The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems

Mahsan Nourani[1], **Joanie T. King**[2], **Eric D. Ragan**[1]
[1] University of Florida, Gainesville, Florida
[2] Texas A&M University, College Station, Texas
mahsannourani@ufl.edu, joanie_king@tamu.edu, eragan@ufl.edu

Abstract

Domain-specific intelligent systems are meant to help system users in their decision-making process. Many systems aim to simultaneously support different users with varying levels of domain expertise, but prior domain knowledge can affect user trust and confidence in detecting system errors. While it is also known that user trust can be influenced by first impressions with intelligent systems, our research explores the relationship between ordering bias and domain expertise when encountering errors in intelligent systems. In this paper, we present a controlled user study to explore the role of domain knowledge in establishing trust and susceptibility to the influence of first impressions on user trust. Participants reviewed an explainable image classifier with a constant accuracy and two different orders of observing system errors (observing errors in the beginning of usage vs. in the end). Our findings indicate that encountering errors early-on can cause negative first impressions for domain experts, negatively impacting their trust over the course of interactions. However, encountering correct outputs early helps more knowledgable users to dynamically adjust their trust based on their observations of system performance. In contrast, novice users suffer from over-reliance due to their lack of proper knowledge to detect errors.

1 Introduction

System designers and practitioners incorporate machine learning and artificial intelligence (ML/AI) models to help end-users achieve their goals and make decisions. Intelligent systems are used across a wide variety of domains, such as medical diagnosis assistance (Goyal et al. 2018; Bussone, Stumpf, and O'Sullivan 2015), cybersecurity monitoring (Goyal and Sharma 2019), and criminal justice (Rudin and Ustun 2018; Berk and Hyatt 2015). The intended end users of such systems often possess different levels of background domain knowledge. For instance, medical decision support systems incorporate AI/ML approaches to help with automated diagnoses for diseases. While doctors and medical practitioners can use these systems to make a diagnosis or verify it, patients may use similar systems to input their symptoms for an early diagnosis.

Previous research has demonstrated that domain expertise and user-reliance on intelligent systems are related. For example, Bussone et al. (2015) have shown that little or no domain knowledge can cause over-reliance on the system and automation. It has also been found that pre-existing knowledge of an automated system can influence user's initial trust on the system (Hoff and Bashir 2015). We can conclude that domain experience plays an important role on when users decide to trust a system and when not. To encourage a more trustworthy intelligent system, researchers suggest incorporating explanatory techniques to improve transparency and help users understand how the model is making its predictions. (Hoff and Bashir 2015; Doshi-Velez and Kim 2017).

Though improving model transparency is viewed as a partial solution to the trust problem, users normally develop a sense of a system's accuracy through their own experiences and observations using the system over time. Thus, when beginning to use the system, and before fully developing an appropriate mental model, early impressions of the system can affect how users perceive the system's accuracy (Nourani et al. 2020a). Since there is almost no AI/ML algorithm with 100% accuracy in meaningful real-world contexts, we expect all intelligent systems will eventually make errors—but *when* users observe the errors is crucial.

As users interact more with an automated system over time, their trust evolves based on user observations and experiences (Hoffman et al. 2013). We also might expect a different initial trust from users with domain experience while their ability in detecting errors can also affect how their trust changes—as also suggested by (Merritt et al. 2015).

Motivated by these challenges, we designed an experiment to explore how users with different levels of domain expertise develop trust over time as the observed accuracy changes. Incorporating a simulated image classification task in the entomology domain, we recruited domain-experienced and novice participants in an online study. To study the relationship between domain knowledge and first impressions, we defined two extreme scenarios to control when users experience system errors: 1) in the beginning and 2) at the end of the usage. Each user reviewed *the same* set of arthropod images—with their associated labels and

explanations—while the order of the set was determined by one of the two assigned scenarios in a between-subjects setting. We measured and compared trust and its calibration over time for the novice and experienced participants based on their initial impressions of the system, as well as their perception of the overall system accuracy. Our results provide novel and significant findings of the importance of domain knowledge in the formation of first impressions and experience with the system over time.

2 Related Work

Social and psychological researchers have been studying human trust for many years. Although there is not one agreed upon definition of trust in this area, human-human trust is commonly based on believing that the trustee will do what is expected (Good 2000). Similarly, trust in automation is a user's ability to rely on and predict the results from the automated system. Similar to human-human trust, once human-machine trust is lost, it is hard to reestablish it (Hoffman et al. 2013). However, research has shown that humans are more forgiving towards humans than machines when their invested trust is violated (de Visser et al. 2012), which highlights the importance of maintaining trust in automation.

Researchers have looked into different modes of trust in automated and intelligent systems. For instance, Merrit et al. (2015) examined trust calibrations (users' adaptation of trust over time) and its outcomes on task-performance and error detection. One of the major design decisions that is mutually accepted in the research community is model transparency through explanatory systems and intelligent user interfaces (Hoffman, Klein, and Mueller 2018). Papenmier et al. (2019) studied the interplay between model accuracy and explanation fidelity, and how they affect user trust in intelligent systems. Their results show that model accuracy plays a more important role on user trust than explainability. They also found that users cannot be tricked into trusting a bad classifier when the system provides high fidelity explanations. Our work is similar to their work in that we explore user trust through a low accuracy classifier system with high fidelity explanations. However, we are looking into whether a user's domain knowledge can affect their perception of the system accuracy. In other relevant work, Yu et al. (2017) studied changes of trust over time based on different levels of model accuracy through a decision-support system, targeting novice users. They found that with lower overall accuracy, trust tends to decrease over time. Our research also studies the changes of trust over time, while we focus on how first impressions can affect this change, specifically with domain experience.

Different factors (prior and during interactions) can affect user trust and reliance on intelligent systems (Hoff and Bashir 2015). These factors can make it more challenging to design intelligent systems and agents. One known challenge is when users tend to over-trust and over-rely on the system predictions, heavily depending their decisions on the system outputs (also known as *automation bias*) (Mosier and Skitka 1999; Alberdi et al. 2005). Previous research demonstrate that novice users can suffer from this problem (Mosier and Skitka 1999; Bussone, Stumpf, and O'Sullivan 2015). On the contrary, mistrust and distrust can cause users to underestimate the system and rely on themselves, eventually causing them to stop using the automated system in the future. For example, getting people to provide feedback to an intelligent system to fix errors can amplify user mistrust in the system (Honeycutt, Nourani, and Ragan 2020). More related to this paper, Nourani et al. (2020b) found that after observing system's weakly-justified predictions, users tend to disagree with the system even when it is right. In this paper, we explored the interplay of domain expertise and observed performance on user reliance behaviours.

A number of researchers in human-computer interaction have researched how domain expertise can affect user behaviours with intelligent systems. For example, Zhang et al. (2020) studied how users' trust calibration can be affected by knowing that their domain knowledge is higher than the model's. Although they raise an important question, their results were inconclusive. The resulting studies have been focused on different domains, such as medical (Bussone, Stumpf, and O'Sullivan 2015; Vaidyanathan et al. 2014; Cai et al. 2019), data science (Kaur et al. 2020), visual analytics (Dasgupta et al. 2016), and aviation (Mosier and Skitka 1999). In this paper, unlike the previous work, we study how domain expertise can affect impression formation and trust calibration. It is important to bear in mind that novice and expert terminology is task and domain dependent and might vary from one system to another. For example, some researchers define novice users as students or those who have a ground-level of knowledge in the domain (Bussone, Stumpf, and O'Sullivan 2015; Mosier and Skitka 1999), while many times terms as novice or lay users are referring to the general public with close to zero knowledge in the domain(Doshi-Velez and Kim 2017).

Trust is subjective by nature and therefore challenging to quantify. Several methods have been suggested to measure trust in intelligent systems (Mohseni, Zarei, and Ragan 2018). There are many suggested *trust in automation* questionnaires such as (Hoffman et al. 2018) that can be used to measure trust explicitly. Some of the implicit trust measurements include checking user agreement with wrong system outputs (Papenmeier, Englebienne, and Seifert 2019; Nourani et al. 2020b); repeatedly asking for trust ratings (Yu et al. 2017); and measuring user perception of system accuracy as an indication of user trust (Yin, Wortman Vaughan, and Wallach 2019; Nourani et al. 2019). In this paper, we measure trust through both implicit and explicit measures.

In this paper, we study first impression formation based on domain expertise. Previous research on first impressions has shown that human's early observations and judgements can bias and affect their behaviours towards people (Zebrowitz 2017; Fourakis and Cone 2020), systems (Nourani et al. 2020a), and/or agents (Petrak et al. 2019; Desai et al. 2013). However, to the best of our knowledge, there has been little focus on how the ordering of user experiences with different model outputs can affect user trust. Without focusing on the order, Dietvorst et al. (2015) found that users tend to avoid an algorithm and favor humans over systems once they observed the system made an error. In our own previous work, we found evidence that positive and negative first im-

pressions can affect user reliance and mental models of an intelligent system. However, the prior study did not account for relationships including domain expertise and changes in trust over time (2020a), which serves as the focus for our new study of first impressions in the current paper.

3 Experiment

We conducted a user study with a simulated multi-class image classification scenario to understand how domain knowledge and order of observing system errors can affect user trust. In this section, we discuss the study design, goals, and participants in more detail.

3.1 Research Goals and Hypotheses

The primary motivation of this study was to understand how first impressions of an intelligent system can affect user trust, and whether and how domain expertise can help bypass the influences of these first impressions. We focused on systems with local outputs and explanations, i.e., systems that show one output at a time to their users – e.g., (Ribeiro, Singh, and Guestrin 2016) – rather than intelligent systems at a global scope where users see a representation of how the model works on a high level – e.g., (Hohman et al. 2019). Considering these systems, our goal was to understand how user domain expertise affects the formation of first impressions, changes of trust over time, and estimation of system accuracy. To address this research question, we summarized the following set of hypotheses:

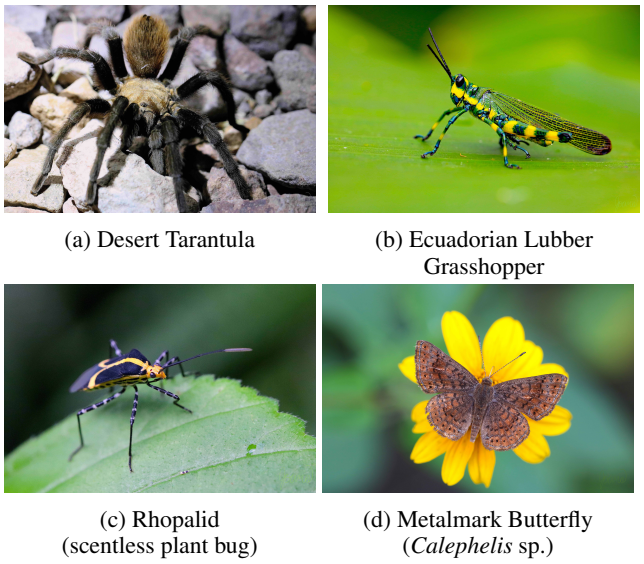
- **H1**: Ordering bias only affects first impression formation in users with domain expertise, whereas novice users are more prone to automation bias due to having constantly high trust on the system.
- **H2**: Users with domain expertise and positive first impressions have a higher overall trust on the system compared to those with negative first impressions.
- **H3**: Users with domain expertise and positive first impressions will adjust their trust over time based on their observation of system errors, while those with negative first impressions will continue mistrusting the system, regardless of their observation of the system performance.

To test these hypotheses, we controlled two different orders of presenting system outputs: (1) Participants observed all the correct predictions in the beginning and all the mispredictions at the end (i.e., a positive first impression) and (2) observations follow the opposite order (i.e., a negative first impression). Note that the only difference in these conditions was the order of presenting the output, while the accuracy and observed trials remained the same. Figure 2 shows an example of an image with its corresponding label and explanation.

3.2 Experimental Design

To test our hypotheses, we designed a user study where participants were asked to review a set of images from a multi-class simulated classification scenario. Based on our research questions and goals, we sought an image classification domain where background knowledge is not a requirement, while having it could help a lot in completing the task. We chose entomology—the study of insects and other terrestrial arthropods—as a domain where novice users can partially identify system errors, but domain experts are expected to excel at this task. The *arthropod review task* consisted of a set of 40 trials with a classification accuracy of 47.5%, and was designed to allow the participants to experience and use the classification system over time in order to measure their level of trust.

We defined two independent variables for the study: *domain knowledge* and *order of observing correct outputs*.



(a) Desert Tarantula

(b) Ecuadorian Lubber Grasshopper

(c) Rhopalid (scentless plant bug)

(d) Metalmark Butterfly (*Calephelis* sp.)

Figure 1: Four examples of raw images from the study dataset. (a) and (b) are examples of easy-to-detect images while (c) and (d) are examples of hard-to-detect images.

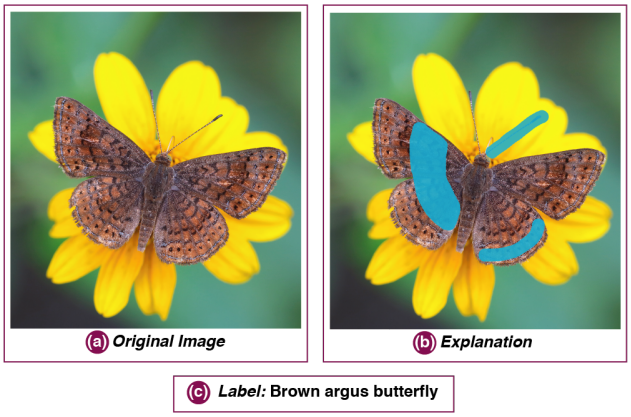


Figure 2: An example of what participants observed in the study. (a) the original image (which is a *Calephelis* sp.); (b) a blue saliency map to explain which regions of the image were used by an expert to determine the name and species of the arthropod; (c) the image classification label, which in this case is *incorrect*.

First, *domain knowledge* refers to a user's level of familiarity with and knowledge of entomology, for which we defined two levels: *novice* and *experienced*. Second, we controlled the *order of observing outputs* in two different manners. For the *correct first* level, all the correctly-classified outputs were observed in the beginning while all the misclassifications were shown afterwards. The reverse order was provided for the *wrong first* level.

The study followed a 2x2 between-subjects design, where each participant from *novice* and *experienced* group completed and observed the trials in one of the two defined orders. Since the subjects were exposed to the same set of trials, only with a different order, we incorporated a between-subjects design over a within-subjects design in order to avoid biases and learning effects.

3.3 Dataset

For the purposes of the study, the experiment's data used 40 high-quality macro images of different arthropods, photographed by an entomologist on the research team. Different regions in the world have bug species that are specific to each area and might not be found in other places. Since we were running the study in the US and our target entomology participants were mostly familiar with the arthropods in this country, all the selected arthropod images were from arthropod families that are found in the United States. Figure 1 shows examples of the raw images used in the study.

As our study was designed for both *novice* and *experienced* users, we designed the image set to contain a mix of both easy-to-detect and hard-to-detect arthropods in order to make the task fair for the *novices* as well. After selecting the images for the study, our expert entomologist generated a textual classification label for each image. The label contained the name of the arthropod and, in some cases, the family and species of the arthropod in brackets. For each image, our expert also created high-fidelity explanations in the form of saliency maps on top of the image. These saliency maps were chosen as portions of the image that the expert would use herself to detect the bug in the image. However, to address our goals and hypotheses of the study, we selected the classification accuracy of the simulated system as 47.5%, i.e., 19 images included correct labels and 21 images included false labels.

3.4 Participants

We recruited a total of 116 participants for this study, with 48 females, 61 males, and 7 others (non-binary, non-listed, or unknown). For the purposes of this study, we distinguish two groups of participants: 1) people who had *at least* 1 year of university coursework in entomology (i.e., the *experienced*), and 2) people with little or no familiarity with entomology (i.e., the *novices*). These two groups were recruited separately. The *novice* participants were recruited from undergraduate and graduate level university students, most of whom studying in computing majors. The *experienced* subjects were university students and practitioners in entomology or related fields. Among these participants, 71.23% held or were pursing a graduate degree.

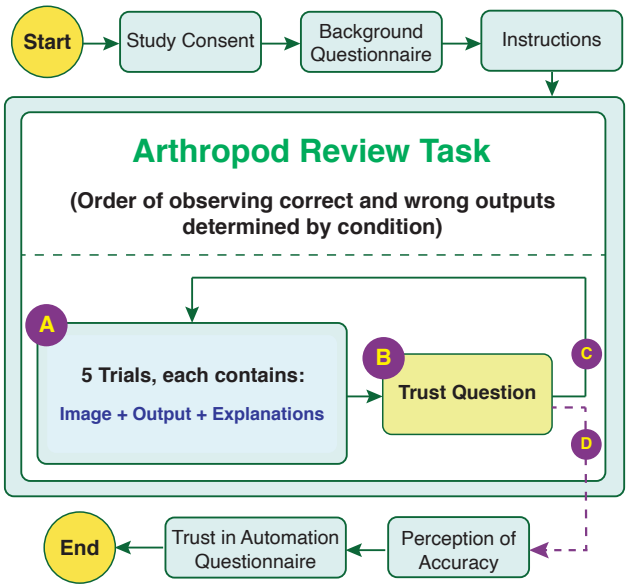


Figure 3: An overview of the study procedure. The Arthropod Review Task starts with *(A)*, consisting of 5 trials, and continues to a trust question *(B)*. By following *path (C)*, participants iterate through *(A)* and *(B)* seven more times. After answering to the trust question for the 8^{th} time, subjects continue to the post-study questionnaire through *path (D)*.

To help verify participants were considered in the appropriate group for expertise level, participants self-reported their level of familiarity with entomology as well as their occupation or major. Familiarity was measured through a seven-point Likert scale from 1 to 7 for *no knowledge* to *expert*, respectively. Since this self-reported measure is subjective, novices might overestimate their knowledge, whereas experts might underestimate it (Aqueveque 2018; Dunning 2011). A two-way factorial ANOVA found significant differences between these groups, with $F(1, 107) = 712.99$, $p < 0.001$, showing the domain-experienced group significantly rated their familiarity higher than novices.

3.5 Study Procedure and Measures

The user study was conducted online through a custom web application and took roughly 20 minutes. Participants were asked to complete the study in a single session using a preferred web browser on a desktop computer. The study was approved by the organization's Institutional Review Board (IRB). Figure 3 shows the overall procedure of the study.

The participants filled out a background questionnaire about demographic information, education, occupation, and familiarity with machine learning and entomology. They were shown instructions about the study and the task. After reviewing the instructions, participants started the *arthropod review task*, where they reviewed 40 trials. Each trial consisted of: 1) an image of an arthropod, 2) a textual label for the classification of the image (which might also include the family and species of the bug), and 3) a feature explanation for the classification in form of a saliency map (as described

(a) Average Trust Rating	Main Effect
	Domain Knowledge: $F(1, 107) = 39.07, p < 0.001$ *
	Order of Trials: $F(1, 107) = 17.66, p < 0.001$ *
	Interaction Effect: $F(1, 107) = 26.14, p < 0.001$ *
	Post-Hoc Test
	ExpCorrectFirst vs. *ExpWrongFirst* ($p < 0.001$) *
	NovWrongFirst vs. *ExpWrongFirst* ($p < 0.001$) *
	NovCorrectFirst vs. *ExpWrongFirst* ($p < 0.001$) *
(b) Change of Trust Rating	Main Effect
	Domain Knowledge: $F(1, 107) = 17.58, p < 0.001$ *
	Order of Trials: $F(1, 107) = 39.02, p < 0.001$ *
	Interaction Effect: $F(1, 107) = 39.02, p < 0.001$ *
	Post-Hoc Test
	ExpCorrectFirst vs. *ExpWrongFirst* ($p < 0.001$) *
	ExpCorrectFirst vs. *NovCorrectFirst* ($p < 0.001$) *
	ExpCorrectFirst vs. *NovWrongFirst* ($p < 0.001$) *

Table 1: Summary of results for average trust and change of trust. We used a two-way factorial ANOVA test for the main effect and a Tukey HSD test for pairwise comparison. For the post-hoc results, bold texts represent the conditions with (a) higher trust and (b) more change.

in section 3.3).

For each trial, participants were required to rate their agreement with two statements on a 5-point Likert scale, as seen below. In order to answer these questions, participants were advised to use their best judgement for identifying the arthropod in each trial, and in case the bug was unfamiliar, they were advised to refer to the provided explanations.

1. *I believe the highlighted explanation is appropriate with regards to the system answer.*
2. *I am confident that the system answer is correct.*

These questions were meant to focus user attention to the label and explanation of each image before moving on and to build an understanding of how the classifier works. After every five trials, participants were asked to report their level of trust on the system based on their observations up to that point. Trust was rated on a 7-point Likert scale from 1 (*distrust*) to 7 (*trust*).

After the *arthropod review task*, participants answered a questionnaire on trust in explainable AI (Hoffman et al. 2018) and estimated the system accuracy in percent. They also answered two free-response questions, asking them to explain how the system accuracy and their trust changed over time.

4 Results

We analyzed study results for the presented metrics based on the data collected from the *arthropod review task* and post-study questionnaire. For simplicity, we use *NovCorrectFirst* and *NovWrongFirst* condition names for *novice* participants, as well as *ExpCorrectFirst* and *ExpWrongFirst* condition names for *experienced* participants, with *correct first* and *wrong first* order trials, accordingly. For analysis, we used a two-way factorial ANOVA to test the main effects and Tukey HSD tests for posthoc pairwise comparisons.

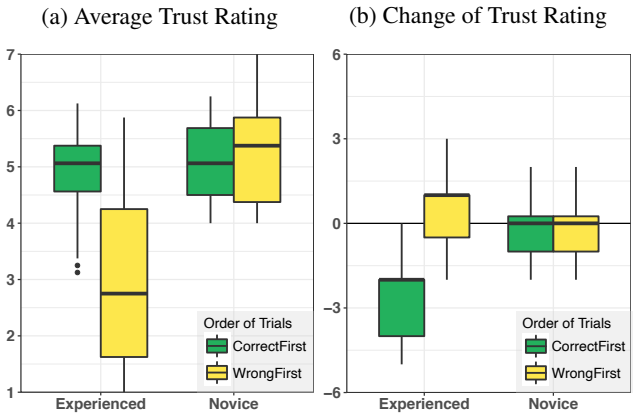


Figure 4: Results from self-reported trust in the *arthropod review task*. (a) average of the 8 self-reported trusts, and (b) difference between the first and the last reported trust. Negative values indicate trust declining over time and positive points show increasing trust.

4.1 Data Pre-processing

For quality verification, we removed the results from five participants due to data collection errors or evidence of lack of appropriate attention judged by their responses to the final open-ended questions. This left us with data from 111 participants, with *NovCorrectFirst*, *NovWrongFirst*, *ExpCorrectFirst*, and *ExpWrongFirst* having 28, 28, 28, and 27 data points, accordingly.

4.2 Average Self-reported Trust

We tested the effects of background domain knowledge and order of observing system correctness on user self-reported trust. Participants rated their trust in the system eight times during the *arthropod review task*. To address our first two hypotheses (H1 and H2), we calculated the average trust for each participant and compared them to find differences across the conditions. Table 1a and Figure 4a show the summary of the results and distribution of this data. The findings provide strong evidence that the average trust was significantly affected by domain knowledge and order of observing correct outputs. However, a significant interaction effect indicates these factors are interdependent. The pairwise comparison reveals that the order of observing correct system outputs is only significantly affecting trust for knowledgeable participants, which aligns with H1. Moreover, users with domain expertise and positive first impressions report significantly higher trust compared to those with negative first impressions (H2).

4.3 Changes in Trust over Time

To test our third hypothesis (H3), participants rated their trust throughout the study so we could track how it evolves over time. For each condition, we calculated the average trust of all participants per time-step, resulting in one trust value at each of the eight time-steps. Figures 5a and 5b show line charts of changes over time for the *novice*s and the *experienced* groups, respectively.

(a) Changes of trust over time for *novice* participants.

(b) Changes of trust over time for *experienced* participants.

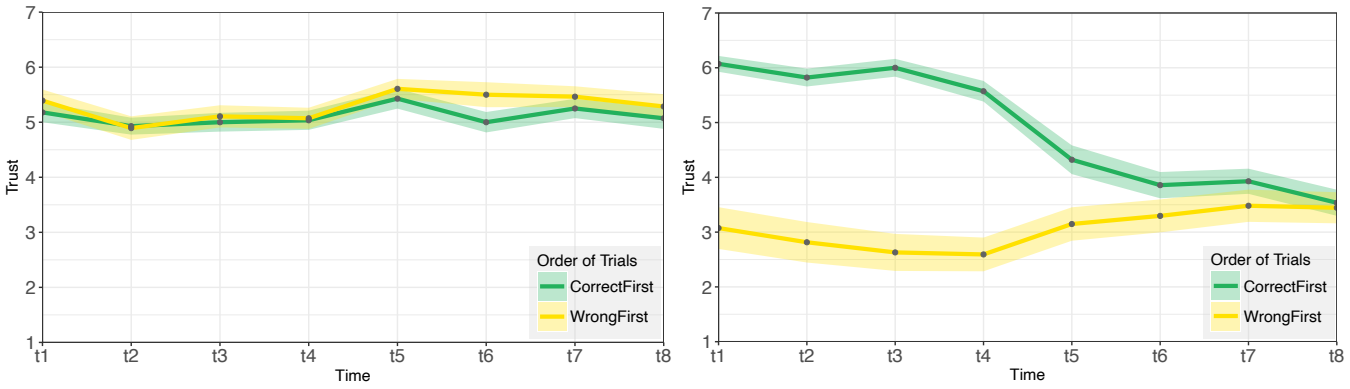


Figure 5: Average of participants' self-reported trust after every 5 trials in the main task. The y-axis indicates level of trust, where 1 indicates distrust and 7 indicates total trust. The ribbon around each line shows the standard error of the mean. The x-axis shows the time-step when trust question is asked (every 5 trials).

Ordering Condition	Common Themes		
	Strong Distrust	**Trust Decreased**	**Trust Increased**
CorrectFirst	0	21	0
WrongFirst	11	5	8

Table 2: Most common qualitative themes identified for how trust changed over time for participants from the experienced condition. The themes were retrieved from an open-ended question at the end of the study where participants were asked how their overall trust changed over time.

In order to statistically compare the magnitude and direction of change in trust over time across the conditions, we calculated the difference of initial trust and final trust. With this measure, negative values indicate declining trust and positive values indicate increasing trust. Table 1b and Figure 4b show the results for this comparison. Experienced participants in the *correct first* condition, had a significantly larger change-of-trust than those in the *wrong first* condition. The direction of this change was negative, indicating a loss of trust. To understand these results further, refer to Figure 5b. Participants from the *correct first* condition start with higher trust, which decreases over time; this is expected as the system accuracy lowers with time. In contrast, those from the *wrong first* condition start off with lower trust due to their negative first impressions, and the magnitude of their change-in-trust is significantly less than their counterparts', slightly going up while remaining relatively low. We did not observe any significant differences for novice participants.

We further reviewed the open-ended responses from the *experienced* participants to analyze how their trust in the system changed over time to understand the trends and themes based on first impressions. A summary of the main observations is presented in Table 2. We expected *experienced* participants to have a proper understanding of how and when the system accuracy changed. According to Yu et al. (2017), this understanding should reflect in their change of trust. Thus, we counted the number of participants whose comments indicated their assessments matched our expectation; that is, a decrease in trust for *correct first* and an increase in trust for *wrong first* participants. In total, 21 out of 28 *experienced* participants in the *correct first* condition stated that their trust was high in the beginning but lowered over time. From those in the *wrong first* condition, only 8 out of 27 indicated a slight increase in their trust. However, different themes were observed among these responses. For example, one participant who detected a slight increase in accuracy noted:

> *"Strangely, the system accuracy got much better towards the end, but by then I distrusted the system's [outputs, as] the misidentifications it made were too scandalous."*

Similarly, 11 out of 27 participants stated that regardless of the change in their trust, they did not trust the system at all. For instance, one participant noted:

> *"I didn't trust the system in the beginning and as the test continued, I only became surer that my distrust was the appropriate response."*

Moreover, 5 participants mentioned that their trust decreased over time, which is unexpected since the system grew more accurate towards the end.

These observations—backed by the statistical analysis for changes of trust—support our hypotheses (*H1* and *H3*) that first impressions of a system with local scope only matter when the user has background knowledge of the domain. Positive first impressions provide a chance for users to build trust and not give up on the system when it makes mistakes, while negative first impressions can cause an overall distrust in the system.

4.4 Post-Study Questionnaire

After the *arthropod review task*, participants estimated the accuracy of the classification system. It is important to keep in mind that all participants (regardless of the condition) observed the same simulated classification results with the same controlled accuracy across observed instances. The only difference was the order of observing the correct clas-

(a) Error of Perceived Accuracy	Main Effect
	Domain Knowledge: $F(1, 107) = 76.88, p < 0.001$ *
	Order of Trials: $F(1, 107) = 14.48, p < 0.001$ *
	Interaction Effect: $F(1, 107) = 13.13, p < 0.001$ *
	Post-Hoc Test
	ExpCorrectFirst vs. *ExpWrongFirst* ($p < 0.001$) *
(b) Trust in XAI questionnaire	Main Effect
	Domain Knowledge: $F(1, 107) = 87.84, p < 0.001$ *
	Order of Trials: $F(1, 107) = 3.75, p = 0.055$ (NS)
	Interaction Effect: $F(1, 107) = 2.10, p = 0.149$ (NS)

Table 3: Summary of results for error of perceived accuracy and trust questionnaire. We used a two-way factorial ANOVA to test the main effect and a Tukey HSD test for pairwise comparison. (a) for the post-hoc results, the condition with bold text shows higher overestimation of accuracy.

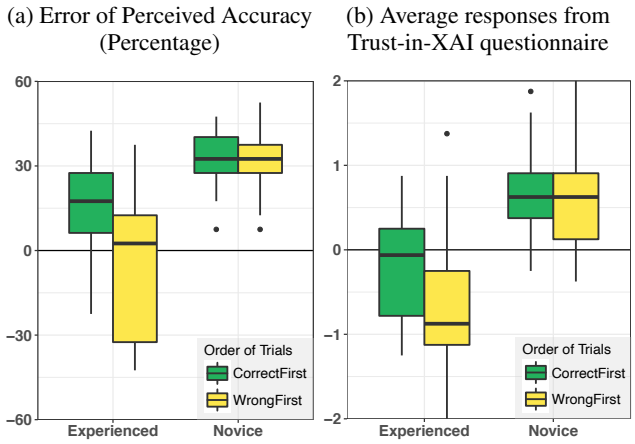


Figure 6: (a) Error of user-estimation of system (difference between user estimations and the actual observed system accuracy). Positive numbers represent overestimation and negative numbers represent underestimation of system accuracy. (b) Average of user responses to trust questionnaire results. Higher numbers indicate higher level of trust.

sifications. We assessed the error of each participant's perceived system accuracy by calculating the difference between the estimated accuracy and actual observed system accuracy. In addition, participants answered a set of questions about trust in automation and explainable AI systems (Hoffman et al. 2018) through a 5-point Likert scale, where higher values indicate higher trust in the intelligent system. We calculated the average of all questionnaire responses into one single score per participant and analyzed them to test for differences among conditions. Table 3 and Figure 6 show a summary and distribution of the results.

The results show that novices significantly overestimated the system while experienced participants may underestimate or overestimate the accuracy based on the order of observing the correct trials. The results from the trust questionnaire demonstrate that novice participants tended to trust the system significantly more than those with domain experience, which aligns with our first hypothesis (*H1*).

5 Discussion

Overall, our results demonstrate the importance of first impression formation with users with domain expertise and how it affects their trust. In this section, we discuss the results more generally, their importance to system designers, and possible future directions.

5.1 Interpreting the Results

Our first goal in this study was to understand whether first impression formation is influenced by domain expertise. Different implicit and explicit trust measures clearly indicate that novice users over-trusted the intelligent system although its overall accuracy was low. However, since domain experienced users tend to be more skeptical due to their knowledge, their overall trust depended on their early observations of the system performance. Domain experts' perception of system accuracy also varied by these first impressions, while novices always overestimated the accuracy. Previous research by Papenmeier et al. (2019) demonstrated users cannot be tricked into trusting a low accuracy intelligent system with high fidelity explanations. Our work builds on their findings and shows that given a domain-specific task where domain knowledge might be beneficial, novice users are prone to trusting such systems as they do not have enough knowledge to detect errors.

When comparing changes of user trust over time, domain experienced users showed different trends of trusting the system depending on the order of observed errors. Experienced users with positive first impressions had a significantly higher magnitude of change in their trust compared to those with a negative first impression (Figure 4b). This observation indicates that with positive first impressions, expert users tend to adjust their trust, whereas those with negative first impressions start with lower trust which stays low throughout the usage.

As previous work in automation bias and psychology shows, it is easier to lose trust than to reestablish trust (Hoffman et al. 2013). Starting with a system with good performance, experts are likely to form trust initially and to adjust their trust over time—also known as swift trust (Hoffman et al. 2013). However, our results show that expert users with negative first impressions lose trust in the system and stated that they are not willing to use it in the future (see Table 2). This has implications on real-world systems, as users might not continue using a system they do not trust (Dietvorst, Simmons, and Massey 2015).

5.2 Implications for Intelligent System Designers

This study presents important findings for intelligent system designers who are involved with designing domain-specific systems with various target users. Designing one system for all users is indeed tricky and requires certain considerations with user domain knowledge. Failing to account for such considerations can cause various problems such as under-reliance and over-reliance. Our results indicate that while experienced domain users tend to be more skeptical of the system, novices might not be able to catch system problems and suffer from automation bias.

System designers can incorporate techniques to help fill the knowledge gap for novice users and utilize techniques to guide domain experts into observing system performance, e.g., by using a more detailed explanation interface to provide more information. Rather than allowing users to use a system with zero understanding or a poor mental model of how the system works, designers can incorporate introductory sessions to alert users about system strengths and shortcomings so that users can decide whether and when they should trust the system's outputs. Alternatively, designers can provide a high-level overview of the model with key information (e.g., system accuracy and known weaknesses) that can influence impression formation. Additionally, one approach for consideration could be showcasing examples of both correct and incorrect predictions or outcomes in the beginning of usage to help experts develop first impressions, covering the variability of system capabilities to reduce the risk of encountering an unrepresentative sample by chance. Further research would need to explore the implications of such an approach. The showcasing method could also consider attempts to more strongly encourage users to review explanations for both correct and incorrect examples. Our results indicate that users tend to check the explanation for further information when they encounter errors, but not necessarily when they perceive the system to be correct. For novice users, however, errors need to be shown and explained to circumvent automation bias.

5.3 Limitations and Future Work Opportunities

This study contributed novel findings regarding how domain expertise can affect first impression formations and user trust. Our results show that novice users trusted the system regardless of the order of observing system errors and the overall low system accuracy. One possible explanation for this observation is novice users' inability to identify errors. A user's assigned level of expertise might vary with the domain and task at hand. While some systems consider novices as students or inexperienced domain knowledgeable users, we expected novices to have little or no domain knowledge at all. Thus, our study results from domain experts might also be observed for novices if they have the ability to judge system errors correctly.

Measuring trust is tricky, and any chosen evaluation methodology will have limitations. Directly asking subjects to rate their trust might bias them about the purposes of the study and hence, affect their response. Another issue relates to the use of Likert-scales for self-reported trust estimations and whether participants are able to differentiate the values in their response. Although 7-point Likert-scales are generally reliable for ordinal self-reported measures (Oaster 1989), we cannot control nor can we precisely know how each participant differentiates each point (e.g., value 5 from 6). In our study, we looked to qualitative data and free-response questions to help address this limitation and provide a better understanding for the collected explicit quantitative trust metrics. While the qualitative and quantitative results align, our study is limited in its ability to dissect specific characteristics of the observed trust and mistrust due to the open-ended nature of the free response data collection.

To maintain experimental control for our study, we selected a task where there is a definition for correct predictions. In other words, for these tasks, there is a concrete distinction between when the system makes a mistake and when not. To achieve this, the study is based on a multi-class image classifier with visual explanations. While such tasks are quite common in different domains and our results can be generalized for such intelligent systems, future work is required to verify if these findings hold for more exploratory and complex tasks. Specifically, for these tasks, errors might be challenging for users to detect, while they can also be difficult or impossible to define. For example, missing information or missing values can cause a model to predict different outcomes, all of which may be correct based on different hypothetical values for the missing information. As another example, recommendation systems strive to help users by making suggestions, but how these suggestions fit a user's needs is not easy to assess, and "false suggestions" are often impossible or difficult to define. Future research can extend the study of our research questions to such alternative analytical systems and tasks.

Finally, our findings show strong impression-formations for expert users based on instance-level observations of system performance. The presented study used a system with simple representations of model outputs. Future research of higher-level representations or visualizations can investigate how our findings generalize to contexts that allow deeper expert analysis of the model as-a-whole.

6 Conclusion

In this paper, we present a controlled human experiment to understand how user domain knowledge can affect first impression formation and trust calibration over time. Choosing entomology as our domain, we recruited domain-knowledgeable and novice participants to review outputs of a simulated arthropod-classification system with a low accuracy. Through the course of the study, we asked the participants to rate their trust in the system, and in the end, we measured overall trust implicitly and explicitly. Our significant results show that only those with domain knowledge form first impressions of the system. With a low accuracy system, we expected low overall trust from the subjects. However, encountering errors early-on resulted in a lower trust over time and a reluctance to use and rely on the system in the future. Though with positive first impressions, subjects calibrated their trust as they observed the system performance and were more likely to use the system in the future.

7 Acknowledgments

This work was supported by the DARPA Explainable Artificial Intelligence (XAI) Program under award number N66001-17-2-4032 and by NSF award 1900767. We would like to thank Dr. Vincent Bindschaedler for providing constructive feedback and suggestions on this study, as well as Emma Drobina, Brianna Richardson, and Prashant Singh for their initial efforts on this project.

References

[Alberdi et al. 2005] Alberdi, E.; Ayton, P.; Povyakalo, A.; and Strigini, L. 2005. Automation bias and system design: a case study in a medical application. In *2005 The IEE and MOD HFI DTC Symposium on People and Systems-Who Are We Designing For (Ref. No. 2005/11078)*, 53–60. IET.

[Aqueveque 2018] Aqueveque, C. 2018. Ignorant experts and erudite novices: Exploring the dunning-kruger effect in wine consumers. *Food Quality and Preference* 65:181–184.

[Berk and Hyatt 2015] Berk, R., and Hyatt, J. 2015. Machine learning forecasts of risk to inform sentencing decisions. *Federal Sentencing Reporter* 27(4):222–228.

[Bussone, Stumpf, and O'Sullivan 2015] Bussone, A.; Stumpf, S.; and O'Sullivan, D. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*, 160–169. IEEE.

[Cai et al. 2019] Cai, C. J.; Winter, S.; Steiner, D.; Wilcox, L.; and Terry, M. 2019. " hello ai": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW):1–24.

[Dasgupta et al. 2016] Dasgupta, A.; Lee, J.-Y.; Wilson, R.; Lafrance, R. A.; Cramer, N.; Cook, K.; and Payne, S. 2016. Familiarity vs trust: A comparative study of domain scientists' trust in visual analytics and conventional analysis methods. *IEEE transactions on visualization and computer graphics* 23(1):271–280.

[de Visser et al. 2012] de Visser, E. J.; Krueger, F.; McKnight, P.; Scheid, S.; Smith, M.; Chalk, S.; and Parasuraman, R. 2012. The world is not enough: Trust in cognitive agents. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 56(1):263–267.

[Desai et al. 2013] Desai, M.; Kaniarasu, P.; Medvedev, M.; Steinfeld, A.; and Yanco, H. 2013. Impact of robot failures and feedback on real-time trust. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 251–258. IEEE.

[Dietvorst, Simmons, and Massey 2015] Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1):114.

[Doshi-Velez and Kim 2017] Doshi-Velez, F., and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

[Dunning 2011] Dunning, D. 2011. The dunning–kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology*, volume 44. Elsevier. 247–296.

[Fourakis and Cone 2020] Fourakis, E., and Cone, J. 2020. Matters order: The role of information order on implicit impression formation. *Social Psychological and Personality Science* 11(1):56–63.

[Good 2000] Good, D. 2000. Individuals, interpersonal relations, and trust. *Trust: Making and breaking cooperative relations* 31–48.

[Goyal and Sharma 2019] Goyal, Y., and Sharma, A. 2019. A semantic machine learning approach for cyber security monitoring. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, 439–442. IEEE.

[Goyal et al. 2018] Goyal, H.; Khandelwal, D.; Aggarwal, A.; and Bhardwaj, P. 2018. Medical diagnosis using machine learning. *Bhagwan Parshuram Inst Technol* 7.

[Hoff and Bashir 2015] Hoff, K. A., and Bashir, M. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57(3):407–434.

[Hoffman et al. 2013] Hoffman, R. R.; Johnson, M.; Bradshaw, J. M.; and Underbrink, A. 2013. Trust in automation. *IEEE Intelligent Systems* 28(1):84–88.

[Hoffman et al. 2018] Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

[Hoffman, Klein, and Mueller 2018] Hoffman, R. R.; Klein, G.; and Mueller, S. T. 2018. Explaining explanation for "explainable ai". *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 62(1):197–201.

[Hohman et al. 2019] Hohman, F.; Park, H.; Robinson, C.; and Chau, D. H. P. 2019. S ummit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE transactions on visualization and computer graphics* 26(1):1096–1106.

[Honeycutt, Nourani, and Ragan 2020] Honeycutt, D. R.; Nourani, M.; and Ragan, E. D. 2020. Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. In *Eighth AAAI Conference on Human Computation and Crowdsourcing*.

[Kaur et al. 2020] Kaur, H.; Nori, H.; Jenkins, S.; Caruana, R.; Wallach, H.; and Wortman Vaughan, J. 2020. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

[Merritt et al. 2015] Merritt, S. M.; Lee, D.; Unnerstall, J. L.; and Huber, K. 2015. Are well-calibrated users effective users? associations between calibration of trust and performance on an automation-aided task. *Human Factors* 57(1):34–47.

[Mohseni, Zarei, and Ragan 2018] Mohseni, S.; Zarei, N.; and Ragan, E. D. 2018. A survey of evaluation methods and measures for interpretable machine learning. *ACM Transactions on Interactive Intelligent Systems*.

[Mosier and Skitka 1999] Mosier, K. L., and Skitka, L. J. 1999. Automation use and automation bias. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 43, 344–348. SAGE Publications Sage CA: Los Angeles, CA.

[Nourani et al. 2019] Nourani, M.; Kabir, S.; Mohseni, S.; and Ragan, E. D. 2019. The effects of meaningful and meaningless explanations on trust and perceived system accuracy

in intelligent systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7(1):97–105.

[Nourani et al. 2020a] Nourani, M.; Honeycutt, D. R.; Block, J. E.; Roy, C.; Rahman, T.; Ragan, E. D.; and Gogate, V. 2020a. Investigating the importance of first impressions and explainable ai with interactive video analysis. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–8.

[Nourani et al. 2020b] Nourani, M.; Roy, C.; Rahman, T.; Ragan, E. D.; Ruozzi, N.; and Gogate, V. 2020b. Don't explain without verifying veracity: An evaluation of explainable ai with video activity recognition. *arXiv preprint arXiv:2005.02335*.

[Oaster 1989] Oaster, T. 1989. Number of alternatives per choice point and stability of likert-type scales. *Perceptual and Motor Skills* 68(2):549–550.

[Papenmeier, Englebienne, and Seifert 2019] Papenmeier, A.; Englebienne, G.; and Seifert, C. 2019. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652*.

[Petrak et al. 2019] Petrak, B.; Weitz, K.; Aslan, I.; and Andre, E. 2019. Let me show you your new home: studying the effect of proxemic-awareness of robots on users' first impressions. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 1–7. IEEE.

[Ribeiro, Singh, and Guestrin 2016] Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

[Rudin and Ustun 2018] Rudin, C., and Ustun, B. 2018. Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces* 48(5):449–466.

[Vaidyanathan et al. 2014] Vaidyanathan, P.; Pelz, J.; Alm, C.; Shi, P.; and Haake, A. 2014. Recurrence quantification analysis reveals eye-movement behavior differences between experts and novices. In *Proceedings of the symposium on eye tracking research and applications*, 303–306.

[Yin, Wortman Vaughan, and Wallach 2019] Yin, M.; Wortman Vaughan, J.; and Wallach, H. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

[Yu et al. 2017] Yu, K.; Berkovsky, S.; Taib, R.; Conway, D.; Zhou, J.; and Chen, F. 2017. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 307–317.

[Zebrowitz 2017] Zebrowitz, L. A. 2017. First impressions from faces. *Current directions in psychological science* 26(3):237–242.

[Zhang, Liao, and Bellamy 2020] Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. *arXiv preprint arXiv:2001.02114*.