

Exploring User Evaluations of Machine Learning Models: A Qualitative Study on the Impact of Confidence Intervals

Scott K. Meyers
DCS Corporation
Dayton, OH
smeyers@dcscorp.com

Paige E. Murry
DCS Corporation
Dayton, OH
pmurry@dcscorp.com

Sarah A. Jessup
Air Force Research Laboratory
Wright-Patterson AFB, USA
sarah.jessup.ctr@us.af.mil

Gene M. Alarcon
Air Force Research Laboratory
Wright-Patterson AFB, USA
gene.alarcon.1@us.af.mil

Krista N. Harris
DCS Corporation
Dayton, OH
kharris@dcscorp.com

Abstract

Research on artificial intelligence and machine learning models has burgeoned in the last decade. However, research has seldom utilized qualitative methods for assessing user-based experiences and system evaluations of AI/ML models. This study aims to provide an example of how thematic text analysis can be used to provide greater insight into user experiences with these systems and examine how varying levels of model transparency affects evaluations. Participants ($N = 130$) completed an image binning monitoring task with either an uncalibrated classification model (UCM), which displayed high confidence regardless of classification accuracy or a calibrated classification model (CCM), which had greater calibration between accuracy and confidence. Results revealed detailed information on user evaluations for both models including various performance perceptions, impressions, and strategy behaviors. Furthermore, we identified key differences in user evaluations between these models and our confidence manipulation, such as greater trust and confidence display use. Qualitative analysis has been shown to be an effective approach for detailed investigation of user experiences and model evaluation.

Keywords: machine learning, calibrated classification model, thematic analysis, qualitative data, user perceptions.

1. Introduction

Artificial intelligence (AI) and machine learning (ML) have seen rapid growth in development and application (Arrieta et al., 2020). The presence and continued advancement of these fields is apparent as the AI market is expected to reach \$407 billion by 2027, a staggering 368.2% increase from the estimated revenue in 2022 (Haan, 2023). In terms of industry application,

more than a third of companies (35%) across various industries reported AI/ML adoption according to a recent IBM report (IBM, 2022). The widespread application of these AI/ML models has led to an increase in the need for model transparency (Alarcon & Willis, 2023). Although research has explored these models with Likert-type response scales, calls for qualitative data analysis in the literature have gone relatively unanswered (Visser et al., 2023). The current study explores user perceptions of two types of AI/ML models via qualitative data.

2. Background literature

2.1. AI/ML system transparency

One common AI/ML type that provides particularly low transparency and system explanations to users (i.e., “black-box” models) are deep neural networks (DDNs) (Alarcon & Willis, 2023; Zhou et al., 2021). DDNs consist of many interconnected nodes/neurons that are organized in multiple hierarchical layers, with each layer receiving input from the previous layer. In addition to the complex organization, the inputs are also transformed across the layers through a series of mathematical functions to produce the final output (Aouichou et al., 2021; Joo et al., 2023). This results in a model that is largely opaque to the user regarding the overall system and how it makes decisions (Loyola-Gonzalez, 2019). Researchers have begun to explore transparency in AI/ML such as explainable artificial intelligence (XAI), which aims to increase system transparency and user understandability within these opaque models (Zhou et al., 2021).

AI/ML models have varying levels of transparency — a property of the system involving displaying information about the decision processes of the model (Lee & See, 2004). This information can relate to the operational logic and the processes of the system, such

as how it's designed or the relationship between its inputs and outputs (Ostmann & Dorobantu, 2021). There are differing levels of transparency requirements for effective human-machine interaction, such as the complexity of the task and the capabilities of the users involved.

2.2. Classification Models and Decision Confidence

In the context of classification models, one recent development regarding DNN transparency is the use of calibrated classification models (CCM). Uncalibrated classification models (UCMs) are standard convolutional neural networks (CNN), that classify stimuli when the model has been trained on a dataset with known classification categories and a standard loss function. In contrast, recently developed CCMs are similar CNN classifier models that have been trained on datasets with known classification categories, but with exposure to "Out of Distribution" samples (i.e., outliers) during training and have a modified loss function (Bennette et al., 2020). In other words, UCMs are trained on the images and encounter out of distribution data. In contrast, CCMs are trained with the known categories and other images to more accurately represent data that is not in a known classification.

After being trained on an initial training dataset, the model is tested on a separate test dataset to assess classification performance (Krizhevsky et al., 2012). The model classifies stimuli into one of at least two categories, along with a confidence interval indicating the model's certainty in its classification decision. UCMs typically report higher decision confidence regardless of actual classification performance (Dhamija et al., 2018), which is especially problematic when these models encounter novel stimuli and misclassify them with high confidence (Guo et al., 2017). On the other hand, CCMs provide more accurate decision certainty regarding classification decisions as they are more sensitive to stimuli external to the known categories from the training dataset (Bennette et al., 2020). This provides a unique system explanation to users regarding decision confidence and misclassifications, increasing transparency and trust outcomes (Bennette et al., 2020; Scheutz et al., 2022).

Recently, human-centered research has been conducted on CCMs and UCMs in image-based tasks (Alarcon et al. 2024a; Alarcon et al., 2024b). When users were paired with a CCM, they had better performance outcomes: higher task performance (Alarcon et al., 2024b) and lower decision time (Alarcon et al. 2024a). CCMs also corresponded to better trust outcomes: higher performance perceptions (the what of the system), higher purpose perceptions (the why of the

system), and higher reliance intentions (Alarcon et al. 2024a). Interestingly, process perceptions (the how of the system) only differed in an image classification task when participants could gain insight into how the CCM was flagging decisions as low confidence (i.e., the image was always outside of the known categories; Harris et al., in press). Finally, Alarcon et al. (2024a) noted the importance of a reference group to accurately perceive trustworthiness using inferential statistics. These results all provide evidence for the importance of the increased transparency and utility of CCMs over UCMs, but if users do not have a comparison model, inferential statistics might not capture differences across models. Therefore, examining user perceptions of CCMs and UCMs with their qualitative data could be a fruitful area of interest.

2.3. Evaluation of Machine Learning

With the increase in ML decision-making, it is imperative to evaluate the methods used to make those decisions and how users perceive the system. Human-centered perspectives contain studies that are human-in-the-loop, involving end-users or human-subject study designs (Vilone & Longo, 2020; Zhou et al., 2021). Zhou et al. (2021) split human-centered evaluations into two categories: application-grounded evaluations and human-grounded explanations. Application-grounded evaluations conduct end user experiments in real world settings to better understand how explanations assist actual users in completing tasks, whereas human-grounded evaluations are simpler experiments with lay-persons to better understand the accuracy of the ML and how well lay-persons understand the ML.

Human-centered evaluation often concerns trust, usefulness, understandability, or performance (Lopes et al., 2022) and can be further split into qualitative or quantitative studies (Vilone & Longo, 2020; Zhou et al., 2021). Vilone and Longo (2020) mention the importance of both types of studies because while quantitative studies provide results that are easily analyzed statistically, qualitative studies provide deeper insights and additional feedback.

Although the technical side of system development and evaluation has been well researched (Myllyaho et al., 2021), the human-centered perspectives with AI/ML is relatively sparse (Alarcon & Willis, 2023; Miller, 2019). Most methods developed for gaining a better understanding of ML decisions are created in computationally focused environments and do not consider a multidisciplinary approach, such as the inclusion of HCI literature (Lopes et al., 2022). The HCI literature emphasizes the importance of transparency in ML because it focuses on user understanding (Alarcon & Willis, 2023) and is a key aspect to developing trust

in machines (Lyons, 2013). Therefore, it is important to examine user considerations and the identification of model features that are important for inducing transparency in these systems. Indeed, there have been calls in the literature for more human-centered evaluation on AI/ML with qualitative studies being a useful approach in understanding how users perceive the DNNs (Visser et al., 2023).

2.3. Qualitative Analysis

Although there has been some research on UCMs and CCMs, little research has answered the call for qualitative data put forth by Visser et al. (2023). Qualitative data can provide deeper insights into the thought processes of the user (Tenny et al., 2022). Zhou and colleagues (2021) note that objective (Likert scales) and subjective (qualitative data) metrics are imperative to understanding the human-centric evaluations of AI/ML.

Qualitative analysis methods have demonstrated great utility in areas such as examining factors that influence ChatGPT adoption (Li et al., 2024), social media use (Snelson, 2016), and technology acceptance during human-robot interactions (Jessup et al., 2023). A major benefit of analyzing qualitative data is that it allows researchers to gain insight into users' perceptions, thoughts, and feelings that is afforded by open-ended questions. Unlike qualitative data, Likert-type response scales are cost- and time-effective to collect and analyze. However, participants are not able to elaborate on their responses like they are with open-ended prompts, which can reveal other constructs, themes, or factors that can influence users' decisions and perceptions (Almalki, 2016).

One of the most prominent approaches to text analysis is thematic analysis (Lochmiller, 2021). Thematic analysis is a method to identify themes and categories of information within a dataset, as well as organize the data systematically and intuitively (Braun & Clarke, 2012). Thematic analysis involves inductive (or bottom-up) reasoning (Figgou & Pavlopoulos, 2015), providing researchers the advantage of exploratory, rather than theory-driven, text mining approaches.

3. Current study

The current study utilizes qualitative analyses to explore the differences in user's perceptions, evaluations, and experiences with traditional UCM and CCM partners in an image monitoring task. We conducted a thematic analysis that utilizes text data to provide richer detail in user evaluations of ML interactions, answering a specific call in the literature.

The results of this research can inform researchers on what factors users consider important when interacting with algorithms, how users perceive information provided by two different algorithms, and strategies they adopt during repeated interactions.

4. Method

4.1. Participants

A total of 134 online participants were recruited using Amazon's Mechanical Turk (MTurk) and CloudResearch's MTurk Toolkit platform (Litman et al., 2017) as a part of a larger data collection effort (Alarcon et al., 2024b). Eligible participants had to be 18 years of age or older, reside in the United States, and vouch they were not color-blind. We also filtered for participants who had completed at least 100 human intelligence tasks (HITs) and had at least a 95% approval rate for previous HITs. Four cases were removed during data cleaning for failed training quiz and inactive task performance (i.e., did not change any images), leaving a total of 130 participants for subsequent analyses. The majority of the final sample was 56.15% male and 76.15% Caucasian. Three participants' ages were not recorded so the average age was 39.62 years ($SD = 10.03$) for the sample ($N = 127$). Participants were compensated with \$3.00 in addition to any performance bonus earned in the task. The average time to complete the task was 41.10 minutes. This study was approved by the Air Force Research Laboratory Institutional Review Board.

4.2. Task

Participants completed two training and 10 task experimental rounds of an image binning monitoring task. Within each round, participants had 15 seconds to review 15 displayed images of cats and dogs that were classified by either a CCM or UCM. Participants were asked to either approve the algorithm's classification decision or override the algorithm's decision with additional classification choices. An average of three images on each page were incorrectly classified by the model(s). Participants started with a bonus amount of \$10.00 and were penalized \$0.25 for every incorrectly classified image.

4.3. Classification models

Participants were randomly assigned to partner with traditional uncalibrated classification models (UCMs) or calibrated classification models (CCMs) in a between-subjects design. Confidence intervals of the

models were displayed with a color border surrounding the stimuli. High confidence decisions (80% or higher) were bordered in green, whereas low confidence decisions (below 80%) were bordered in yellow (see Fig. 1). This threshold was based on a conservative estimate of the cutoff for automation reliance identified in prior research (Parasuraman & Manzey, 2010). We note that although CCMs were developed for cases when the ML model has been trained on a dataset with known classification categories (e.g., dogs and cats) but encounters images outside its dataset (e.g., other animals), we only used the primary stimuli categories (i.e., cats, dogs) in our study to focus primarily on the aspect of confidence display transparency.

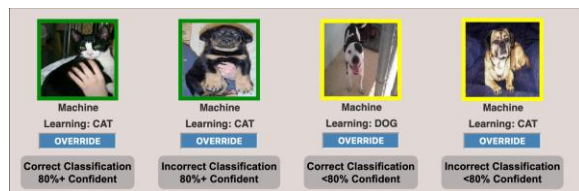


Fig. 1. Four types of images participants encountered during the task, which varied in correctness and confidence.

4.4. Data

Participant text responses to questions regarding their perceptions and evaluations of teaming with their UCM/CCM partner were used for analysis. After partnering with their respective model, we asked participants four free-response question prompts including, 1.) Please describe your interactions with the machine learning algorithm, 2.) Please describe the accuracy of the machine learning algorithm, 3.) Please describe the reliability of the machine learning algorithm, and 4.) What strategy or strategies did you use to monitor the machine learning algorithm? We used a thirty-character minimum response length requirement. We chose these questions based on relevant HCI theory (Alarcon & Willis, 2023).

4.5. Qualitative analysis process

We utilized an iterative process of coding and review – a conventional approach within qualitative analysis for codebook development (Saldaña, 2021). Within this analysis and validation, we employed six raters with academic/professional backgrounds in Human Factors Engineering and Psychology (two PhDs, two PhD candidates with Masters degree, one Masters degree, one Masters candidate with Bachelors degree). The two primary raters began by individually assessing the text responses to create the “first-level” codes. These codes served as brief summarizations of

repeating key points, purely at the descriptive level to initially assess the data. Participant responses were randomized and the model condition was blinded from the raters to reduce biases in analysis. The raters then discussed their descriptive observations across multiple sessions and consolidated their findings to create an initial codebook. This codebook comprised various individual codes (keywords and phrases) that were condensed and organized by similarity under higher-level concepts (i.e., themes). A third unbiased rater was used to evaluate the initial codebook and ensure the grouped themes of the codes and overall organization made sense and was interpretable.

Next, the two primary raters then began coding at the second level, which involved marking the occurrences of these individual codes for each response when appropriate. Raters individually assigned codes for all responses to the first question, before meeting to discuss their analyses and consolidate their results. During result consolidation, raters compared total frequency across each code and overall theme. Any discrepancies were discussed until unanimous agreement could be reached for the coding scheme. This process typically involved reviewing each individual response that was coded under a specific code, until both raters reached agreement. When agreement could not be reached, a third rater served as a tiebreaker. This process was repeated until all four question responses were completed and consolidated. After codebook validation, the three raters conducted a final review to examine and discuss overall theme/code organization.

4.6. Codebook validation

We implemented multiple validation checks to reduce biases associated with human interpretation of qualitative data and to provide a more systematic approach to analysis. The first approach involved using a third rater for codebook development. The third rater was a subject matter expert (SME) in Human Factors Psychology and had prior experience with qualitative research of AI evaluation. The SME served as an unbiased decision-maker for codebook evaluation and final judgement when agreement could not be reached for specific coding schemes between the two main raters. After codebook development, we also utilized four raters (which included the SME rater), separate from the primary two raters, to validate the final codebook and ensure inter-rater reliability amongst the higher-level themes. Each rater was assigned two question responses and the agreement between raters who analyzed the same responses was assessed. Across the questions, there was an agreement percentage greater than 75% for all major sub-themes for at least one question, indicating moderate inter-rater reliability.

The only exception to this was for the *miscellaneous and non-applicable (NA)* text responses sub-theme (discussed below).

5. Results

5.1. Thematic analysis

After coding each question, a total frequency output for each code and theme was calculated. These frequencies were then split by condition and cutoff scores were created to identify code frequency differences greater than 130% and less than 70% (i.e., 30% absolute value cutoff), using the UCM values as a reference. We decided to use this percentage benchmark, rather than a uniform cutoff value, as it provided a more conservative and standardized comparison, while taking context sensitivity and relative comparison into consideration. We found considerable condition differences using this benchmark in the following themes. Themes, sub-themes, and codes are presented in Table 1.

5.1.1. Model performance perceptions. The first theme we identified were various perceptions of UCM/CCM *model performance*. We did not find differences in overall performance perception responses between the two conditions; however, there were differences in sub-themes.

The first sub-theme identified was *perceived accuracy* with five codes (see Table 1), which represented how participants perceived the algorithm's classification accuracy during the task. The "accurate when confident" code referred to the algorithm's accuracy in assessing its own confidence regarding its decisions. For example, one participant said, "The green was correct approximately 95% of the time but the yellow was wrong 95% of the time." When examining condition differences, this code appeared within the CCM condition 430% more often than the UCM condition. We also found that those in the CCM condition had 214% more neutral evaluations of perceived accuracy compared to those in the UCM condition. All other perceived accuracy codes were similar across conditions.

The *perceived reliability* sub-theme had three associated codes and involved participants' perceived reliability of the system during the task. When examining condition differences, participants in the CCM condition reported low perceived reliability 54% less often compared to UCM condition participants. One UCM participant stated, "I don't think it's too reliable given it got a good amount of answers wrong." Neutral reliability perceptions was also different between conditions; CCM condition responses contained 183%

more descriptions of neutral perceived reliability than UCM condition responses. This finding demonstrates that those in the CCM condition had more ambiguous responses and unsure perceptions towards their algorithm regarding reliability. Note, some participants in both conditions mentioned a lack of understanding regarding the distinction between reliability and accuracy of the algorithm in this context.

The final performance sub-theme was *readiness* with three codes, which represented participant's perceptions of the algorithm's level of preparedness for completing the study task and provided insight into their impressions of the algorithm's performance as a whole. When examining condition differences for the overall sub-theme, the CCM condition had less descriptions than the UCM condition (50%). When examining specific codes, CMM's were reported as being "not ready" 200% more often and "ready" 100% less often than their UCM counterpart. In contrast, we found condition differences with the "needs training" code, where CCM condition participants reported the need for additional training 42% less often. The number of UCM respondents who felt the algorithm required additional improvement was higher than the number who felt it was ready.

5.1.2. Model information impression/evaluation. The next theme we identified involved *users' impression(s) and evaluation(s) of the information* provided during ML interactions.

The first sub-theme was *transparency/understanding of ML decisions* with two codes, which represented whether participants believed they understood the process and performance of their respective algorithm. Overall, CCM users had 59% less responses for this sub-theme than UCM users. CCM users described low perceived understanding/transparency 30% less often than their UCM counterparts. For instance, one participant stated, "I did think a few mistakes were very obvious so I questioned WHY it made its choices," indicating low levels of model understanding and transparency. All other codes were similar across conditions.

Next, we identified an *impression* sub-theme with seven codes, which encompassed all user impression and interaction descriptions such as experienced emotions, expectations, and mental schemas. Overall, UCM users had more responses under this sub-theme (68% difference) than their CCM counterparts. When examining codes, "positive/negative emotion" encompassed users explicitly describing experienced emotions or bad interactions with the models (e.g., "impressed," "enjoyed"). We found 50% less responses involving the negative emotion code by users in the CCM condition. Users described expectations for their

algorithm partner, with some responses indicating the model's performance met/exceeded expectations or unmet expectations. CCM users had less responses for both of these expectation codes (40% and 56%, respectively). Related to this, was a first impression code, which represented responses when users' initial impressions of the CCM/UCM (beforehand or during practice rounds) were different from subsequent impressions after model interaction in the official task rounds. The next code was all-or-none thinking — a factor from the perfect automation schema (PAS) — which represents a belief that automation either works perfectly or not at all (Merritt et al., 2015). Many participants with these responses indicated poor impressions of the ML partner despite satisfactory or high perceptions of its performance such as, "It is reliable to do a fair job - definitely not a perfect or near perfect job, which would be the only acceptable reliability level for me to use it." The last code was human-in-the-loop, which we characterized as user evaluations that the model requires additional human input and oversight (Parasuraman et al., 2000). For instance, one user said, "It is good for the first pass at the items, but needs [human] supervision since it makes too many mistakes." We found 54% less descriptions of this code for CCM users. No other condition differences were found.

The final sub-theme in this group was *trust* with six codes, which included all user trust perceptions, intentions, and behaviors towards the ML models. Although we did not find an overall condition difference at the sub-theme level, we did find differences at the code level. Trust and distrust codes represented general trust perceptions towards the models. CCM users had 300% more descriptions of trust and 47% less descriptions of distrust. Future-oriented trust evaluations of the models were coded as high or low trust/reliance intentions. Users interacting with the CCM had 200% more responses indicating high trust/reliance intentions. Condition differences were minimal for the latter code. These responses described whether users intended to use or trust the CCM/UCM in future contexts such as, "I wouldn't rely on it for anything important," indicating low trust intentions for future model reliance. The last set of codes indicated participant trust behaviors in regards to their model with model use or disuse in task. Model use in task was indicated by participants providing descriptions of effective teamwork with the CCM/UCM or explicitly stating they utilized the model. We found CCM users provided 193% more descriptions of model use in the task than their UCM counterparts. Model disuse in task was indicated by participants explicitly stating low or no model utilization in the study, such as by completely

ignoring the model's colored-confidence displays such as if they, "didn't pay much attention to if it was green or yellow." These latter trust behavior codes describe how users utilize their perceptions and assessments toward specific actions. Additional behaviors relevant to the context of strategy and model evaluation are expanded on in the following section. Together, these theme and codes generally represent user impression/evaluation of the models and the thinking behind their described behavior.

5.1.3. Model strategy behavior. The next theme involved the behaviors and actions of participants resulting from their perceptions and evaluations of the ML models. This primarily involved the *strategy behaviors* users implemented when completing the task.

The first strategy sub-theme we identified was *color/confidence* with four codes, which related to descriptions of participants utilizing the colored confidence border around the image stimuli. Overall, there were 176% more responses with this sub-theme from CCM users compared to UCM users. Participants in the CCM condition described each of these color/confidence codes 165%-195% more often than those in the UCM condition. For instance, one CCM user described how they, "...checked the images it was <80% confident on and corrected them as needed."

The next strategy sub-theme we identified was *physical scan/focus* with five associated codes, that described characteristics, focuses, or strategies involving a visual scanning behavior. Many of these responses involved a focus on the image/label rather than the ML-associated colored confidence and involved basic descriptions of task completion. For instance, one user wrote, "I looked at the picture and then looked at the label." We did not find condition differences at the sub-theme level, but did observe differences with specific codes. We found CCM users reported 144% more descriptions of "scan check" behaviors in their strategy, which included any broad behaviors indicating scans, double-checks, and review actions users engaged in when completing the task. Similarly, we found CCM users described focusing on every individual image to complete the task manually 53% less than UCM users. For instance, one participant stated, "...I just zipped through line by line, image by image. If more responses had been wrong, I probably would have relied on the confidence level to prioritize, but I had enough time to consider all of the options." All other codes were similar across conditions.

5.1.4. Miscellaneous. The last set of sub-themes involved *miscellaneous and non-applicable (NA)* text responses. This included responses that provided

Table 1. Thematic Analysis Results

Theme	Sub-Theme	Codes
Model Performance Perceptions	Perceived Accuracy	perceived accuracy [low/neutral*/high]; made errors/mistakes; accurate when confident*
	Perceived Reliability	perceived reliability [low*/neutral*/high]
	Readiness*	not ready*; ready*; needs training/improvement*
Model Information Impression and Evaluation	Transparency/Understanding of ML Decisions*	perceived understanding/transparency [low*/high]
	Impression*	positive emotion; negative emotion*; unmet expectations*; met/exceeded expectations*; first impression; all-or-none thinking; human-in-the-loop*
	Trust	trust*; distrust*; trust/reliance intentions [low/high*]; did use in the task*; did not use in the task
Model Strategy Behavior	Color/Confidence*	yellow*; green*; confidence level [low*/high*]
	Physical Scan/Focus	scan check*; pattern; manual check of every image*; image; label
Miscellaneous	Time	time limit; time management
	Adaption*	learning*; strategy adjustment; bonus*
	Miscellaneous*	n/a*

Note. * denotes condition differences in user evaluations of UCMs and CCMs utilizing the 30% difference threshold

additional contextual details regarding the study task or strategy, or were unrelated/unintelligible/bot responses.

We first identified the overall sub-theme of *time* with two codes. We did not find any condition differences for this sub-theme or associated codes. The time codes involved any descriptions of the time limit, experience of time pressure, or strategies for time pressure when completing the task, such as users referencing the “timer” or the “short amount of time available.”

Next, we identified an *adaption* sub-theme with three associated codes related to strategy adjustments in response to the task. There were 180% more descriptions of this sub-theme in CCM users. Additionally, CCM users described “learning” or observing and figuring out aspects of the task and model, 267% more often than UCM users. Similarly, there were 400% more responses involving mentions of the task “bonus” in the CCM condition. As performance and bonus amount feedback were provided to participants after each round, some users described utilizing that information to influence their process or strategy such as, “I was burning through my bonus until I decided to just use the machine learning algorithm and just focus on the images with the yellow border.”

Finally, we had a separate sub-theme for *miscellaneous and non-applicable (NA)* text responses. This primarily included “bot” responses consisting of standard definitions unrelated to the question prompts,

unintelligible responses, and responses with ambiguous meanings, in which a clear theme could not be extracted from or agreed upon by the raters. There was a smaller presence of miscellaneous responses in CCM users (53%), which we likely attribute to random error.

6. Discussion

The current study explored user perceptions of UCMs and CCMs in an image binning task. In general, there were more positive performance perceptions ascribed to CCMs. CCMs had greater descriptions of being accurate when displaying decision confidence, indicating participants noticed the additional calibration. Furthermore, CCMs had fewer descriptions of low perceived reliability and a need for additional training compared to UCMs. Interestingly, we did not find condition differences in the extreme accuracy perception codes (high/low) or reliability code (high), which was surprising as lower performance perceptions towards UCMs was expected due to its high-confident misclassifications. However, we attribute these mixed results to user uncertainty and a lack of a model comparison of each novel algorithm, leading to similar levels of extreme opinions regarding model performance between conditions. CCMs did have a higher prevalence of “neutral” perceptions of model accuracy and reliability performance, potentially

indicating increased uncertainty associated with this algorithm. Overall, we determined CCMs were generally associated with greater performance perceptions, although differences in the various sub-themes and codes should be explored further.

UCM users generally had more negative impressions and evaluations of their model including negative associated emotions, unmet expectations, and a desire for human input. Furthermore, participants associated UCMs with more descriptions of low transparency/understanding and higher levels of general distrust, which was expected due to the associated misalignment between misclassifications and confidence. In contrast, users had higher trust in CCMs including general perceptions, intentions, and actual behaviors. This key finding indicates the utility of CCMs over more traditional models, particularly within the user trust/reliance domain. Interestingly, we did not find condition differences for all-or-none thinking, indicating users had similar cognitive beliefs regarding a low tolerance for mistakes for both model systems (Merrit et al., 2015).

The condition differences across all codes and the overall theme of color/confidence indicated a greater use of the algorithm for the strategy of CCM users. As color was an inherent part of the ML display in this study, the more frequent descriptions of “yellow” and “low confidence,” demonstrated utilization of the CCM’s low confidence decision cases. The transparency in the CCM misclassifications were both noticed and utilized by users, as they diverted their focus to model errors in response to the colors, whereas previous research with poor display design did not find any impact of confidence intervals (Ling et al., 2024). In contrast more UCM users described utilizing the specific “manual check of every image” strategy, which indicated these participants did not focus on specific aspects of the display or rely on the algorithm. This manual scanning behavior of each stimuli indicated little to no utilization of the UCM algorithm confidence intervals during task completion. Interestingly, we found that CCM users had more descriptions of the general “scan check” behavior code. This could be attributed to a two-fold effect, where CCM users described this behavior when specifically scanning for color/confidence, which had a greater presence in this condition, as previously mentioned. Although it should be noted both color focus and manual check strategies involved this general scan/check behavior. This is unsurprising given that these codes involved the general nature of the study task rather than model-specific interaction behaviors. The results of this study also demonstrate the importance of display design as the results may have been vastly different if we had chosen

a more cognitively demanding display of the confidence intervals (Ling et al., 2024).

6.1. Practical and theoretical implications

The current study provides a case example of utilizing qualitative thematic analysis methods to provide additional insight from user’s text responses. In fact, the results of this study align with prior CCM research regarding increased trust perceptions and a stronger presence of positive attitudes and intentions in CCM users (Alarcon et al., 2024a). However, past studies using self-report scales are limited to the questionnaires, and have not yet explored deeper into user trustworthiness perceptions such as the mechanisms behind them and the resulting evaluations and behaviors. Additionally, the nature of post-hoc theme development allows CCM researchers to capture other perceptions such as model readiness and transparency. Previous research also lacked information of user’s assessment of information (Alarcon et al., 2024b; Meyers et al., 2024), which we were able to ascertain in this study such as user’s emotional reactions, unmet expectations, and perfect automation schema.

6.2. Limitations and future research

There were various limitations to this approach of measuring user’s ML evaluations. First, our “manual” approach to text data is both labor and time intensive, involving a continuous process of coding, consolidation, and review for the primary raters. This approach relies solely on human-raters, each having their own set of biases during interpretation and analysis. The various steps we outlined in our method section attempts to address these considerations by utilizing constant analysis review and standardization. Second, the presence of code and theme inflation during this more granular-level approach was a limitation. As codes were not mutually exclusive, one individual response could be assigned to multiple codes, which could be organized under the same sub-theme. Thus, there were cases where an individual response could account for inflated frequencies subject to further accentuation when collapsing response results across the multiple question prompts.

Alternative approaches can also be utilized to assess the qualitative data. Although it is not in the scope of the current study, performing sentiment analysis, and word counts among other quantitative assessments of qualitative data could be useful in assessing perceptions of AI/ML (Jessup et al., 2020; Jessup et al., 2023). However, these methods may not provide the richness of data as the current analyses. Nevertheless, the rapid

advancement of text mining and natural language processing models calls for further exploration in the use of AI and quantitative methods of ML evaluation text data in the future (Cambria & White, 2014). These alternative methods can automate the effort-intensive approach shown in this study, reduce rater biases, and utilize key information that may not be salient for human raters. These models also address issues with more basic text analytic techniques (e.g., word counts), by modeling the contextual connections between words. There have been recent use cases of these models including employee selection (Campion & Campion, 2020; Thompson et al., 2023) and chatbots (Adamopoulou & Moussiades, 2020). These examples evidence not only the current success but future capabilities of these systems. A more in-depth and controlled comparison of traditional thematic analysis, sentiment analysis, and extensions with more advanced LLMs would be useful in extending the literature regarding how we should approach and design research in this domain.

Similarly, rather than utilizing classification models as in this study, expanding the evaluation to these LLMs and more complex ML models, could lead to more rich and dynamic analyses. The unique advantages of qualitative methods, such as providing rich detail, could be better showcased through LLM evaluation, as the task and associated model outputs would be more complex and multifaceted.

7. Conclusion

With the advancement of AI/ML development and adoption, utilizing comprehensive approaches for system evaluation will be increasingly important. In this study, we used qualitative methods to capture new aspects of user experiences with UCM/CCM from text response data. With this approach, we were able to identify more detailed information about user perceptions, impressions, and behaviors in regards to these models and our confidence display manipulation. Furthermore, we were able to parse out significant condition differences within these user evaluations. Major findings included higher trust perceptions, intentions, and behaviors from users interacting with CCMs. Furthermore, we found CCM users reported higher utilization of the color display and consequently the model itself in their task strategy. With this study, we aim to showcase the benefits of traditional thematic analysis and answer the call for more exhaustive detail of user interaction experiences in this research domain. We hope the exploratory evidence in this case example with image classification models will encourage greater use of qualitative and mixed-method approaches for human-machine interaction and system evaluation.

8. Acknowledgements

Distribution A. Approved for public release: RH Cleared 22 Jul 2024; RH-24-125825. The views expressed are those of the authors and do not reflect the official guidance or position of the United States Government, the Department of Defense, or the United States Air Force. The work was supported by the Air Force Office of Scientific Research under Grant # 22RICOR001.

9. References

- Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2, Article 100006.
- Alarcon, G., & Willis, S. (2023). Explaining explainable artificial intelligence: An integrative model of objective and subjective influences on XAI. *Proceedings of the Hawaii International Conference on System Sciences*, 56, 1095–1104.
- Alarcon, G. M., Harris, K. N., Jessup, S. A., Meyers, S. K., Ryan, T. J., Willis, S. M., Johnson, D., & Bennette, W. (2024a). Trust in machine learning: Comparing calibrated and uncalibrated image classification models. [Manuscript in preparation].
- Alarcon, G. M., Jessup, S. A., Meyers, S. K., Johnson, D., & Bennette, W. D. (2024b). Trustworthiness perceptions of machine learning algorithms: the influence of confidence intervals. *Proceedings of the IEEE International Conference on Human-Machine Systems*, 4.
- Almalki, S. (2016). Integrating quantitative and qualitative data in mixed methods research—challenges and benefits. *Journal of Education and Learning*, 5(3), 288–296.
- Aouichaoui, A. R., Al, R., Abildskov, J., & Sin, G. (2021). Comparison of group-contribution and machine learning-based property prediction models with uncertainty quantification. *Computer Aided Chemical Engineering*, 50, 755–760.
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.
- Bennette, W., Maurer, K., & Sisti, S. (2020). Harnessing adversarial distances to discover high-confidence errors. *Proceedings of the IEEE International Joint Conference on Neural Networks*, 1–10.
- Braun, V., & Clarke, V. (2012). Thematic analysis. *APA handbook of research methods in psychology: Research designs: Quantitative, qualitative, neuropsychological, and biological* (Vol. 2, pp. 57–71). American Psychological Association.
- Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9(2), 48–57.

- Campion, E. D., & Campion, M. A. (2020). Using computer-assisted text analysis (CATA) to inform employment decisions: Approaches, software, and findings. *Research in Personnel and Human Resources Management* (Vol. 38, pp. 285–325). Emerald Publishing Limited.
- DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Sage Publications.
- Dhamija, A. R., Günther, M., & Boulton, T. (2018). Reducing network agnostophobia. *Proceedings of Advances in Neural Information Processing Systems*, 9157–9168.
- Figgou, L., & Pavlopoulos, V. (2015). Social psychology: Research methods. *International Encyclopedia of the Social & Behavioral Sciences* (Vol. 2, pp. 544–552). Elsevier.
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the International Conference on Machine Learning*, 1321–1330.
- Haan, K. (2023). 24 Top AI statistics and trends in 2024. *Forbes*. https://www.forbes.com/advisor/business/ai-statistics/#sources_section
- Harris, K. N., Capiola A., Johnson, D., Alarcon, G. M., Jessup, S. A., Willis, S., & Bennette, W. (in press). Investigating the Effects of Classification Model Error Type on Trust-relevant Criteria in a Human-Machine Learning Interaction Task. Paper accepted for presentation at the *Hawaii International Conference on System Sciences*, USA.
- IBM. (2022). *IBM Global AI Adoption Index 2022*. IBM. <https://www.ibm.com/downloads/cas/GVAGA3JP>
- Jessup, S. A., Alarcon, G. M., Capiola, A., & Ryan, T. J. (2020). Multi-method approach measuring trust, distrust, and suspicion in information technology. *Proceedings of International conference on human-computer interaction*, 412–426. Springer International Publishing.
- Jessup, S. A., Willis, S. M., & Alarcon, G. M. (2023). Extending the Affective Technology Acceptance Model to human-robot interactions: A multi-method perspective. *Proceedings of the Hawaii International Conference on System Sciences*, 56, 491–500.
- Joo, C., Kwon, H., Kim, J., Cho, H., & Lee, J. (2023). Machine-learning-based optimization of operating conditions of naphtha cracking furnace to maximize plant profit. *Computer Aided Chemical Engineering* (Vol. 52, pp. 1397–1402). Elsevier.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1–9.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80.
- Ling, S., Zhang, Y., & Du, N. (2024). More is not always better: Impacts of AI-generated confidence and explanations in human–automation interaction. *Human Factors*, Advance online publication.
- Li, Y., Zhao, Y., Min, D.-J., & Payne, D. (2024). A qualitative inquiry into the adoption of ChatGPT in the early stages. *Proceedings of Hawaii International Conference of System Sciences*, 57, 4942–4951.
- Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 49, 433–442.
- Lochmiller, C. R. (2021). Conducting thematic analysis with qualitative data. *The Qualitative Report*, 26(6), 2029–2044.
- Loyola-Gonzalez, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7, 154096–154113.
- Merritt, S. M., Unnerstall, J. L., Lee, D., & Huber, K. (2015). Measuring individual differences in the perfect automation schema. *Human Factors*, 57(5), 740–753.
- Meyers, S., Capiola, A., Alarcon, G. M., & Bennette, W. (2024). Transparency and trustworthiness: Exploring human-machine interaction in an image classification task. *Proceedings of the IEEE International Conference on Human-Machine Systems*, 4.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.
- Myllyaho, L., Raatikainen, M., Männistö, T., Mikkonen, T., & Nurminen, J. K. (2021). Systematic literature review of validation methods for AI systems. *Journal of Systems and Software*, 181, Article 111050.
- Ostmann, F., & Dorobantu, C. (2021). *AI in financial services*. The Alan Turing Institute.
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381–410.
- 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.
- Saldaña, J. (2021). *The coding manual for qualitative researchers* (4th ed.). Sage Publications.
- Scheutz, M., Thielstrom, R., & Abrams, M. (2022). Transparency through explanations and justifications in human-robot task-based communications. *International Journal of Human-Computer Interaction*, 38(18–20), 1739–1752.
- Snelson, C. L. (2016). Qualitative and mixed methods social media research: A review of the literature. *International Journal of Qualitative Methods*, 15(1), 1–15.
- Tenny, S., Brannan, J. M., & Brannan, G. D. (2022). *Qualitative study*. StatPearls Publishing.
- Thompson, I., Koenig, N., Mracek, D. L., & Tonidandel, S. (2023). Deep learning in employee selection: Evaluation of algorithms to automate the scoring of open-ended assessments. *Journal of Business and Psychology*, 38(3), 509–527.
- Visser, R.W., Peters, T.M., Scharlau, I., & Hammer, B. (2023). *Trust, distrust, and appropriate reliance in (X)AI: A survey of empirical evaluation of user trust*. ArXiv. <https://arxiv.org/abs/2312.02034>
- Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5), Article 593.