 982–1003. https://doi.org/10.1287/mnsc.35.8.982

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331–349. https://doi.org/10.1037/xap0000092

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013, March). *Impact of robot failures and feedback on real-time trust* [Conference session]. Proceedings of the 8th ACM/IEEE International Conference on Human–Robot Interaction, Tokyo, Japan. https://doi.org/10.1109/HRI.2013.6483596

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79–94. https://doi.org/10.1518/0018720024494856

Gilpin, L. H., Testart, C., Fruchter, N., & Adebayo, J. (2019). *Explaining explanations to society*. arXiv preprint arXiv:1901.06560. https://arxiv.org/abs/1901.06560

Godau, C., Vogelgesang, T., & Gaschler, R. (2016). Perception of bar graphs–A biased impression? *Computers in Human Behavior*, *59*, 67–73. https://doi.org/10.1016/j.chb.2016.01.036

Goddard, K., Roudsari, A., & Wyatt, J. C. (2011). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Goddard, K., Roudsari, A., & Wyatt, J. C. (2014). Automation bias: Empirical results assessing influencing factors. *International Journal of Medical Informatics*, *83*(5), 368–375. https://doi.org/10.1016/j.ijmedinf.2014.01.001

Goëau, H., Joly, A., Yahiaoui, I., Bakić, V., Verroust-Blondet, A., Bonnet, P., Barthélémy, D., Boujemaa, N., & Molino, J. F. (2014). *Plantnet participation at lifeclef2014 plant identification task*. https://hal.archives-ouvertes.fr/halsde-01064569

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in psychology, human mental workload* (Vol. 52, pp. 139–183). Nova Science Publishers. https://doi.org/10.1016/S0166-4115(08)62386-9

Hertz, N., & Wiese, E. (2019). Good advice is beyond all price, but what if it comes from a machine? *Journal of Experimental Psychology: Applied*. Advance online publication. https://doi.org/10.1037/xap0000205

Hoff, K. A., & 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

Howard, A. G. (2013). *Some improvements on deep convolutional neural network based image classification*. arXiv preprint arXiv:1312.5402. https://arxiv.org/abs/1312.5402v1

Israelsen, B. W., & Ahmed, N. R. (2019). "Dave ... I can assure you ... that it's going to be all right ... " A definition, case for, and survey of algorithmic assurances in human- autonomy trust relationships. [CSUR]. *ACM Computing Surveys*, *51*(6), 1–37. https://doi.org/10.1145/3267338

Jing, P., Xu, G., Chen, Y., Shi, Y., & Zhan, F. (2020). The determinants behind the acceptance of autonomous vehicles: A systematic review. *Sustainability*, *12*(5), Article 1719. https://doi.org/10.3390/su12051719

Kay, M., & Wobbrock, J. (2020). *ARTool: Aligned Rank Transform for Nonparametric Factorial ANOVAs* (R package version 0.10.7). https://github.com/mjskay/ARTool

Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., & Nass, C. (2015). Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, *9*(4), 269–275. https://doi.org/10.1007/s12008-014-0227-2

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In M. I. Jordan, Y. LeCun & S. A. Solla (Eds.), *Advances in neural information processing systems* (pp. 1097–1105). Neural Information Processing Systems Foundation, Inc. https://cs.nju.edu.cn/zhangl/alexnet.pdf

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., & Ferrari, V. (2020). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, *128*, 1956–1981. https://doi.org/10.1007/s11263-020-01316-z

Larson, L., & DeChurch, L. A. (2020). Leading teams in the digital age: Four perspectives on technology and what they mean for leading teams. *The Leadership Quarterly*, *31*(1), Article 101377. https://doi.org/10.1016/j.leaqua.2019.101377

Lenth, R. (2020) *emmeans: Estimated marginal means, aka least-squares means* (R package version 1.4.8), The Comprehensive R Archive Network. https://CRAN.R-project.org/package=emmeans

Lyell, D., Magrabi, F., & Coiera, E. (2018). The effect of cognitive load and task complexity on automation bias in electronic prescribing. *Human Factors*, *60*(7), 1008–1021. https://doi.org/10.1177/0018720818781224

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vincente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., . . . Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors*, *53*(4), 356–370. https://doi.org/10.1177/0018720811411912

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors*, *55*(3), 520–534. https://doi.org/10.1177/0018720812465081

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. https://doi.org/10.1016/S0020-7373(87)80013-5

Murphy, M. (2017, December 18). *Artificial intelligence will detect child abuse images to save police from trauma*. Retrieved November 1, 2019, from https://www.telegraph.co.uk/technology/2017/12/18/artificial-intelligence-will-detect-child-abuse-images-save/

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics. Part A, Systems and Humans*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). *Why should I trust you?: Explaining the predictions of any classifier* [Conference session]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Ryan, T. J., Alarcon, G. M., Walter, C., Gamble, R., Jessup, S. A., Capiola, A., & Pfahler, M. D. (2019, July). Trust in automated software repair. In A. Moallem (Ed.), *International conference on human–computer interaction* (pp. 452–470). Springer. https://doi.org/10.1007/978-3-030-22351-9_31

Sauer, J., Chavaillaz, A., & Wastell, D. (2016). Experience of automation failures in training: Effects on trust, automation bias, complacency and performance. *Ergonomics*, *59*(6), 767–780. https://doi.org/10.1080/00140139.2015.1094577

Schwark, J., Dolgov, I., Graves, W., & Hor, D. (2010, September). The influence of perceived task difficulty and importance on automation use. In *Proceedings of the and ergonomics society annual meeting* (Vol. 54,, pp. 1503–1507). SAGE Publications. https://doi.org/10.1177/154193121005401931

Selkowitz, A. R., Larios, C. A., Lakhmani, S. G., & Chen, J. Y. (2017). Displaying information to support transparency for autonomous platforms. In T. Ziemke, K. E. Schaefer & M. Endsley (Eds.), *Advances in human factors in robots and unmanned systems* (pp. 161–173). Springer; https://doi.org/10.1007/978-3-319-41959-6_14

Sievert, C. (2020). *Interactive web-based data visualization with R, plotly, and shiny* (R Package). Chapman and Hall. https://plotly-r.com

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., & Kaplan, L. (2020). Rapid trust calibration through interpretable and uncertainty- aware AI. *Patterns*, *1*(4), Article 100049. https://doi.org/10.1016/j.patter.2020.100049

Venkatesh, V., Thong, J. Y., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, *36*(1), 157–178. https://www.jstor.org/stable/41410412

Venkatesh, V. (2015). Technology acceptance model and the unified theory of acceptance and use of technology. *Wiley Encyclopedia of Management*, *7*, 1–9. https://doi.org/10.1002/9781118785317.weom070047

Verame, J. K. M., Costanza, E., & Ramchurn, S. D. (2016, May). *The effect of displaying system confidence information on the usage of autonomous systems for non-specialist applications: A lab study* [Conference session]. Proceedings of the 2016 Chi Conference on Human Factors in Computing Systems, San Jose, California, USA. https://doi.org/10.1145/2858036.2858369

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer. https://ggplot2.tidyverse.org

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011, May). *The aligned rank transform for nonparametric factorial analyses using only anova procedures* [Conference session]. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, BC, Canada. https://doi.org/10.1145/1978942.1978963

Yu, K., Berkovsky, S., Taib, R., Zhou, J., & Chen, F. (2019, March). *Do I trust my machine teammate? An investigation from perception to decision* [Conference session]. Proceedings of the 24th International Conference on Intelligent User Interfaces, Marina del Ray, California. https://doi.org/10.1145/3301275.3302277

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020, January). *Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making* [Conference session]. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain. https://doi.org/10.1145/3351095.3372852

(*Appendix follows*)

Appendix

Debriefing Questionnaire

1. How helpful did you think the classifier was?
2. <Not at all/A Great Help>
3. How predictable was the classifier's behavior?
4. <Predictable/Unpredictable>
5. How specific did you think the classifier's labels were?
6. <Too Specific/Too General>
7. If you had to describe it to someone, how you would characterize the classifier?
8. <Teammate/Tool>
9. If you had to classify another set of images, would you want to work with the classifier again?
10. <With Classifier/Alone>
11. If you had to classify another set of images, which type of collaborator would you prefer?
12. <Computer/Human>
13. If you had to quickly classify another set of 1,000 images, which version of the interface would you prefer?
14. <Numerical, Iconography, Graphical, or Control Interface>

Received January 20, 2020
Revision received December 21, 2020
Accepted January 24, 2021 ■