THESIS

WEATHER FORECASTING AUTOMATION ERROR TYPE, RELIABILITY, AND TRANSPARENCY AFFECT USE AND CORRESPONDING ATTITUDES

Submitted by

Haley L. Short

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2026

Master's Committee

Advisor: Jessica Witt

Co-Advisor: Christopher Wickens

Nathaniel Blanchard

ABSTRACT

WEATHER FORECASTING AUTOMATION ERROR TYPE, RELIABILITY, AND TRANSPARENCY AFFECT USE AND CORRESPONDING ATTITUDES

In two experiments, 208 and 163 participants completed a series of trials in which they were to decide if a school should remain open or close due to expected snowfall. These experiments differed in type of error that automation made (errors due to the challenge of predicting a noisy environment in Experient 1 and errors due to algorithm miscalculations in Experient 2). Participants were given a weather forecast automation prediction of snowfall whose predictions were either 70% or 90% reliable and were either accompanied by raw data (transparency) or not. Participants self-reported trust, and outcome measures of dependence and accuracy were also recorded. Overall, participants reported high trust of weather forecasts, regardless of the presence of transparency or level of reliability. Increasing reliability increased trust, dependence, and accuracy. We found trends that transparency is most helpful at lower reliability and that participants do not tend to depend on highly reliable automation as much as they should. Further, there are implications regarding the amount of uncertainty with a prediction decision by the user that automation does not account for regarding decision making.

ACKNOWLEDGEMENTS

I would like to express my sincerest gratitude to my advisor, Dr. Jessica Witt for her feedback, support, time and mentorship during this project. Without her, this work would not have been possible. Many thanks, also to my co-advisor Dr. Christopher Wickens for his shared expertise, patience, and feedback. Special thanks to my thesis committee member Dr. Nathaniel Blanchard, without which this work would not have been possible without his contributions of knowledge and expertise.

TABLE OF CONTENTS

WEATHER FORECASTING AUTOMATION ERROR TYPE, RELIABILITY, AND TRANSPARENCY AFFECT USE AND CORRESPONDING ATTITUDES

Over the past few decades, there has been an increased use of day-to-day automation. Weather forecasting, healthcare, and even food services are also beginning to utilize automation more, when automation is defined as completing tasks that used to be completed by humans (Parasuraman & Riley, 1997). It is becoming more pertinent that users of automation use it effectively. Typically, automation is used when it can operate better than a human can without the aid of automation (Wickens et al., 2023). It seems, however, that people are unsure how to utilize automation most effectively, either by depending on it too much or too little (Wischnewski et al., 2023). To facilitate teamwork between the user and automation, it is imperative that the user trusts and uses the system based on how effective it is (Wischnewski et al., 2023).

Automated weather models are often applied by weather forecasting services. Individuals may consult the weather app on their phone or the internet to determine what clothes they should wear that day, or if they will be able to drive to work due to expected snowfall. School age children might consult the weather app to determine if they will get a “snow day” the following day. Decisions related to weather forecasting are made with many factors in mind, but automation can help simplify this decision. To decide whether to close a school, weather forecasting automation can pool factors with increased reliability and efficiency to help humans make critical decisions. Weather-related mortality in the context of driving on icy roads or hurricane warnings could potentially be mitigated by improved communication of risk to the public or better calibration of user’s trust in the automation (Losee & Joslyn, 2018; Elder et al., 2007).

Users can be influenced by two components of automation: its performance (reliability) and the display of its inner workings (automation transparency). We explored human interaction with automation (measured by trust in automation, dependence on automation, and performance with automation in the context of a weather-related decision (modeled from Burgeno & Joslyn, 2020). The task in the current experiment required participants to decide whether to close a school due to forecasted snowfall. The school was to be closed if snowfall exceeded 6 inches.

Automation Reliability

Automation reliability is defined in terms of how often the automated system errs. When predicting an uncertain future, automation errors will happen, so reliability is unlikely to be 100% for any given system (Sargent et al., 2023). It is expected that weather systems will err at times because of the nature of predicting a noisy environment (Sanchez et al., 2014). A system with a higher reliability rating corresponds with an improved relationship between the system and the user (Hancock et al., 2011), until a system may be so reliable that an individual is unable to respond quickly or sufficiently when the automation fails, because of overdependence on the system (Parasuraman & Riley, 1997). Furthermore, increasing reliability typically produces increased performance, trust, and dependence on the system (Pharmer et al., 2021; Holthausen & Walker, 2022; Wickens et al., 2023; Patton & Wickens, 2024). An individual should trust an automation system at a rate that is proportional to the reliability of the system. This proportionality is defined as optimal trust calibration.

It is crucial to highlight the context within which an automation error may occur and the type of error. Errors may be plausible and subsequently more forgivable (such as there being 4 inches of rain when only 3 inches were forecasted), or they may be catastrophic (such as when a hurricane hits harder or in a different location than expected). In application to the current

experiment, 2 inches might have been predicted when 8 inches fell. Therefore, different types of automation errors can have different effects on the human automation team. In the current experiment, automation can make an error that does not influence the decision made based on an automation diagnosis. For example, in a scenario for which 6 inches of snow occurred after only 4 inches were predicted. The deviation may have little significance. But in some situations, this deviation between the outcome and the forecasted value could be consequential. As an example, in the school closure task devised by Burgeno & Joslyn (2020), schools should be closed if 6 or more inches of snow occur. In this case, a deviation of 2 inches means the user is likely to keep the schools open based on the forecast of 4 inches, but the outcome of 6 inches means the schools should be closed. Instead, if the forecast was 3 inches but the outcome was 5 inches, the deviation is still 2-inches but the difference does not cross the decision threshold. For both 3 inches and 5 inches the schools should remain open, so the automation error has little consequence.

In terms of Signal Detection Theory, an automation error is classified as when the automation predicts an amount of precipitation that causes the incorrect decision to be made regarding school closure, (see Table 1) either a miss or a false alarm.

Table 1

Signal Detection Theory Matrix for the current experiment regarding when automation makes an error.

	Outcome states that the school should have been closed	Outcome states that the school should have remained open
Prediction denotes that the school should be closed	Hit	False alarm
Prediction denotes that the school should remain open	Miss	Correct rejection

Automation Transparency

Automation transparency may be defined as the information that provides the user with a better mental understanding of what the automation is doing and its reasons for doing it (Sargent et al., 2023). In a meta-analysis of effect sizes, Sargent et al. (2023), reported that introducing transparency or increasing its level significantly increased trust (0.81) and dependence on automation (0.45). Transparency also decreased error rate (-.99). Furthermore, dichotomized effect sizes were reported for routine tasks compared to failure recovery, when automation fails. Trust increased by increasing transparency significantly for both routine tasks (0.72) and failure recovery (1.12). Dependence also increased for both routine tasks (0.22) and failure recovery (0.92) with transparency. The error rate also significantly decreased with transparency for both routine tasks (-1.02) and failure recovery (-0.44). A meta-analysis by Schemmer and colleagues (2022) also found that AI explanations of decision-making improve performance of the user.

A literature review by Van de Merwe et al. (2022) found that while transparency improved performance overall, more transparency did not necessarily equate to better performance. Specifically, when performance was quantified as accuracy, some studies did not report better accuracy with increased automation transparency, (Guznov et al., 2020; Patton et al., 2023; Pharmer et al., 2025; Roth et al., 2020; Wright et al., 2020), though no studies reported a statistical decrease in accuracy with increased automation transparency.

Automation transparency may be presented in multiple different forms, such as by explaining to the user how the system works, providing a self-assessed confidence rating of the automation prediction, or by providing raw data that the system used to derive its model (Sargent et al., 2023). Many experiments that manipulated transparency provided additional information that the automation is using to compute its model, depending on the context of the automation task. Others provided communication to the user from the automation about information that

may be missing. Other transparency studies used mixed forms of transparency and manipulate the levels of transparency rather than transparency being present or not (Van de Merwe et al., 2022). One study found that trust increased most when transparency not only explained how an automated system would react, but why that reaction was decided (Zhang et al., 2025). It seems that transparency is most effective when it provides insight as to why a decision was made or advised by automation (Luo et al., 2022).

Types of Automation Errors

In addition to the magnitude of the automation error discussed above, automation may err for different reasons. In Experiment 1, the assumption is that automation makes errors due to the challenge of predicting a noisy environment. Therefore, the transparency that accompanies the automation matches the prediction of the automation. In other instances, transparency might indicate errors in the algorithm of the automation by displaying raw data that is misaligned with the display of the automation. In this case, users of automation could detect automation errors by the increased understanding that the raw data of the transparency provides.

The type of error that automation commits could have varying implications for how users interact with automation following its failure. For instance, failures due to a noisy environment may be more forgivable for users and not influence their trust in automation, whereas algorithmic errors may significantly decrease the foundational trust that users have with the system. One purpose of automation transparency is to allow users to view raw data associated with automation's predictions and note that automation is making an error, then react or correct accordingly (Sargent et al., 2023).

The difference in Experiment 1 and Experiment 2 models the algorithmic type of error to be made by weather forecasting but can be applied to other types of automation. For instance,

self-driving cars utilize automation transparency, but the user can also utilize their own form of automation transparency by looking at the road ahead of them and ignoring the display of the car to determine if they want to follow the guidelines of automation.

Trust and Dependence in Automation

Trust and dependence are conceptually related but function differently within the context of automation. Dependence is typically related to trust, as the attitude of trust that one has for a given automation system corresponds with how much they will depend on the automation system. That is, if one trusts an automation system a lot, they will likely be highly dependent on the system and vice versa (Endsley, 2017; Patton & Wickens, 2024; Sargent et al., 2023). In a study by Joslyn & LeClerc (2012) participants were assigned a road salting task that was based on the forecast of overnight low temperatures. Participants reported significantly higher trust in more reliable conditions as compared to less reliable conditions and were able to properly take action. In another experiment, participants rated higher competence and trustworthiness in reliable or accurate financial analysts and were more likely to continue purchasing reports from them, that is, increasing their dependence (Kadous et al., 2009). In a third study, when mammography patients were asked to imagine being given an initial false positive breast cancer report, they indicated that this would yield decreased trust and delay of future use of mammography test, that is, decreased dependence (Kahn & Luce, 2003). Burgeno & Joslyn (2020) found that users trusted inaccurate or unreliable weather forecasts less than accurate or reliable weather forecasts. These studies demonstrate that trust and dependence are related to one another and to reliability in multiple contexts.

There are some additional factors that may improve trust aside from system reliability. According to a study by Losee & Joslyn (2018), severity, consistency, expectation of harm, and

familiarity influenced trust specifically in weather forecasts. Furthermore, some individuals are simply more trusting as a baseline and are more likely to be trusting of automation (Hoff & Bashir, 2015).

However, trust and dependence do not always increase at the same rate. In a meta-analysis, Patton & Wickens (2024) found that increasing reliability increased trust by an effect about twice as large as the increase in dependence (Davenport & Bustamante, 2010; Merritt & Ilgen, 2008**;** Wickens et al., 2020). Therefore, trust and dependence, though related, should be measured separately.

Furthermore, Ku and colleagues (2025) found evidence that the relationship between trust and dependence could be moderated by workload. Trust predicts dependence more significantly when workload is low and less when workload is high. Thus, trust may be a worse predictor of dependence for a more complex task than for a simpler task.

Dependence on automation is a behavioral measure of how frequently a person uses automation when it is available (Patton, 2023). A high level of dependence on an automation system decreases performance quality and speed if the automation fails (Sargent et al., 2023). In correspondence, over-trust can lead to no reaction or a delayed reaction in the context of automation errors (Kunze et al., 2018). Both of these effects illustrate automation bias.

As explained by Parasuraman & Riley (1997), there are different types of automation bias and reasons that these types of biases may occur, such as decision biases, use of decision-making heuristics, lack of feedback, or lack of monitoring automation. Automation bias or misuse can be found with high levels of dependence by the user. Since the implementation of automation transparency increases monitoring and can provide feedback for the user. Therefore, we would predict that transparency could decrease automation bias if implemented and attended to.

If a user has no trust in automation system, they will not use the automation (Lee & See, 2004; Hoff & Bashir, 2015). However, the goal of automation use is to reach a level of trust that is calibrated to the reliability of the system. Users trust a low-reliability system less than optimal, even if the reliability of the system increases over time (Rittenberg et al., 2024). In the context of weather, users that have limited trust in automation may not use it. For large-scale weather events, this could cause people to disregard evacuation orders.

Performance and Automation

Human performance as part of the human automation team is dependent on the ability of the human to use an automated system and the reliability of the system (Korbalich et al., 2018). It is also true that performance with automation varies widely from person to person (Eriksson & Stanton, 2017; Xu et al., 2007). Additionally, multiple studies have found that human performance often falls at or lower than the reliability of automation, but that the automation and human working together perform better than the human without any automation (Boskemper et al., 2022, Bartlett & McCarley, 2017, 2019, 2021, 2022; Munoz Gomez Andrade et al., 2025). These experiments were conducted utilizing different tasks with some variation in performance, but these trends remain consistent. Especially at high reliability, users of automation perform worse than the automation alone, a term referred to as ironic efficiency (Bartlett & McCarley, 2021) because as reliability of automation decreases, users perform as well as automation. This means that users of automation depend on it less when it is more reliable.

Automation Transparency and Reliability

Few prior studies have investigated the relationship between reliability and transparency, in particular with dependent variables of trust, dependence, or performance. Hussein et al (2020) manipulated reliability at 70% and 90% and transparency being present or not in a simulated UV

task. They found that higher reliability increased performance, dependence, and trust of the automation. They also found that transparency increased trust and performance. They found no interaction between reliability and transparency on any outcome measure. Wright et al. (2020) manipulated reliability at 67% and 100% and transparency being present or not where participants monitored a dismounted soldier team with the help of automation providing advice. They found no interactions with outcome variables and that trust increased with increased reliability. Gegoff et al. (2024) had participants complete UV missions with the help of recommended advice that varied in reliability (65% or 90%) and transparency levels being medium or high. They found an interaction between transparency and reliability, such that transparency mitigated the diminishing effect of reliability on trust. They found that increased reliability corresponded with higher rates of accuracy and trust. Additionally, increased transparency led to higher accuracy rates and lower rates of automation disuse for low reliability, but not high reliability. Finally, Kaltenbach and Dolgov (2017) manipulated reliability (65% or 95%) and transparency (one line or multiple lines of text of what was happening in the system) for a coffee manufacturing task. They found that transparency interacted with reliability for trust, that transparency amplified the effect of reliability on trust. They also did not find that trust increased by increased reliability. More recently, in an unpublished study, the current authors found that automation transparency mitigated the increasing effect of reliability on trust and dependence.

Two studies have found an interaction between reliability and transparency. The Kaltenbach and Dolgov (2017) study was not peer reviewed and was published as a proceeding for the Human Factors annual meeting, but the Gegoff et al. (2024) study was peer reviewed and

published in the Human Factors journal. Therefore, the findings from the latter mentioned experiment are those hypothesized for the current experiment.

Foundation for the Present Study

Burgeno & Joslyn (2020) conducted a series of experiments in which they instructed participants to use a series of weather predictions to determine if a school should be closed due to snowfall. Burgeno & Joslyn (2020) manipulated the accuracy of the forecasts and their consistency across time. They found that inaccurate forecasts yielded decreased user trust.

In the current experiment, participants completed a similar task. However, automation reliability (i.e. its accuracy) and the presence of automation transparency were jointly manipulated. Prior research has established that reliability increases user trust, dependence, and performance. Furthermore, the current experiment examined the influence of automation transparency in the trust and dependence calibration process in the context of a weather-based task. It was expected that there would be an interaction effect of reliability and transparency on trust and dependence. Based on the findings of Gegoff et al. (2024), we expected to find that transparency reduced the diminished effect of low reliability on trust and dependence.

Hypotheses

- H1: Participants will have increased trust, dependence, and performance as automation reliability increases.
- H2: The implementation of transparency will yield increased trust and dependence.
- H3: The presence of transparency will reduce the effect of reliability on trust and dependence.
- H4: The performance of the human automation team will be worse than the performance of the automation, especially at high reliability.

- We expect to find differences in trust, dependence, and performance when comparing the results of Experiment 1 to those of Experiment 2, addressed in the General Discussion.
    - A. Participants will trust automation transparency predictions more in Experiment 2 (than in Experiment 1) because the transparency in Experiment 2 allows participants to detect automation errors.
    - B. Participants will depend on automation with transparency less in Experiment 2 (compared to Experiment 1) because they will find that the automation transparency is not aligned with the forecasted weather predictions.
    - C. Participants will be more accurate with transparency in Experiment (compared to Experiment 1) because they will be able to determine when a forecasted prediction will result in an error.

EXPERIMENT 1

Method

Participants

In total, 208 students completed the survey for the current experiment in exchange for course credit. Of the participants, 79% identified as female, 19% identified as male, and 2% identified as non-binary. The participants had an age range of 18-42 years, with an average age of 19.

Task

In the current experiment, participants were asked to provide decision advice to a hypothetical school regarding whether they should stay open or close due to an upcoming snowstorm. They were told that they should close the school if snowfall is 6 inches or more and remain open if snowfall is less than 6 inches. Several factors are considered when schools make

the decision to close or remain open, but in this task, the decision was based on forecasted snow accumulation alone. Participants were provided with a prediction for snowfall on Tuesday and asked if they would advise the school to close on Wednesday. They were asked to self-report how much they trusted the weather forecast prediction on each trial after receiving the predicted snowfall (before deciding to remain open or close the school, *trust1*) and after reviewing the outcome snowfall (after deciding to remain open or close the school, *trust2*). The former is referred to as trust1 and the latter is referred to as trust2.

Design

The experiment was comprised of four total blocks, but each participant only completed two blocks total. The four blocks were based on a two (reliability 70% or 90%) by two (transparency, present or absent) design. Each block utilized a different fictitious weather forecasting service (such as "Weather Now") so that the blocks would be independent of each other. At the end of each block, participants were asked how much they trusted the fictional weather forecaster with the same self-reported scale used during the trials (*final trust*). The effects of imperfect reliability on trust were expected to be cumulative over trials, so final trust at the end of each block more clearly represents overall trust of a weather forecaster by participants.

Each block contained 10 trials, with 20 trials in total completed by each participant. Of the 10 trials in a given block, 7 predicted values of snowfall that were close to the threshold, but 3 were either low predictions of snowfall or high predictions. The high or low predictions may have erred by a 1-inch discrepancy but did not cross the decision threshold (the school should be closed at 6 inches of snowfall or more). One example of this is the fifth trial in the first block is shown in Table 2. In the case of these trials, there was an error of diagnosis, but not an error of decision recommendation by the automation. Therefore, these trials would be classified as a hit

or correct rejection (see Table 1 above for Signal Detection Theory Matrix). The experimental trials that were the false alarm and the miss were unreliable because they provided a prediction that would lead participants to close the school when it should have remained open (false alarm) and remain open (miss) when they should have closed the school, respectively. There was a 2-inch discrepancy across the decision threshold in these automation error trials. Automation had a reliability rating of either 70% or 90%. In other words, following the predictions provided by the weather forecasters would result in a correct decision 70% and 90% of the time, respectively. With the signal detection theory framework, our focus is on the accuracy of decisions and not the specific amount of snow that occurred.

The stimuli used for predictions and outcomes for each trial are shown in Table 2. The order of trials within a block was the same for all participants, but the order that blocks appeared for participants was counterbalanced, meaning participants were assigned to either the 70% or 90% reliability group and whether they received the transparency or non-transparency block first was randomized.

Table 2

This table presents all the stimuli for Experiment 1. It includes trial numbers, predicted snowfall, outcome snowfall, the weather forecaster, block, and trial type.

Trial Number Within Block	Predicted Snowfall	Observed Snowfall	Weather Forecaster	Block	Trial Type
1	2	2	TruWeather	90% Reliable No Transparency	Correct Rejection
2	6	4			False Alarm
3	5	5			Correct Rejection
4	6	6			Hit
5	1	2			Correct Rejection

6	4	4			Correct Rejection
7	8	8			Hit
8	7	7			Hit
9	4	4			Correct Rejection
10	7	7			Hit
1	7	7	WeatherNow	70% Reliable, No Transparency	Hit
2	9	8			Hit
3	6	4			False Alarm
4	5	5			Correct Rejection
5	4	6			Miss
6	1	2			Correct Rejection
7	4	6			Miss
8	5	5			Correct Rejection
9	9	9			Hit
10	7	7			Hit
1	5	5	Weather Direct	90% Reliable, Transparency	Correct Rejection
2	8	8			Hit
3	2	2			Correct Rejection
4	5	5			Correct Rejection
5	4	6			Miss
6	7	7			Hit
7	6	6			Hit
8	4	4			Correct Rejection
9	2	1			Correct Rejection
10	6	6	The Weather Company	70% Reliable, Transparency	Hit
1	5	5			Correct Rejection
2	1	1			Correct Rejection
3	9	8			Hit
4	6	4			False Alarm
5	7	7			Hit

6	5	7	Miss
7	8	9	Hit
8	6	6	Hit
9	7	5	False Alarm
10	4	4	Correct Rejection

Reliability was defined by the proportion of times that the automation made an error. In this case, an error is represented by a two-inch difference between the prediction and the outcome that, critically, crosses the decision threshold of 6 inches. For instance, in the error of diagnosis trials of the imperfectly reliable blocks, the prediction may have been 2 inches, and the outcome may have been 3 inches of snowfall. Although this prediction errs in the exact amount of snowfall, the decision threshold is not crossed, so the trial is still classified as a correct rejection.

Automation transparency was manipulated by displaying the raw data (see Figures 1 and 2) that was presented along with the prediction estimate by the fictitious weather prediction system. In the blocks with automation transparency, participants were provided with a key that informed them what the icons represented in the raw data meant (see Figure 1). Raw data represents information that the weather forecaster used to formulate their prediction of snowfall based on factors that are used in weather forecasting. Participants were informed that there were four key pieces of information that the system considered when making its prediction of snowfall. They were not given a reliability rating for the weather forecaster, they were instead asked to determine on their own how much trust they should have in the predictions provided. It was assumed that participants would calibrate to the reliability of the automation as they progressed through the sequence of trials in a block as they experienced forecast errors. Participants were able to refer to the key at any point as they examined the map given to them for each trial.

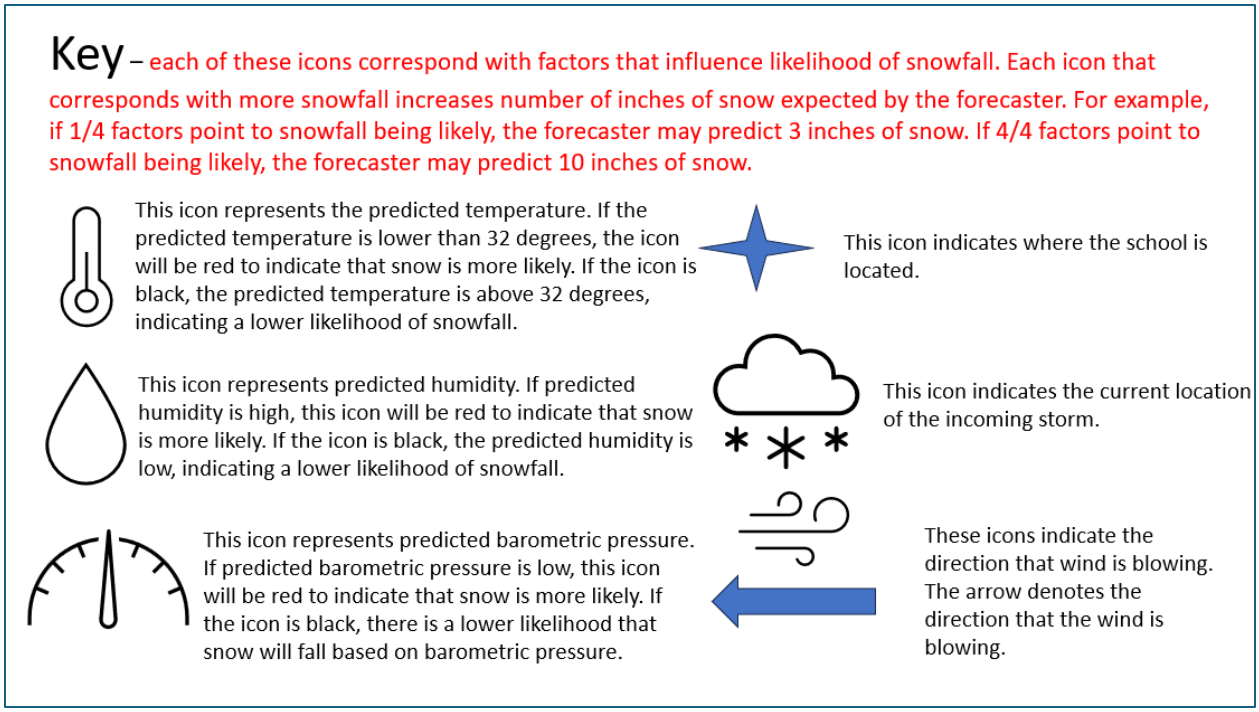

Figure 1

In the blocks with transparency, participants were provided with this key that indicated what the icons on the map meant.

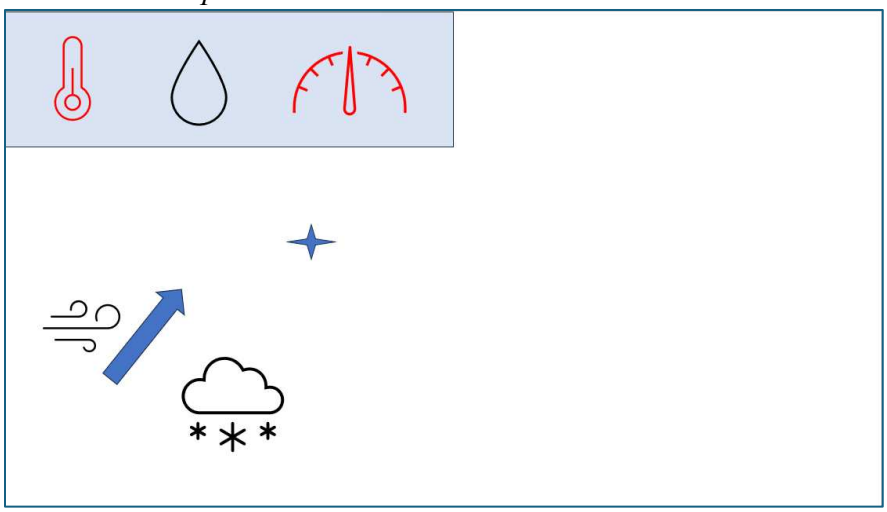

Figure 2

Example of raw data output for transparency block. This arrangement of icons represents 6 inches of snowfall.

The raw data were represented through a series of dichotomous presented icons that represented components of weather forecasting. These were temperature, humidity, and

barometric pressure. These elements were presented as indicating that the amount of snowfall would be heavier or lighter, which was demonstrated by the color of the icons. If the temperature was low, humidity high, or barometric pressure low, that icon would present in red. If the temperature was high, humidity low, or barometric pressure low, the corresponding icon would be black. Red icons indicated more evidence for heavy snowfall and black icons indicated more evidence for light snowfall. Additionally, a map was provided which showed the location of the school, location of the storm, and current wind direction. The fourth cue is whether the storm will hit the school the next day based on the current wind direction relative to the location of the school. The graphics that represented the raw data were created on the basis that the greater number of the four components pointing towards more snowfall increased the likelihood of larger amounts of snowfall. Participants were informed that there were four components considered by the weather forecast automation to make their decision. If one of the four components indicated that snow was likely, the predicted snowfall was 2-4 inches. If four out of four components indicated that snow was likely, the prediction would be 8 inches of snowfall (see Table 3). Participants were not told this about the cues; however, this is how the transparency stimuli were generated.

Table 3

Indication of stimuli generation for transparency trials.

Number of cues indicating heavy snowfall	1	2	3	4
Predicted snowfall	2-4 inches	4-6 inches	6-8 inches	8-10 inches

Procedure

The survey for the experiment was administered online through Qualtrics. First, participants were provided with a description of what they would be doing in the experiment (see Figure 3). Next, in the introduction block, participants were asked to go through two practice trials with additional instructions to familiarize themselves with the trials. The first practice trial

was representative of a trial with no transparency, and the second practice trial was representative of a trial with transparency. Both practice trials were representative of a trial from a reliable block.

Welcome to the experiment!

Please expect to spend 30 minutes of continuous attention on this task.

Please imagine you are in charge of connecting the weather forecast related to snow accumulation (how many inches of snow) to the decision of **whether or not to close schools.**

In this situation, schools should be **closed** whenever there is **6 inches or more** of snow.

You will make these decisions over the course of a winter season.

For each week in which snowfall is predicted, you will be shown a forecast for predicted snow accumulation on Tuesday, then make your decision about whether schools should be closed the next day, on that Wednesday.

Remember that if you choose to close the school and it should have remained open, this **costs the school unnecessary money** because they will have to have makeup days. Also, remember that if you choose to remain open when there turns out to be 6 inches of snow or more, **road conditions are dangerous and safety is a concern**.

You will first go through **two practice trials** to help you understand how the trials work. In one practice trial, you will only be shown the predicted forecast. In the other, you will be given additional information as to why the weather forecaster made their prediction. During this experiment, you will go through 10 trials of each type.

You will be given predictions from **two different fictitious weather forecasters**. They may tell you why they made the prediction they made and they may not. **They may be entirely reliable in their predictions, and they may not**. You will receive feedback after your decision to close the school or remain open on each trial that will tell you if the weather forecaster prediction was reliable or not.

Each fictitious weather forecaster will give you predictions for snowfall ten times. At the end of those ten trials, you will be asked to rate your overall trust for the weather forecaster.

Figure 3

The instructions that participants were presented at the beginning of the experiment.

Each trial consisted of three screens viewed sequentially. In a trial without transparency, one the first screen, participants were provided with a prediction of snowfall from a fictitious weather forecaster on Tuesday for Wednesday (for example: “TruWeather predicts 2 inches of

snow on Wednesday"). On the same page, participants were asked, "How much do you trust Tuesday's forecast?". This was recorded using a drop-down menu with 6 option choices of "Not at all", "A little", "Somewhat", "Quite a bit", "Very much", and "Completely". Every time that participants reported trust ratings, the same drop-down menu was used. This measure is described as "trust1"

In a trial with automation transparency, on the first screen, participants were provided with a weather forecast prediction, along with a key and graphic with raw data to give additional insight into the reasons that the weather forecaster made the prediction that it did. The rest of the parts of the trial were the same for transparency and non-transparency trials.

On screen two, participants were reminded of the weather forecast prediction on Tuesday for Wednesday. Then, they were presented with the option to make the decision to close the school or remain open. Participants were asked "Do you want to close the school tomorrow?" Below this question were the answer options as follows "Close (I think the snow accumulation will be 6 inches or above)" and "Remain open (I think the snow accumulation will be less than 6 inches)". In this decision question, participants were able to choose if they wanted to close the school or remain open, regardless of what the prediction indicated. Therefore, they could have chosen to act in accordance with the forecasted predictions or opposite them.

After participants selected a decision regarding keeping the school open or closing the school, on screen three, they were informed that the school followed their advice and provided with the observed snow accumulation that would have been available on Wednesday morning. Participants also saw a note about if the automation was incorrect or not, for example, "Since WeatherNow predicted 7 inches of snow and 7 inches of snow were observed, WeatherNow made a prediction that corresponded with the correct decision" for an automation correct trial;

alternatively “Since WeatherNow predicted 4 inches of snow and 6 inches of snow were observed, WeatherNow made a prediction that may have resulted in the incorrect decision” for an automation incorrect trial, if they followed the automation’s advice. Then, participants were asked how much they trusted the weather forecast automation’s prediction to help them make their decision with the same drop-down menu described earlier. This measure of trust is referred to as “trust2”. This third screen concludes what participants saw for a single trial.

It was expected that at the beginning of a trial, participants have no expectations of the reliability of a weather forecaster’s predictions. However, as participants progressed through the trials in a block, their trust rating of a weather forecaster may have increased or decreased as they got a sense of how much they should have trusted the forecaster’s predictions.

At the end of each block, participants were asked how much they trusted the weather forecaster’s predictions overall. They selected their response using the same drop-down menu described earlier. This rating, referred to as “final trust” was presented after participants have received all relevant information regarding how trustworthy the weather forecaster is. This rating is their final determination of how much they trusted the predictions of each weather forecaster. Participants were informed that the next block would utilize a different weather forecaster.

After completing the experimental blocks, participants completed four additional qualitative questions about their experience in the experiment. They were asked (1) which type of forecast they preferred and (2) which type of forecast they found more accurate. They were also asked about the (3) proportion of errors they remember the weather forecast predictions to have made, and (4) if they found the additional information presented to them (transparency) to be helpful. (5) Participants also self-reported how often they typically trust weather forecasts. Lastly, participants reported demographic questions, such as their gender and age. Then,

participants were directed to a link where they could get course credit for completing the experiment.

Results

Of the 209 participants who completed the experiment, 202 were included due to the removal of 7 outliers. Participants were removed from the dataset if their subject-level accuracy rating was less than or equal to 50%, meaning their accuracy was equivalent to or worse than chance.

All analyses were done in R Studio, the packages that were used can be found in Appendix A. These analyses were conducted using linear and general mixed effects models, where the random effect was the subject due to the within-subjects nature of the experiment.

Trust

Measures of trust are presented in Figures 4, 5, and 6. As described above, trust ratings were taken twice per trial and at the end of each block. Trust was self-reported using a drop-down scale of 1-6. The first time that trust was measured for each trial (trust1) was after the prediction of snowfall was provided by the weather forecaster. The second time that trust was measured for each trial (trust2) was after the participant was shown the outcome snowfall, and hence whether or not the forecast was in error. At the end of each block, participants were asked how much they trusted each weather forecaster overall (final trust). Measures of trust were computed using linear mixed effects models. The dependent measures were trust1, trust2, and final trust, respectively. The fixed effects were reliability, transparency, and their interaction.

Trust1. To test if trust1 differed among conditions, we ran a linear mixed model. Ratings of trust1 followed an approximately normal distribution. We found that transparency increased trust1 ($x = 0.39$, $t = 9.26$, $p < .001$), no effect of reliability on trust1, ($x = 0.01$, $t = 0.20$, $p = .84$),

and a significant interaction between reliability and transparency ($x = 0.17$, $t = 2.93$, $p = .003$). The marginal R^2 value for the model was .04 and the conditional R^2 was .33.

Participants trusted the prediction of weather forecasters more when there was transparency versus when there was not transparency at both reliability levels.

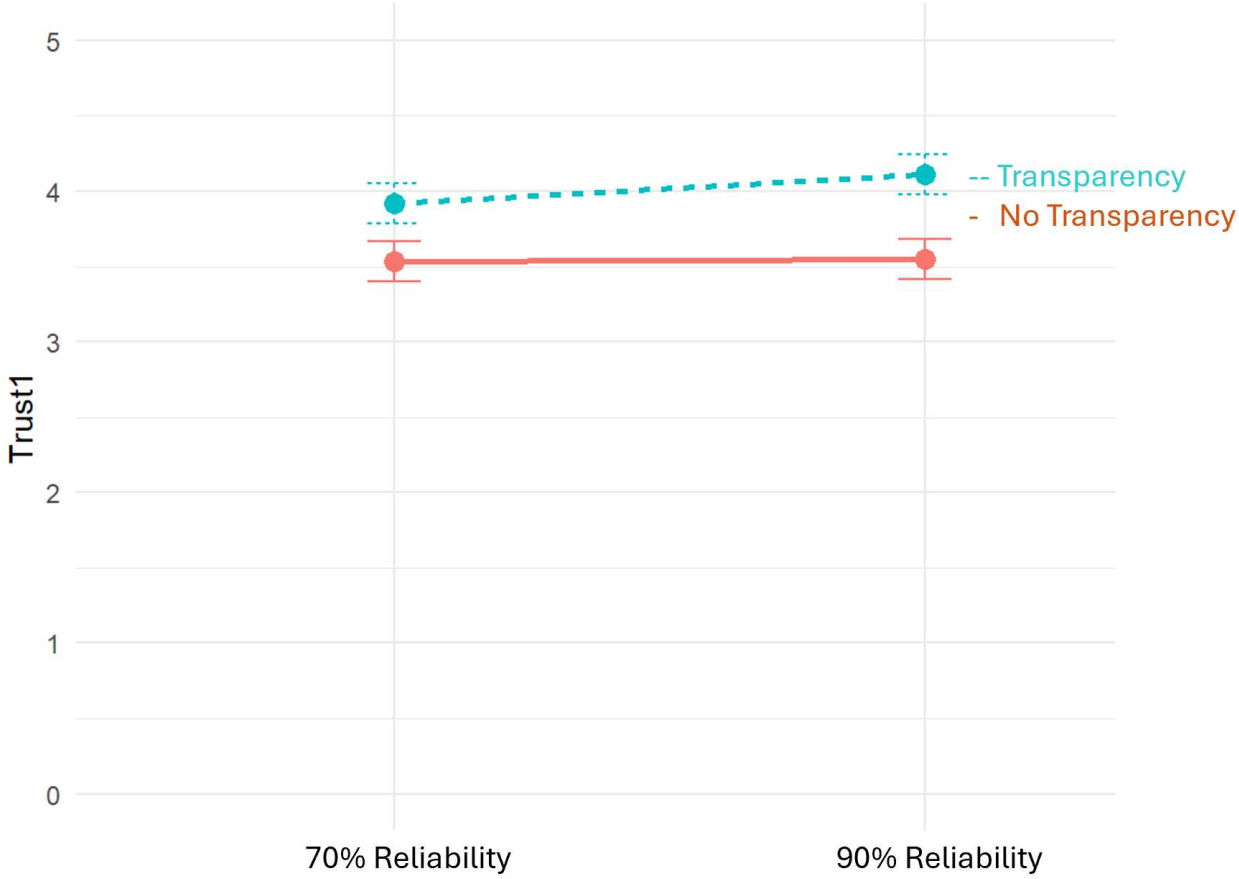

Figure 4

Trust1 as a function of reliability and transparency. Error bars represent 95% confidence intervals of mean trust1 rating.

Trust2. To test if trust2 differed among conditions, we again ran a linear mixed model. Ratings of trust2 followed an approximately normal distribution. We found that transparency increased trust2 ($x = 0.33$, $t = 7.10$, $p < .001$), no effect of reliability on trust2, ($x = 0.19$, $t = 1.82$, $p = .07$), and no interaction between reliability and transparency ($x = 0.07$, $t = 1.15$, $p = .25$). The marginal R^2 value for the model was .03 and the conditional R^2 was .31.

Participants trusted the weather forecasters more when there was transparency versus when there was not transparency. Though not quite significant, there seemed to be a trend that increasing reliability increased trust2.

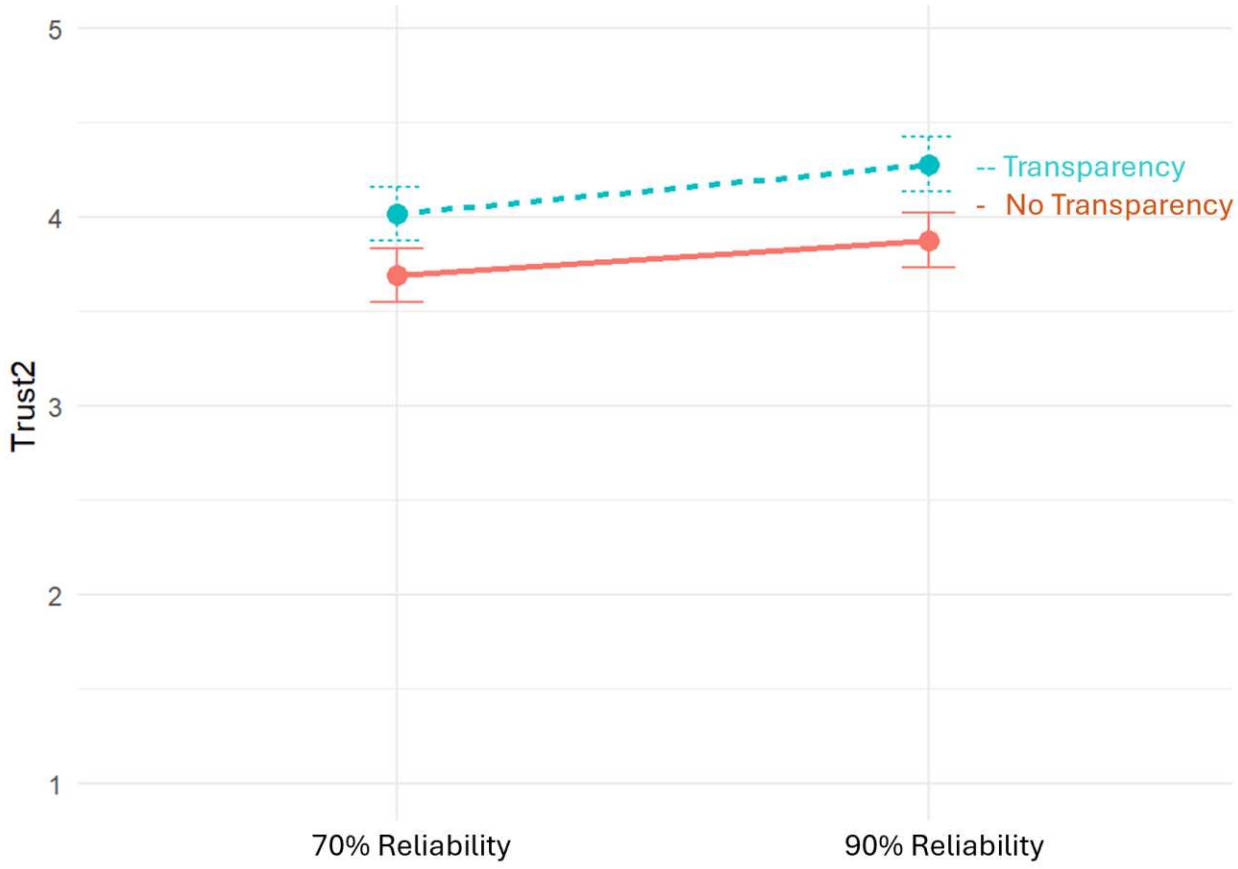

Figure 5

Trust2 as a function of reliability and transparency. Error bars represent 95% confidence intervals of mean trust2 rating.

Final Trust. To test if final trust differed among conditions, we ran a linear mixed model. The dependent measure was final trust. The dependent measure was final trust. The fixed effects were reliability, transparency, and their interaction. The random effect was the subject, as the experiment was within subjects. Ratings of final trust followed an approximately normal distribution. We found no main effect of transparency (x = -0.09, t = -0.83, p = .41), that increasing reliability increased final trust, (x = 0.35, t = 2.52, p =.01), and no interaction between

reliability and transparency ($x = 0.23$, $t = 1.47$, $p = .14$). The marginal R^2 value for the model was .06 and the conditional R^2 was .41.

Participants reported higher final trust ratings with higher reliability. At 70% reliability and 90% reliability, there was no difference in final trust when transparency was present versus when it was not.

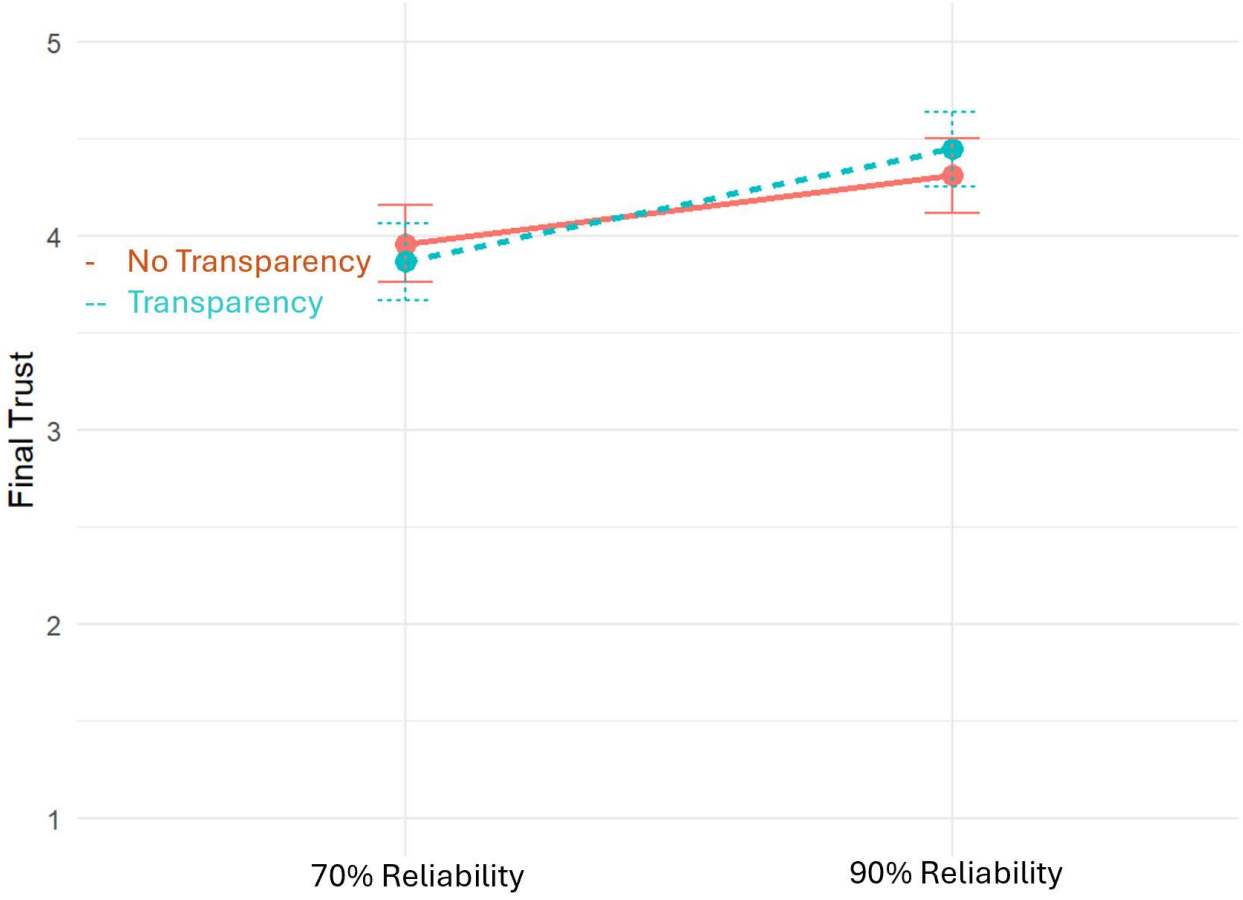

Figure 6

Final trust as a function of reliability and transparency. Error bars represent 95% confidence intervals of mean final trust rating.

Dependence

The measure of dependence is presented in Figure 7. The decision that aligned with the prediction was to decide to close the school if the prediction was 6 inches or more and to keep

the school open if the prediction was less than 6 inches. We ran a general linear mixed effects model. The dependent measure was dependence (coded as 0 when the decision was not aligned and as 1 when the decision was aligned with the prediction). The fixed effects were reliability, transparency, and their interaction. The random effect was the subject. There was no main effect of transparency ($x = 0.03$, $z = 0.22$, $p = .82$). Reliability increased dependence ($x = 0.94$, $z = 5.57$, $p < .001$). There was no significant interaction between reliability and transparency ($x = -0.37$, $z = -1.88$, $p = .06$). The marginal R^2 value for the model was .04 and the conditional R^2 was .14.

As shown in Figure 7, as reliability increased, dependence increased. However, dependence increased slightly less with increasing reliability when transparency was present, though this p-value did not reach the cut-off of .05 for statistical significance.

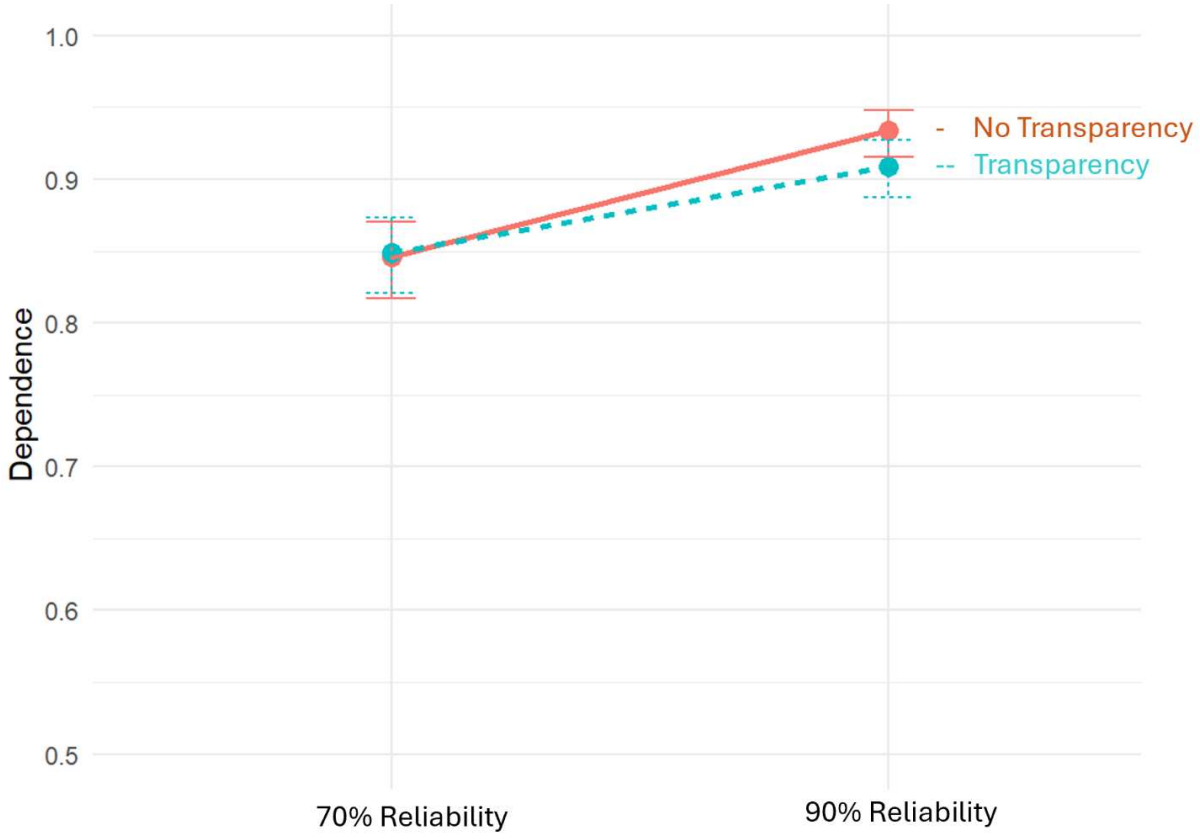

Figure 7

Dependence as a function of reliability and transparency. Error bars represent 95% confidence intervals of mean dependence.

Accuracy

The measure of accuracy is presented in Figure 8. Whereas dependence coded alignment between the decision and the automation prediction, accuracy coded alignment between the decision and the outcome (coded as 0 when the decision did not match the outcome and as 1 when the decision matched the outcome).

A general linear mixed effects model was computed with participant's mean accuracy per block as the dependent measure. The fixed effects were reliability, transparency, and their interaction. The random effect was the subject. Transparency decreased accuracy (x = -0.34, z = -3.71, $p < .001$). Reliability increased accuracy (x = 1.01, z = 9.07, $p < .001$). There was no

significant interaction between reliability and transparency (x = -0.06, z = -0.42, p = .68). The marginal and conditional R^2 values for the model were .08

As reliability increased, accuracy also increased. When transparency was present, participants were less accurate. It is also important to note that when reliability was 90%, meaning the automation was 90% accurate, participants performed significantly worse than the automation in both the transparency and non-transparency conditions (visually represented by the 95% error bars in Figure 8). In the 70% reliability condition (automation had 70% accuracy), participants performed significantly worse than the automation with transparency, but not statistically different from the automation with no transparency.

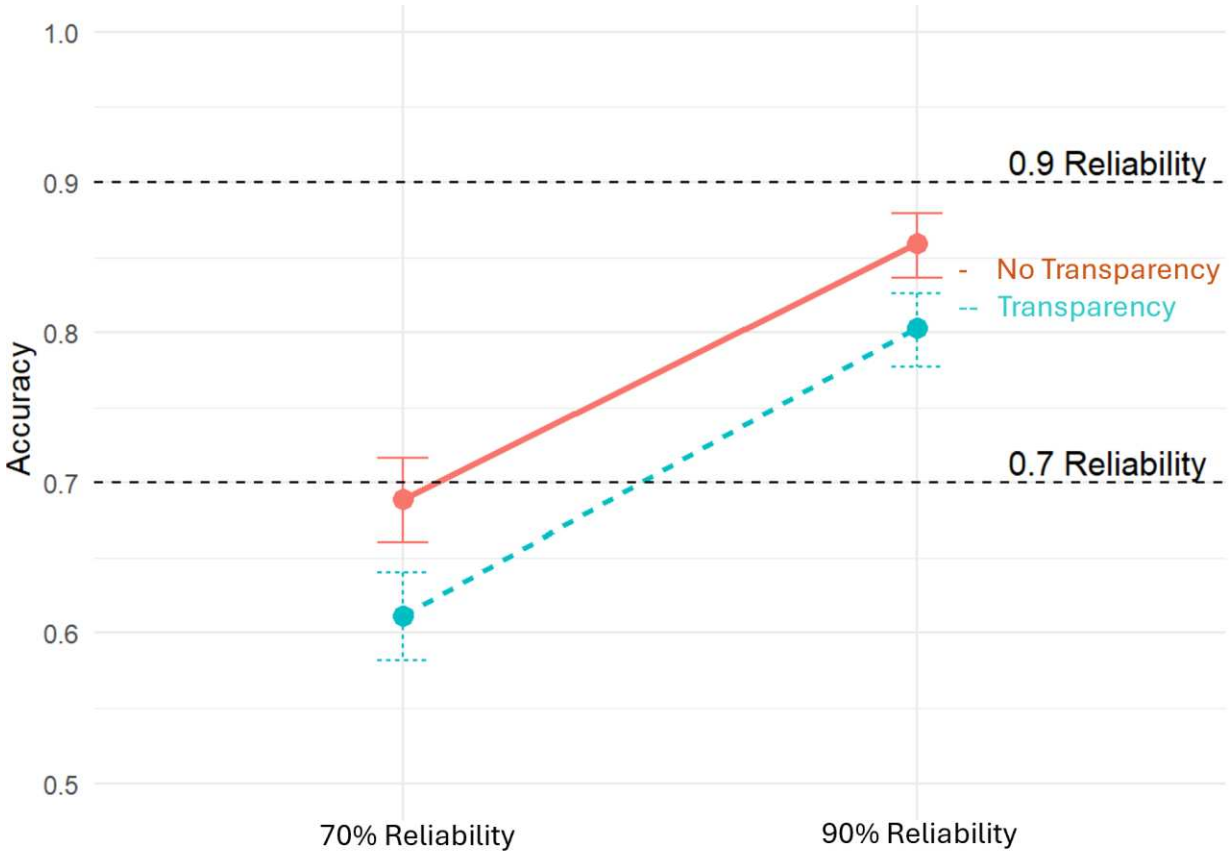

Figure 8

Accuracy as a function of reliability and transparency. Error bars represent 95% confidence intervals of mean accuracy. The horizontal dashed line at the top represents the accuracy of the automation at 90% reliability. The lower horizontal dashed line represents the accuracy of the lower reliability system (70%).

Distance from the Decision Threshold

We hypothesized that trust, dependence, and performance would increase as the deviation of the predicted amount of snowfall's distance from the decision threshold of 6 inches increase. For example, given the decision threshold of 6 inches, one may trust a prediction of 9 inches more than a prediction of 7 inches when it came to confidence that the school should be closed. Additionally, with a prediction of 7 inches, one may be more likely to make a decision that contradicts the automation or make the incorrect decision based on the outcome of snowfall. That is, people may make an underlying assumption that there is more uncertainty closer to the

decision threshold for the decision to make than for predictions that are further from the decision threshold.

For trust as a dependent measure, we ran a linear mixed model. For dependence and performance, we ran general mixed effects models. For each of these three models, the fixed effect was the absolute value of the difference between the forecasted snowfall and the decision threshold of 6 inches. The random effect was the subject for each model.

Measures of trust1, dependence, and performance compared to absolute value are shown in Figures 9, 10, and 11. As absolute value from the threshold increased, trust1 increased (x = 0.08, t = 7.53, $p < .001$). As absolute value from the threshold increased, dependence also increased (x = 0.93, z = 1371, $p < .001$). There was also an increase in performance (x = 0.63, z = 17.52, $p < .001$) as absolute value from the threshold increased. The marginal R^2 values for the models are .01, .31, and .19 respectively.

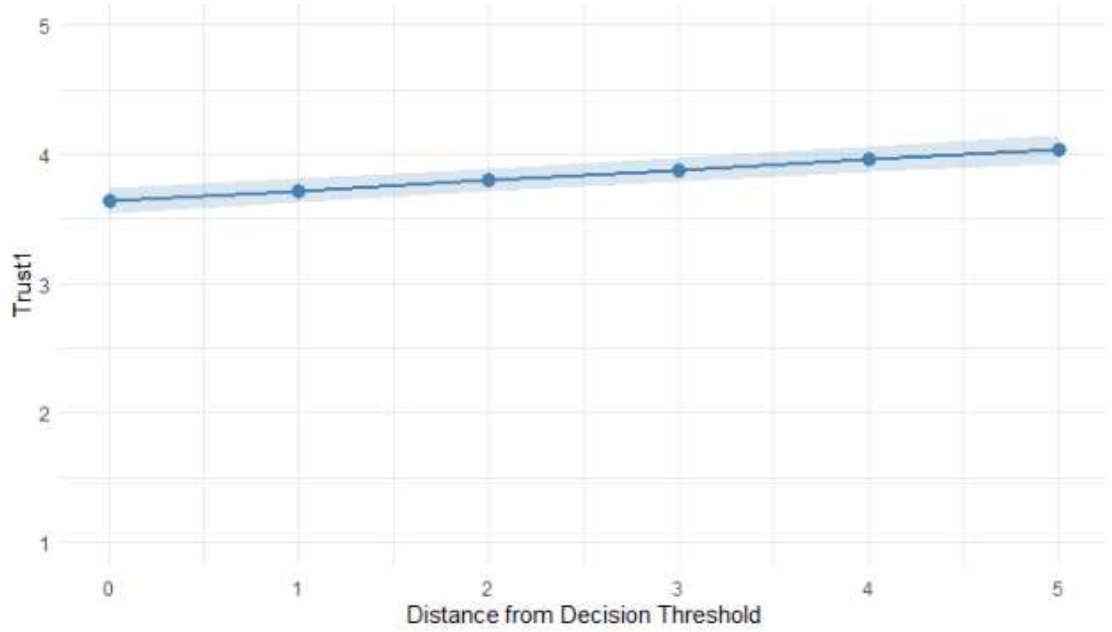

Figure 9

Trust1 as a function of absolute value of prediction distance from the decision threshold. Shading represents 95% confidence intervals.

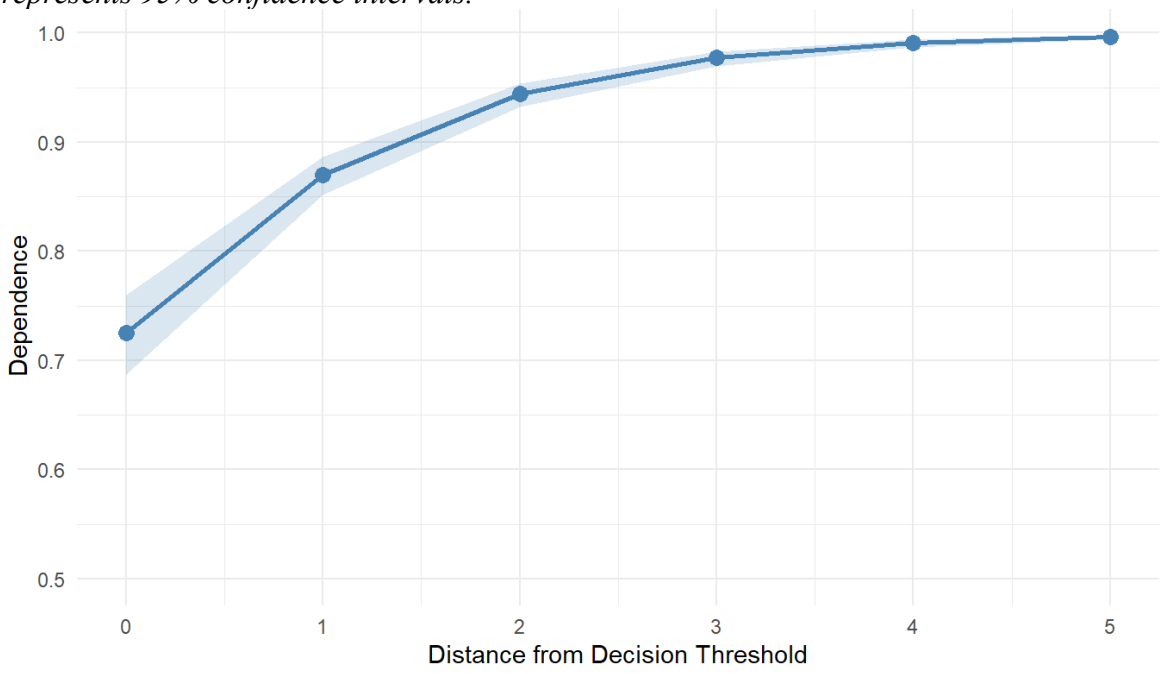

Figure 10

Dependence as a function of absolute value of prediction distance from the decision threshold. Shading represents 95% confidence intervals.

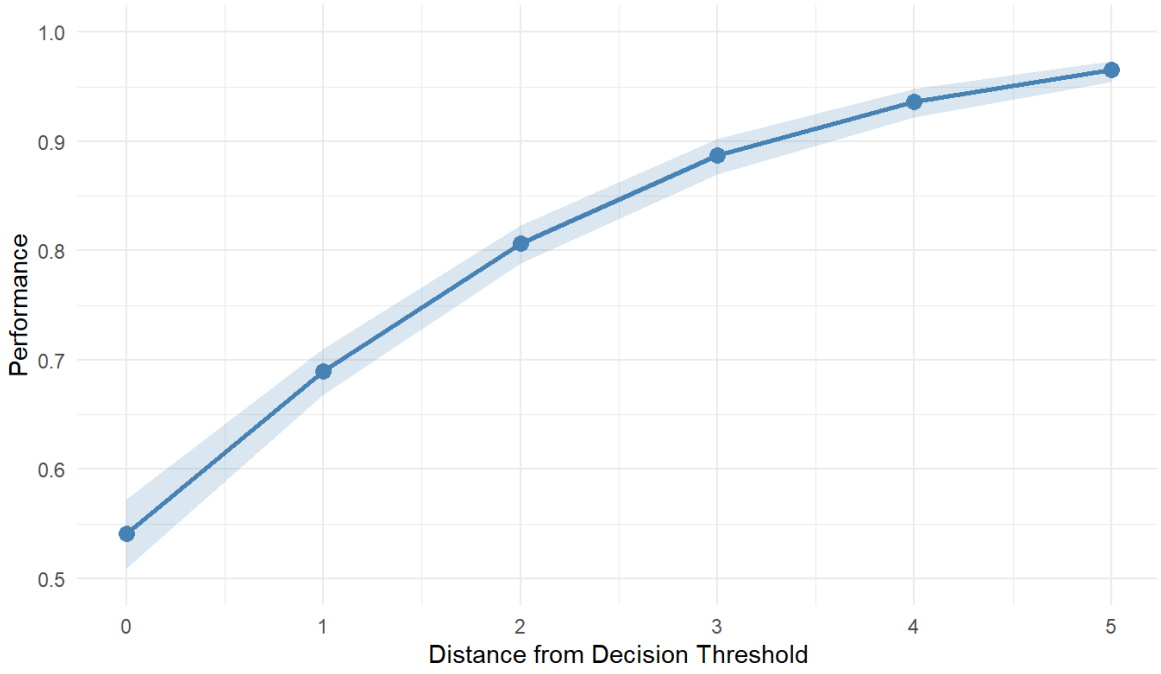

Figure 11

Performance as a function of absolute value of prediction distance from the decision threshold. Shading represents 95% confidence intervals.

Additional Survey Questions

There were five relevant questions asked at the end of the experiment. The intention with these questions was to gather further information about how participants made their decisions during the experiment.

Question 1: You were presented with two different types of weather forecasts. One only gave you a snowfall prediction and the other provided a prediction and additional information (raw data). Which forecast type did you like better?

Question number one was forced choice multiple choice with the possible responses of: "the one that only gave the prediction", "the one with additional information", and "I liked them both equally". The responses to this question are recorded in Table 4. Most participants preferred the transparency forecast type, regardless of reliability (78.11%).

Table 4

Recorded responses to which forecast type participants preferred, separated by their assigned automation reliability.

Reliability	Preferred Forecast Type	Number of participants
70	Equal	9
70	Prediction only	14
70	Transparency	75
90	Equal	9
90	Prediction only	12
90	Transparency	82

Question 2: You were presented with two different types of weather forecasts. One only gave you a snowfall prediction and the other provided a prediction and additional information (raw data). Which type of forecast did you find made more accurate predictions?

Question number two was also forced choice with the possible responses of "the one that only gave the prediction", "the one with additional information", and "I liked them both equally". The responses of which forecast was perceived as more accurate are recorded in table 5. Most participants, regardless of automation reliability, perceived that the transparency forecast type was most accurate (54.73%). However, more participants in the 90% reliability condition correctly identified that both forecast types were equally accurate (31.07%, compared to 21.43% in the 70% reliability condition).

Table 5

Recorded responses of which forecast type participants perceived as more accurate, separated by their assigned automation reliability.

Reliability	Most Accurate Forecast Type	Number of participants
70	Equal	21
70	Prediction Only	22
70	Transparency	55
90	Equal	32
90	Prediction Only	16
90	Transparency	55

Question 3: Overall, how often do you remember the forecasts making errors (making a prediction that corresponded with the incorrect decision)? Please report your response as a percentage or fraction.

Question three was a free response question. Though it was requested that participants report their answer in only a fraction or percentage, many participants did not follow directions. Upon further investigation of the dataset, participants who did not provide a numerical response were recorded as "non-answer" (n = 28). Some participants also provided a range, instead of a single value. For these participants, the midpoint of the range was used so that responses were comparable to one another. Participants were able to respond to the question in fractions or percentages, but for ease of analysis, all responses were converted to percentages. Some participants may have misunderstood the question, noting that they believed the error rate of automation to exceed 50%. Participants who reported an error rate exceeding 80% were removed from the dataset. Figure 12 depicts perceived error rate of automation, separated by the actual reliability of their automation. Participants in the 70% reliability condition did not perceive the

error rate of automation to be statistically different from 30% (x = 32.39, t = 1.29, p = .20). Participants in the 90% reliability condition overestimated the error rate of automation (x = 16.18, t = 5.60, p < .001).

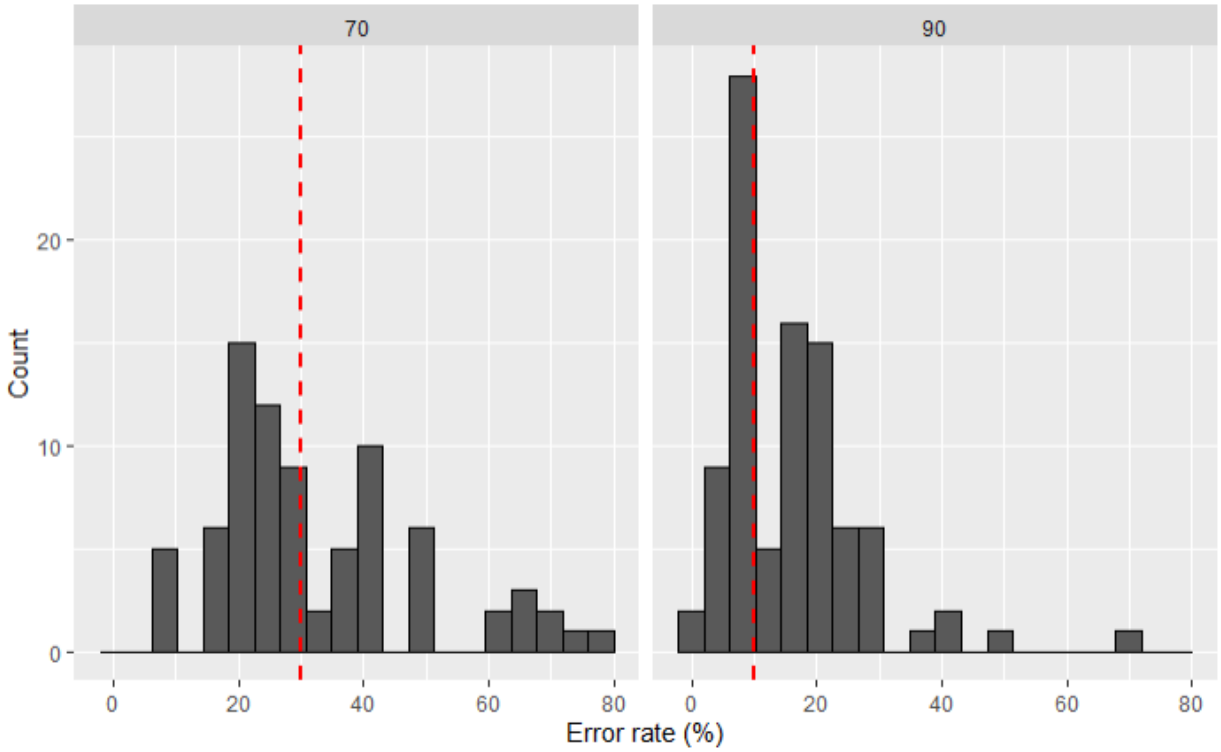

Figure 12
A histogram of perceived error rate of automation by participants, separated by the reliability of the system. The vertical dashed lines represent the actual error rate of the automated system. Question 4: Overall, how did you feel about the additional information (raw data) presented to you in the forecasts? Was the information helpful?

Responses to this question were also free response. There was significant variation in the level of detail in responses, but responses were categorized into three groups: positive, negative, and both. Responses in the positive group only had positive things to say about the automation transparency. Responses in the negative group only had negative things to say about the automation transparency. Responses in the "both" group may have stated that they felt neutral about the transparency or that they liked the transparency, but found that it was not always

accurate, which is true. Some participants with responses that were recorded as both noted that they found the transparency to be confusing.

Responses are recorded in Table 6. Participants overwhelmingly had only positive things to say about the automation transparency (79.10%), regardless of automation reliability. More participants had mixed feelings about transparency than outwardly disliking transparency (14.93% of participants had mixed feelings about transparency, while 5.97% had negative feelings).

Table 6

Responses of participants regarding how they felt about the additional information (transparency) provided to them, separated by the automation reliability group that they were assigned to.

Reliability	Feelings about Transparency	Number of participants
70	Both	16
70	Negative	8
70	Positive	74
90	Both	14
90	Negative	4
90	Positive	85

Question 5: How much do you typically trust weather forecasts to help you make decisions? Uses the same trust scale as the entire experiment.

Question five examines the baseline trust of participants to determine how naturally trusting they think they are of automation in the context of weather forecasts.

Responses are reported in Figure 13. The mean baseline trust was 4.10 (σ = 0.93), which means that participants reported that they tend to trust weather forecast quite a bit, demonstrating high baseline trust in most participants of weather forecasts.

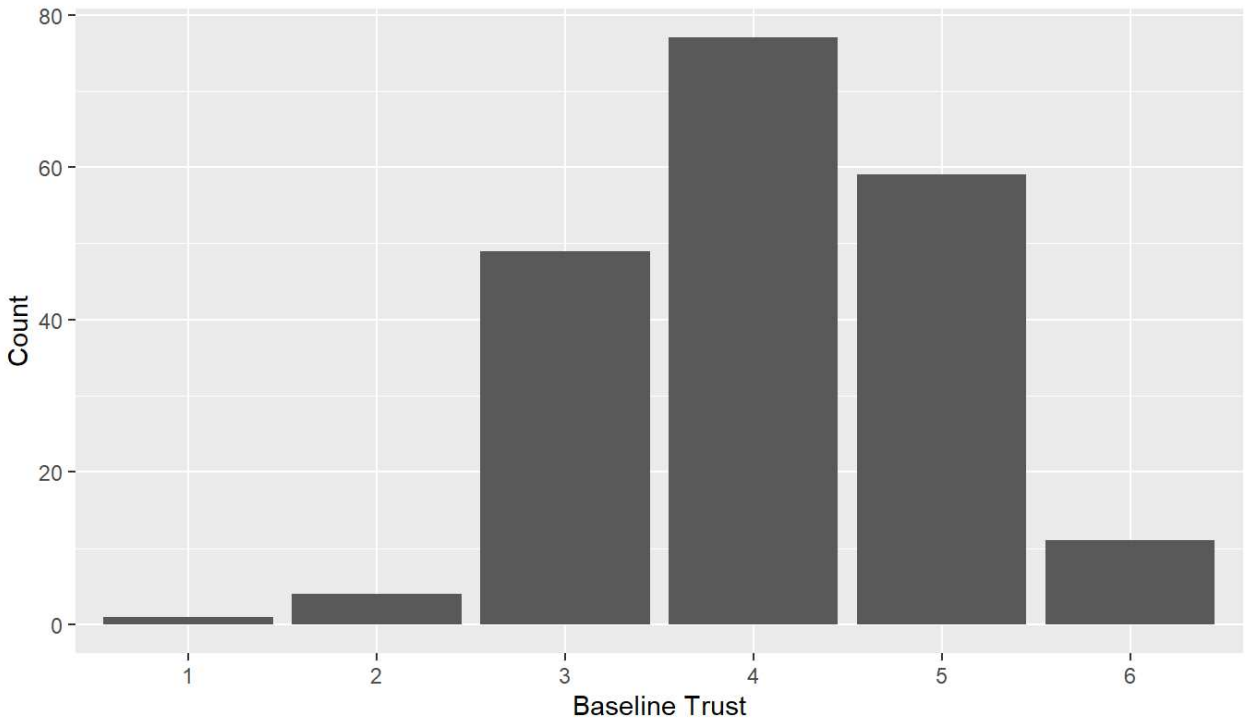

Figure 13

Distribution of participants' self-reported baseline trust in weather forecasts. Responses were recorded on a scale of 1-6, where 1 represents "Not at all", 2 represents "A little", 3 represents "Somewhat", 4 represents "Quite a bit", 5 represents "Very much", and 6 represents "Completely".

Discussion

The current research focused on how trust, performance, and dependence were influenced by transparency and reliability. Further analysis was done regarding the distance from the decision threshold and the questions asked at the end of the experiment.

Hypothesis 1

We hypothesized that participants would have increased trust, dependence, and performance as automation reliability increased. We found support for this hypothesis, as there was a positive significant main effect of reliability for final trust (Figure 6), dependence (Figure 7), and

performance (Figure 8). This finding adds to existing literature *(Pharmer et al., 2021; Holthausen & Walker, 2022; Patton & Wickens, 2024)*, that increasing reliability improves participant's attitudes towards automation, makes them want to use it more, and, as a consequence, use it more effectively.

Additionally, while participants were about as dependent as the reliability of automation at 90%, they were much more dependent on automation (about 85%) that was 70% reliable than they would have been if they were calibrated to the reliability of the system.

Hypothesis 2

We hypothesized that the implementation of transparency would yield increased trust and dependence. We found no main effect of transparency on final trust (see Figure 6) or dependence (see Figure 7), meaning that there was no direct effect of the presence of transparency on self-reported final trust towards the automation or their use of the automation. However, we found that transparency improved trust for both trust1 and trust2 (see Figures 5 and 6), meaning that participants trusted transparency predictions more in the moment, but retrospectively did not report that they trusted transparency automation more than non-transparency automation.

Transparency was not the main factor that resulted in participants trust in or use of a weather forecaster's predictions. This finding is not consistent with the current literature (Sargent et al., 2023). Though participants reported positive feelings about the automation transparency, this is not reflected in their self-reported final trust or behavioral measures of dependence.

In Experiment 1, transparency did not indicate that automation made an error and therefore was not necessarily meant to give reasons for participants to not depend on the prediction of automation. Instead, the transparency was always designed to support the prediction of the forecaster. Therefore, there was not designed to be a reason for participants to

trust or depend on the predictions that were accompanied by transparency more. The type of transparency implemented in this experiment was not beneficial for participants' use of automation or trust in automation. Transparency that does not serve to facilitate any way for participants to detect automation errors did not seem to be helpful for participants regarding final trust, but participants did trust transparency predictions more on a trial by trial basis.

Hypothesis 3

As found in Gegoff et al. (2024), we hypothesized that we would find an interaction between reliability and transparency for trust and dependence, such that adding transparency would reduce the effect of reliability on trust and dependence. In this experiment, we did not find a significant interaction between reliability and transparency for final trust (Figure 6) (p = .14) or for dependence (Figure 7) (p = .06).

When transparency does not benefit trust or dependence, its presence also does not have differing effects at differing levels of reliability for final trust or dependence. However, the interaction between reliability and transparency for dependence did approach significance (p = .06). Had this interaction been significant, the trend would suggest that transparency diminishes the increasing effect of reliability on dependence, an interaction in the direction predicted by the hypothesis (and observed in a prior unpublished study by the authors).

At high reliability, transparency decreases dependence, but at low reliability, there was no effect of transparency on dependence.

If this interaction were to be significant, it would mean that transparency could be viewed as harmful at high reliability because it decreases dependence on an automation system that is more reliable than they are on their own. However, as the goal of automation use is to depend proportional to the automation's reliability, it seems that at high reliability, participants were

closer to the optimal 90% dependence rate on automation with transparency than without. Transparency may be beneficial for calibrating to the reliability of the system.

Hypothesis 4

In accordance with findings by Boskemper and colleagues (2022) and Bartlett & McCarley (2017, 2019, 2021, 2022) we hypothesized that the performance of the human automation team would be lower than the performance of the automation alone, especially when automation is highly reliable. We found support for this finding (Figure 8). For both transparency and non-transparency conditions, at 90% reliability, participants' accuracy was significantly lower than 90% (and accuracy with transparency was significantly lower than accuracy without transparency). At 70% reliability without transparency, we found that participants accuracy was not significantly different than 70%, but with transparency, accuracy was significantly lower than 70%.

In three of the four conditions (that is, excluding 70% reliability with no transparency), it seems that participants depend on automation less than they should, or they act in disagreement with automation at the incorrect times, resulting in accuracy that is below that of the automation regarding decision making. The presence of transparency that did not indicate when automation made an error did not improve participant accuracy as transparency can be expected to do (Sargent et al., 2023).

Distance from the Decision Threshold

We believed that there would be more inherent uncertainty regarding which decision would be best when predictions of snowfall were closer to the decision threshold. Therefore, trust, dependence, and performance would all increase as absolute value of predicted snowfall from the six-inch decision threshold increased. We did find this effect for all three dependent

variables (see Figures 9, 10, and 11). This finding suggests that there is an inherent additional layer of complexity regarding participant trust, dependence, and performance being affected by transparency, reliability, and their interaction.

There is a higher possibility for error when a prediction falls near the decision threshold as there is less "room for error". A prediction of 6 inches, which falls on the decision threshold would indicate that one should close the school if snowfall is at or greater than 6 inches. However, if snowfall is just 5 inches, the decision that should be made changes. If the prediction is 3 inches, there is more room for error before the decision that should be made changes. With a prediction of 3 inches of snowfall, the school should stay open if there were 3 inches of snowfall, 1 inch, 2 inches, 4 inches, and 5 inches of snowfall. The automation would have to make a larger error for the prediction of 3 inches to be wrong compared to a prediction of 6 inches of snowfall. It is important to note that uncertainty regarding the correct decision to be made is highest closest to the decision threshold.

With this reasoning, it makes sense that participants were more trusting of predictions that were further from the decision threshold and more dependent on these predictions. Additionally, because automation failures only occurred within 2 inches of the decision threshold, it makes sense that performance increased as distance from the decision threshold increased as well.

This finding emphasizes the importance of consideration of the stimuli and context in which participants use automation to determine what their trust, dependence, and performance mean for broader application. The user of automation recognizes that the forecasted weather prediction could be incorrect when it is close to the decision threshold, but the automation does

not know this. This finding emphasizes the importance of precision in weather forecasting and the nuance that users demonstrate, but automation can lack.

Additional Survey Questions

We found in the questions asked of participants at the end of the experiment that participants preferred the forecast type that included transparency (see Table 4) and, in many cases, found this type of prediction to be more accurate (see Table 5), though both transparency and non-transparency forecasters had equal reliability. Most participants had positive feelings of transparency (see Table 6). Participants estimated the error rate of 70% reliability well on average but overestimated the error rate of 905 reliable automation (see Figure 12). We also found that participants had a high baseline trust in weather forecasts (see Figure 13).

These findings seem to conflict with the main results of the experiment. Participants did not trust automation transparency more than the lack of transparency. If anything, in the 90% reliability condition they depended on automation transparency predictions less. Participants were less accurate with automation transparency for both reliability groups. It is possible that participants liked the transparency more and it allowed them to make decisions that may not have aligned with the predictions of the weather forecasters more often, though their skepticism of the transparency was not based in detection that the prediction of the weather forecaster was incorrect. In terms of participants detecting the error rate of automation, their dependence seems to be at odds with what they determined the reliability of automation was, as they depended on the high reliability automation about 90% of the time and depended on the low reliability automation about 85% of the time. This finding does correspond with findings of Wickens et al., (2021) on sluggish beta, that users of automation are slow to calibrate to changes in reliability.

Regardless, it is important to note firstly that participants liked having the option of viewing automation transparency, even if it did not seem to improve their final trust, dependence, or performance. Second, participants did fairly well at estimating reliability of automation. Still, they overwhelmingly perceived the transparency conditions to be more reliable and some overestimated or underestimated the error rate of automation. It would be challenging for participants to trust and depend on automation effectively if they cannot detect accurately how reliable the automation is. Third, participants reported high baseline trust in weather forecasts. These participants tend to trust automation as a baseline, rather than distrust it, in the context of weather forecasts. We found that participants consistently reported high trust in automation throughout the experiment.

EXPERIMENT 2

The purpose of Experiment 2 was to determine if the type of error done by automation influenced trust, dependence, and performance. In Experiment 1, automation errors can be attributed to the nature of predicting a noisy environment such as weather forecasting. Therefore, the transparency cues matched the prediction of the forecast and participants should not have been able to determine that automation was going to make an error based on the transparency cues (the raw data of the four variables aggregated by the forecast automation). In contrast, Experiment 2 was meant to model a situation when automation errors are caused by a problem with the algorithm. In such cases, the automation errors were due to a miscalculation of the raw data that was gathered from the environment. Thus, in Experiment 2, if participants understood the transparency cues perfectly, they could determine automation errors by comparing the transparency cues to the automation prediction.

Method

Participants

163 students completed the experiment in exchange for course credit. Of the participants, 75% identified as female, 23% identified as male, and 2% identified as non-binary or did not disclose their gender. Participants had an age range of 17-38 years, with an average age of 20.

Design

Experiment 2 was conducted with nearly identical methodologies to Experiment 1. The only difference between Experiment 1 and Experiment 2 is in the relationship between the transparency cues and the forecast versus the outcome. For Experiment 1, the raw data presented with the prediction of snowfall matched prediction of snowfall. For Experiment 2, the raw data presented with the prediction of snowfall matched the outcome snowfall and may have been at odds with the forecasted prediction. That is, the transparency in Experiment 2 could be used to detect an automation error if participants used interpreted the raw data in the way that it was designed. The errors by automation and transparency used in Experiment 2 represent an instance where the algorithm that the automation used was incorrect, rather than the errors due to a noisy environment as seen in Experiment 1. Because of these differences from Experiment 1 (errors due to a noisy environment) in Experiment 2 (algorithmic error), we predict:

1. Participants will trust automation transparency predictions more in Experiment 2 (than in Experiment 1) because the transparency in Experiment 2 allows participants to detect automation errors.
2. Participants will depend on automation with transparency less in Experiment 2 (compared to Experiment 1) because they will find that the automation transparency is not aligned with the forecasted weather predictions.

3. Participants will be more accurate with transparency in Experiment (compared to Experiment 1) because they will be able to determine when a forecast prediction will result in an error.

Results

Of the 163 participants who completed the experiment, 161 were included due to the removal of two outliers. Participants were removed from the dataset if their subject-level accuracy rating was less than or equal to 50%, meaning that their accuracy was equivalent to or worse than chance.

The methods of analysis for Experiment 2 were identical to those done for Experiment 1. Linear mixed effects models and general linear models were computed for each type of analysis, where the random effect was the subject due to transparency being manipulated within subjects for the experiment.

The following analyses were done in R Studio, the packages that were used were identical to those used in Experiment 1 and can be found in Appendix A.

Trust

Measures of trust are presented in Figures 14, 15, and 16. Trust1 represents the first time that trust was taken, after participants were presented with the prediction of snowfall for the following day. Trust2 was taken after participants decided whether to close the school and were presented with the outcome snowfall. Final trust was taken at the end of each block. Trust was self-reported using a drop-down scale that ranged from 1-6.

Trust1. Trust1 ratings followed an approximately normal distribution. We found that transparency ($x = 0.62$, $t = 13.11$, $p < .001$) and reliability ($x = 0.22$, $t = 2.08$, $p = .04$) increased

trust. There was no interaction between reliability and transparency (x = 0.006, t = -0.09, p =.93). The marginal R^2 value for the model was .08 and the conditional R^2 was .34.

Participants trusted forecasted weather predictions more when there was transparency included in the prediction. Trust1 was also higher for participants in the higher (90%) reliability group.

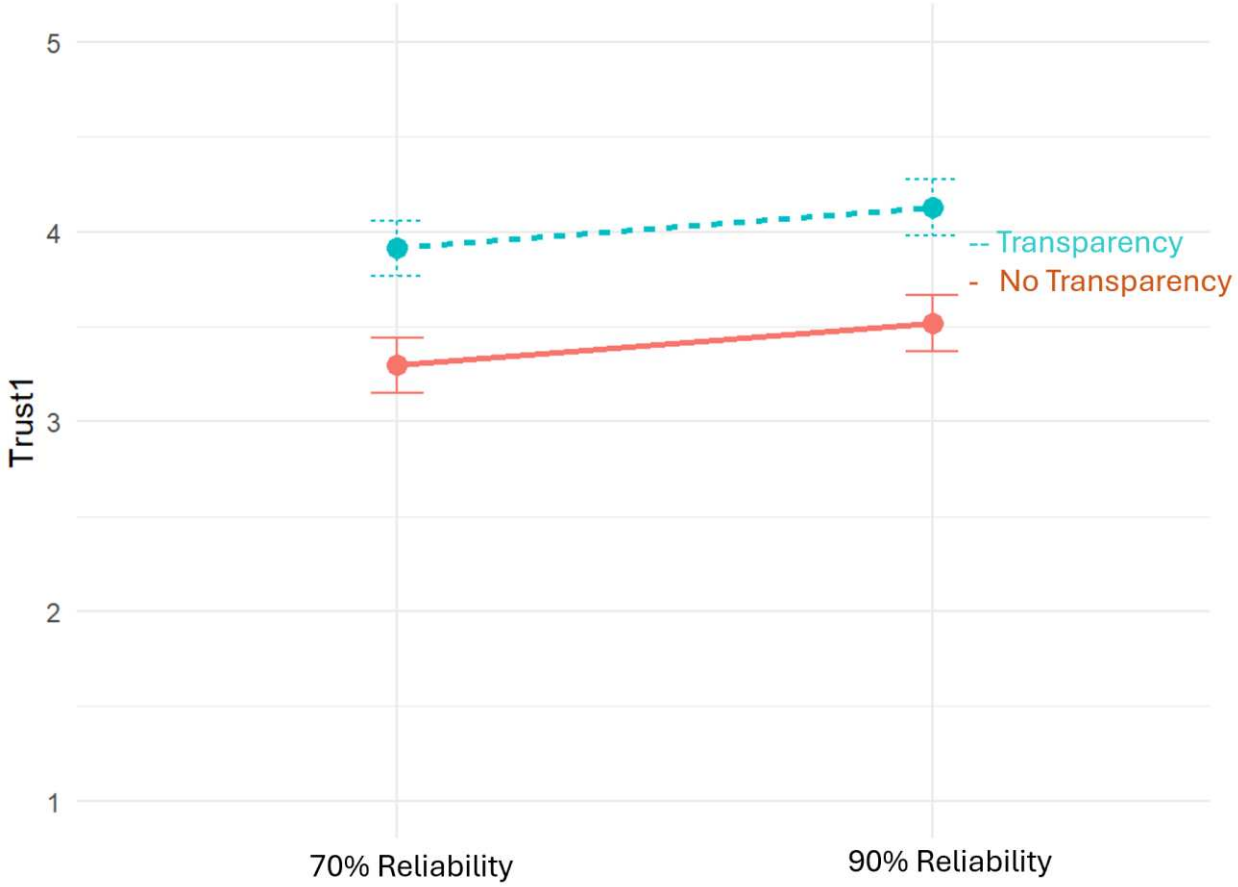

Figure 14

Trust1 as a function of reliability and transparency. Error bars represent 95% confidence intervals of mean trust1 rating.

Trust2. For trust2, we found that transparency (x = 0.52, t = 10.02, p < .001) and reliability (x = 0.36, t = 2.93, p =.003) increased trust, but that there was no interaction between reliability and transparency (x = 0.02, t = 0.27, p =.79). The marginal R^2 value for the model was .06 and the conditional R^2 was .35.

With the same pattern of results as trust1, participants reported that they trusted weather forecasts to make their decision more if it was accompanied by transparency and if reliability was high.

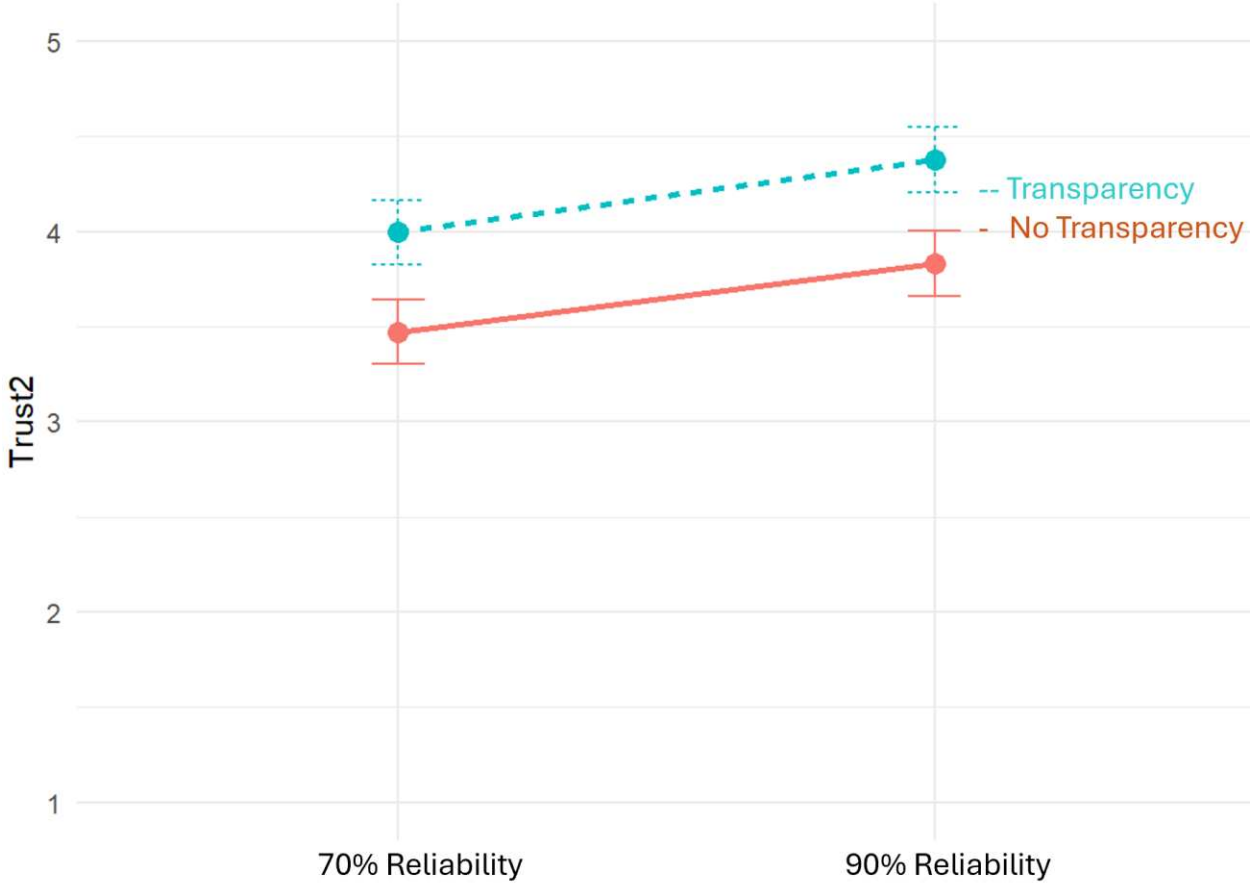

Figure 15

Trust2 as a function of reliability and transparency. Error bars represent 95% confidence intervals of mean trust2 rating.

Final Trust. Final trust ratings followed an approximately normal distribution. We found that transparency ($x = 0.45$, $t = 3.74$, $p < .001$) and reliability ($x = 0.84$, $t = 5.57$, $p < .001$) increased final trust and that there was a significant interaction between transparency and reliability ($x = -0.42$, $t = -2.43$, $p = .02$). The marginal R^2 value for the model was .12 and the conditional R^2 was .44.

We found that reliability and transparency both increase final trust and a significant interaction which means that the presence of transparency reduces the benefit of high reliability on final trust. The benefits of transparency were only evident when reliability was lower.

Figure 16

Final trust as a function of reliability and transparency. Error bars represent 95% confidence intervals of mean final trust rating.

Dependence

The measure of dependence is presented in Figure 17. We found that transparency decreased dependence (x = -0.53, z = -4.82, p < .001), but reliability increased dependence (x = 0.92, z = 5.27, p < .001). There was not a significant interaction between reliability and transparency (x = 0.25, z = 1.11, p = 0.27). The marginal R^2 value for the model was .09 and the conditional R^2 was .10.

We found significant main effects of transparency and reliability, that increasing reliability increased dependence, but that transparency being present decreased dependence. That is, the presence of transparency caused participants to go against the prediction of the automation more often. Though there was not a significant interaction between reliability and transparency on dependence, closer examination reveals that there is no statistical difference between transparency being present and not at 90% reliability, but at 70% reliability, dependence was lower with transparency than without transparency.

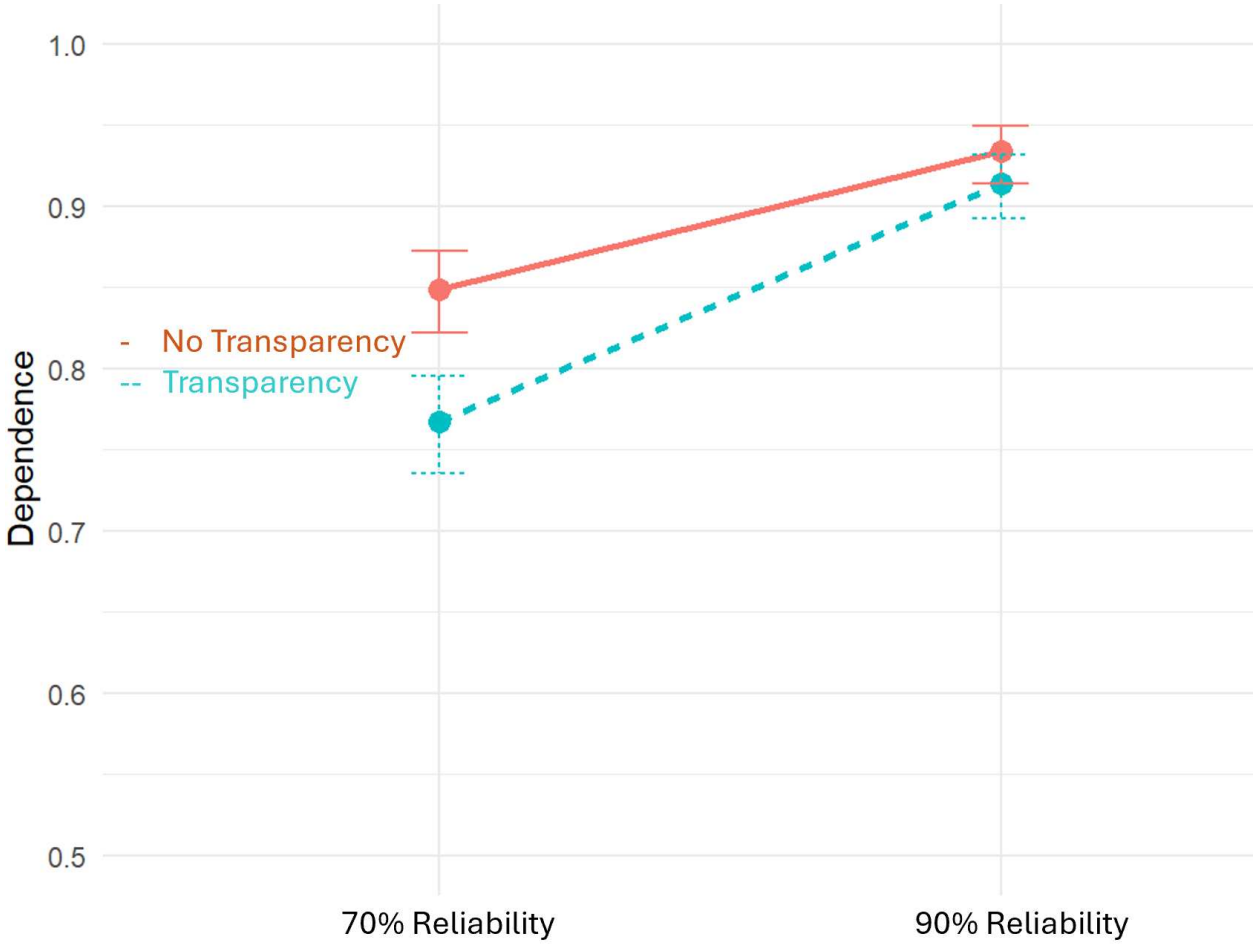

Figure 17

Dependence as a function of reliability and transparency. Error bars represent 95% confidence intervals of mean dependence.

Accuracy

The measure of accuracy is presented in Figure 18. We found no significant main effect of transparency ($x = 0.11$, $z = 1.05$, $p = .29$), but that reliability increased accuracy ($x = 1.12$, $z = 8.91$, $p < .001$). We did not find a significant interaction between reliability and transparency ($x = -0.25$, $z = -1.42$, $p = .16$). The marginal and conditional R^2 values for the model were .07.

We found a main effect of reliability, that increasing reliability increased accuracy. Transparency did not have a significant effect on accuracy, meaning it did not make participants more or less accurate, regardless of the reliability of automation. It is important to also note that at 90% reliability (automation was 90% accurate), participants in both transparency and non-transparency conditions had significantly lower accuracy than the automation. However, at 70% reliability (automation was 70% accurate), participants were not more accurate, but not significantly less accurate than the automation.

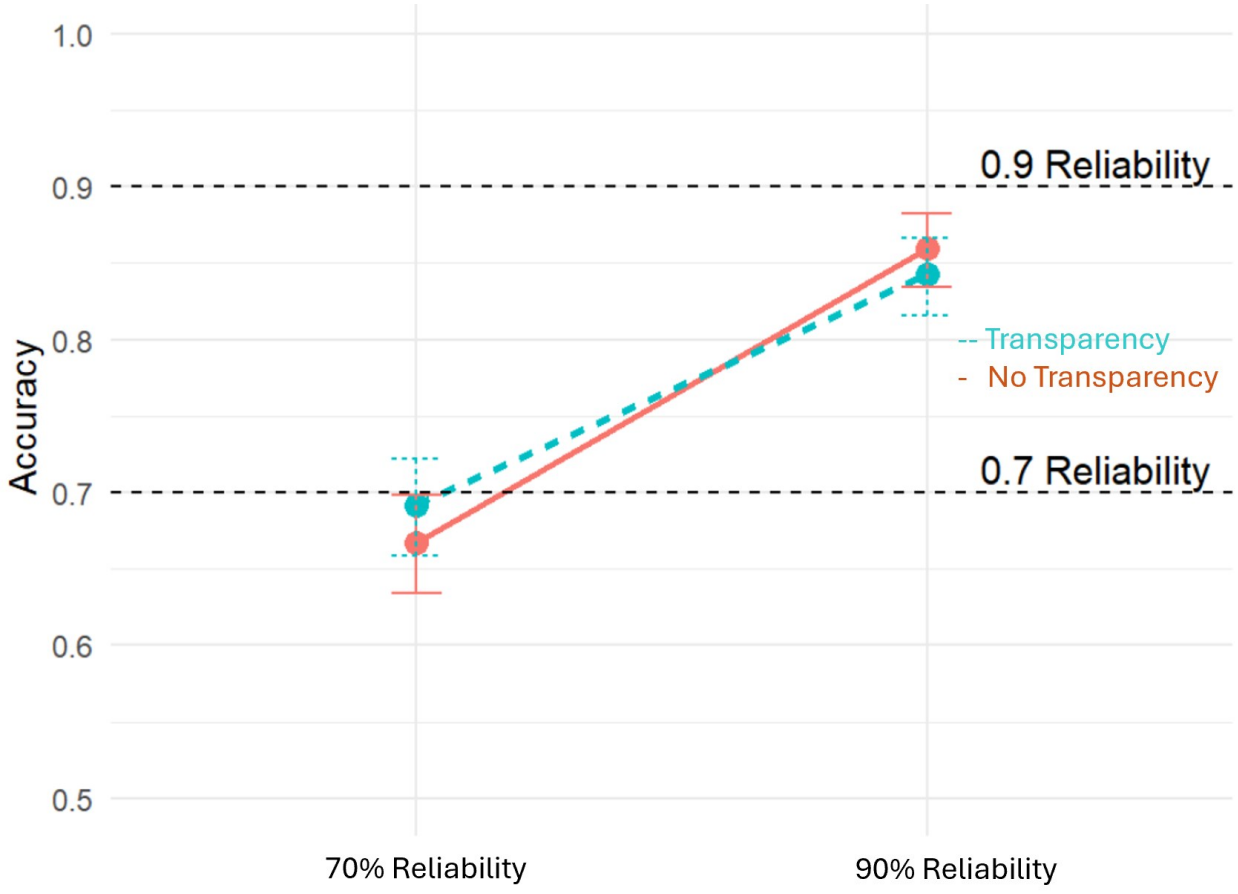

Figure 18

Accuracy as a function of reliability and transparency. Error bars represent 95% confidence intervals of mean accuracy. The horizontal dashed line at the top represents the accuracy of the automation at 90% reliability. The lower horizontal dashed line represents the accuracy of the lower reliability system (70%).

Distance from the Decision Threshold

As in Experiment 1, we hypothesized that trust, dependence, and accuracy would differ by the deviation of the predicted amount of snowfall's distance from the decision threshold of 6 inches. We hypothesized that each outcome variable would increase as distance from the decision threshold increased.

Measures of trust1, dependence, and accuracy compared to the distance from the decision threshold are shown in Figures 19, 20, and 21. As distance from the decision threshold increased, trust1 increased ($x = 0.13$, $t = 10.49$, $p < .001$), dependence increased ($x = 1.16$, $z = 15.79$, $p <$

.001), and accuracy increased ($x = 0.62$, $z = 14.72$, $p < .001$). The marginal R^2 values were respectively .02, .42, and .18.

We found support of our hypothesis that as distance from the decision threshold increased, trust, dependence, and accuracy would also increase.

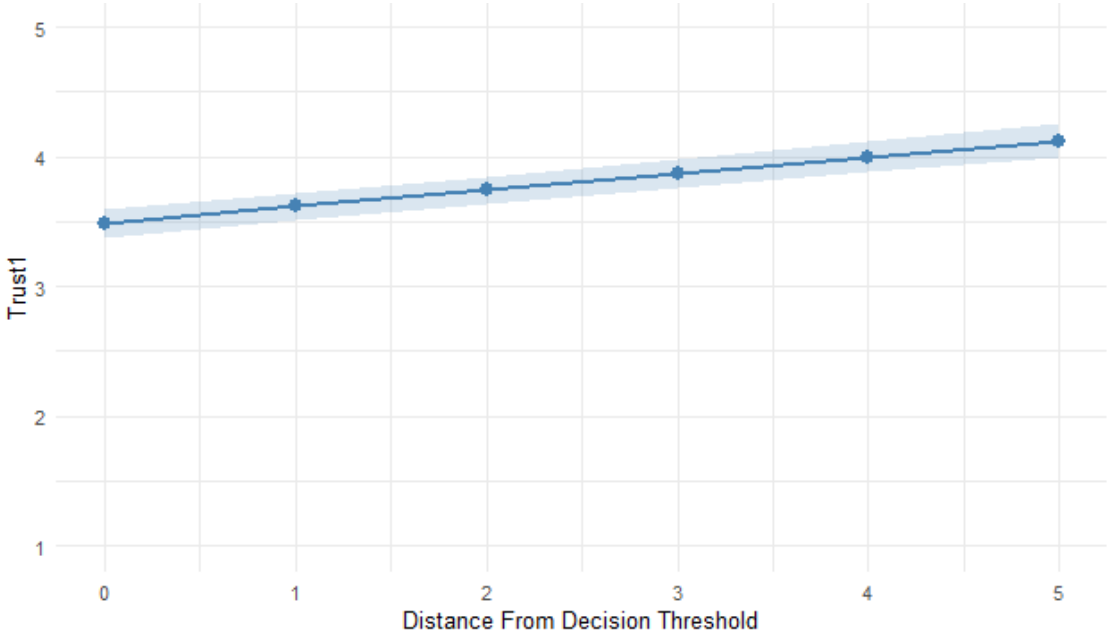

Figure 19

Trust1 as a function of absolute value of prediction distance from the decision threshold. Shading represents 95% confidence intervals.

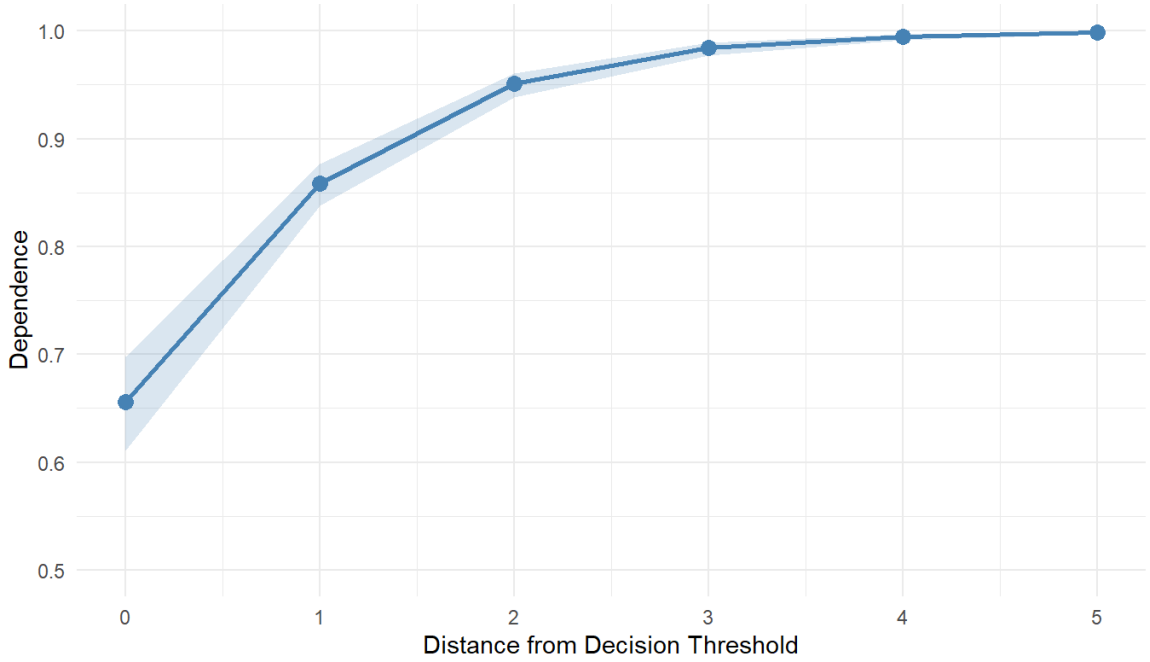

Figure 20

Dependence as a function of absolute value of prediction distance from the decision threshold. Shading represents 95% confidence intervals.

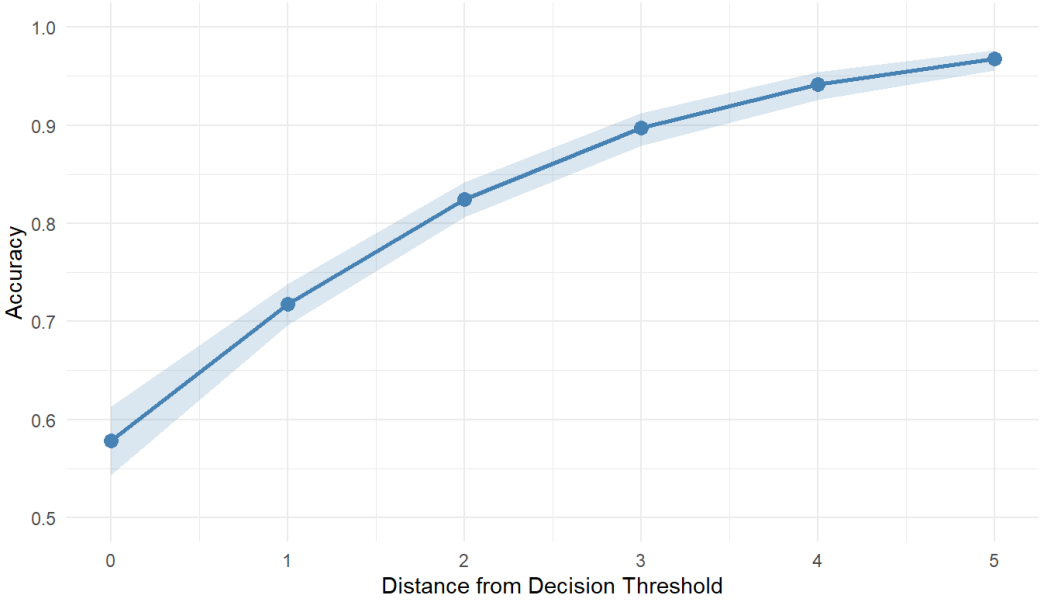

Figure 21

Performance as a function of absolute value of prediction distance from the decision threshold. Shading represents 95% confidence intervals.

Additional Survey Questions

As with Experiment 1, five questions were asked of participants at the end of the experiment to gather more information about why participants' results demonstrated the patterns that they did.

Question 1: You were presented with two different types of weather forecasts. One only gave you a snowfall prediction and the other provided a prediction and additional information (raw data). Which forecast type did you like better?

The first question was forced multiple choice with response options of: "the one that only gave the prediction", "the one with additional information", and "I liked them both equally". The responses to this question are recorded in Table 7. Participants preferred the forecast that had transparency (83.85%).

Table 7

Recorded responses to which forecast type participants preferred, separated by their assigned automation reliability.

Reliability	**Preferred Forecast Type**	**Number of Participants**
70	Equal	5
70	Prediction only	5
70	Transparency	71
90	Equal	7
90	Prediction only	9
90	Transparency	64

Question 2: You were presented with two different types of weather forecasts. One only gave you a snowfall prediction and the other provided a prediction and additional information (raw data). Which type of forecast did you find made more accurate predictions?

The second question was also forced choice with possible response options of "the one that only gave the prediction", "the one with additional information", and "I liked them both equally". The responses of which forecast was perceived as more accurate are recorded in Table 8. We found that regardless of reliability, the majority of participants perceived that the transparency forecast type was most accurate (62.73%), though both forecast types were equal in accuracy. More participants in the 90% reliability condition identified that both forecast types were equally accurate (33.75%, compared to 16.05% in the 70% reliability condition).

Table 8

Recorded responses to the forecast type participants perceived as more accurate, separated by their assigned automation reliability.

Reliability	Most Accurate Forecast Type	Number of Participants
70	Equal	11
70	Prediction only	4
70	Transparency	66
90	Equal	10
90	Prediction only	7
90	Transparency	63

Question 3: Overall, how often do you remember the forecasts making errors (making a prediction that corresponded with the incorrect decision)? Please report your response as a percentage or fraction.

Question three was a free response question. Upon further investigation of the dataset, participants who did not provide a numerical response were recorded as "non-answer" (n = 6). Some participants also provided a range, instead of a single value. For these participants, the mid-point of the range was reported so that responses were comparable to one another.

Participants were able to respond to the question in fractions or percentages, but for ease of analysis, all responses were converted to percentages. Figure 22 demonstrates perceived error rate of automation, separated by the actual reliability of their automation.

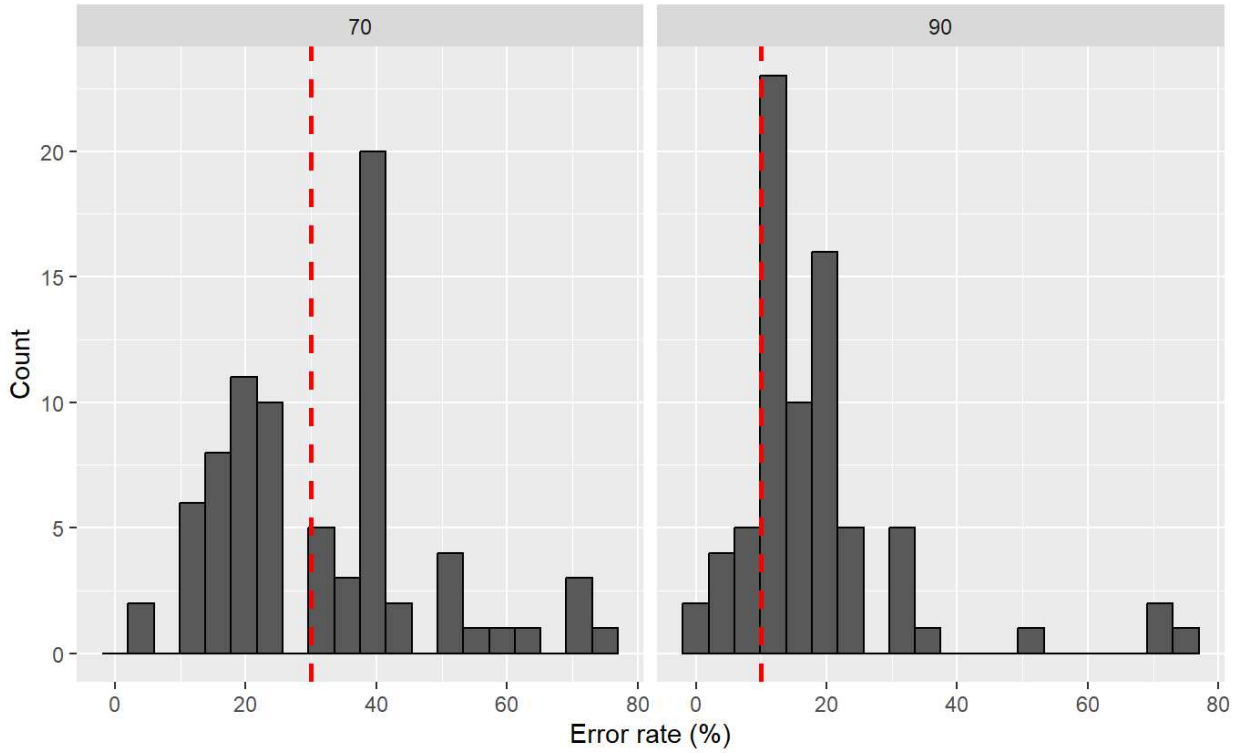

Figure 22

A histogram of perceived error rate of automation by participants, separated by the reliability of the system. The vertical dashed lines represent the actual error rate of the automated system.

Some participants may have misunderstood the question, noting that they believed the error rate of automation to exceed 50%. Participants who reported an error rate exceeding 80% were removed from the dataset. In the 70% reliability condition, participants estimated the error rate of automation statistically equal to its actual error rate of 30% ($x = 31.76$, $t = 0.96$, $p = .34$). Participants in the 90% reliability condition overestimated the error rate of automation of 10% ($x = 17.84$, $t = 4.85$, $p < .001$).

Question 4: Overall, how did you feel about the additional information (raw data) presented to you in the forecasts? Was the information helpful?

Question four was also recorded via free response. There was variation in the detail provided in these responses, so responses were categorized into three groups: positive, negative, and both (meaning positive and negative). Responses in the "positive" group only had positive things to say about the automation transparency. Responses in the "negative" group only had negative things to say about the automation transparency. Responses in the "both" group may have stated that they felt neutral about the transparency or that they liked the transparency but found that it was not always accurate.

Responses are presented in Table 9. Most participants only reported positive feelings about transparency (80.12%), regardless of automation reliability.

Table 9

Responses from participants regarding how they felt about the additional information (transparency) provided to them, separated by the automation reliability group that they were assigned to.

Reliability	Feelings about Transparency	Number of Participants
70	Both	11
70	Negative	4
70	Positive	66
90	Both	10
90	Negative	7
90	Positive	63

Question 5: How much do you typically trust weather forecasts to help you make decisions? Uses the same trust scale as the entire experiment.

Question five aims to detect baseline trust that participants generally have in weather forecasts. Responses are reported in Figure 23. The mean baseline trust was 3.94 (σ = 0.93),

which means that participants reported that they tend to trust weather forecasts slightly less than quite a bit, demonstrating a high baseline trust in most participants of weather forecasts.

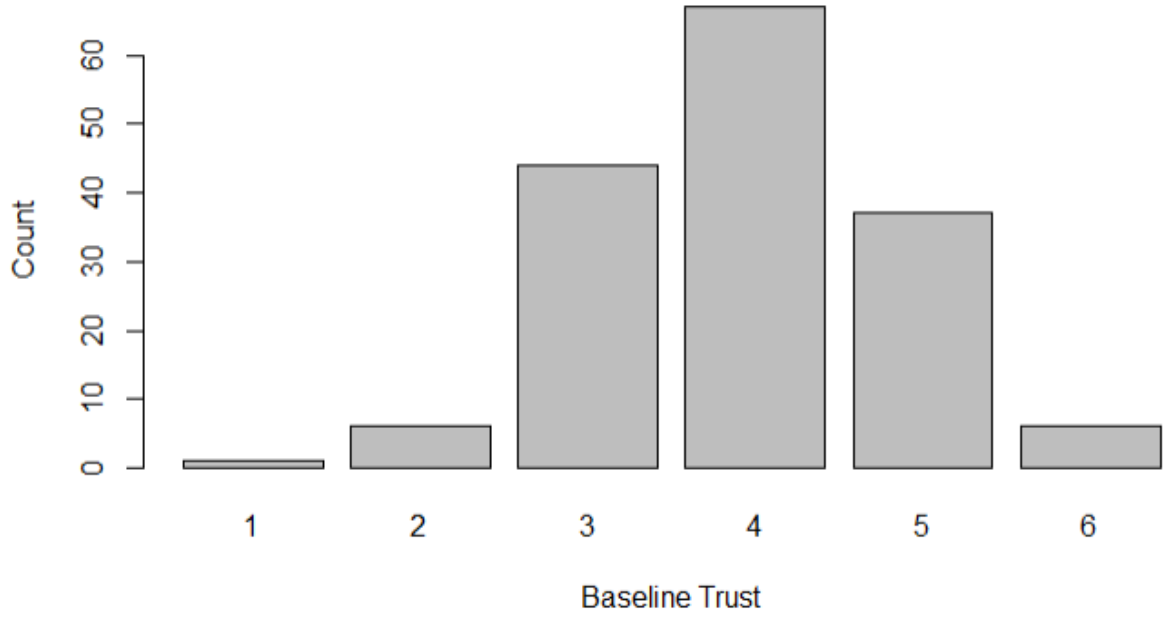

Figure 23

Distribution of participants' self-reported baseline trust in weather forecasts. Responses were recorded on a scale of 1-6, where 1 represents "Not at all", 2 represents "A little", 3 represents "Somewhat", 4 represents "Quite a bit", 5 represents "Very much", and 6 represents "Completely".

Discussion

Experiment 2, like Experiment 1, focused on how trust, performance, and dependence were influenced by transparency and reliability. Further post-hoc examination was done regarding the distance from the decision threshold.

Hypothesis 1

We hypothesized that participants would have increased trust, dependence, and performance as automation reliability increased. We found support for this hypothesis, as there was a significant main effect of reliability for these three variables (see Figures 14, 15, 16, 17, and 18). This finding aligns with existing literature (Pharmer et al., 2021; Holthausen & Walker,

2022; Patton & Wickens, 2024), that higher reliability as a property of automation increases participant trust, dependence, and performance with automation.

Additionally, while participants were about as dependent as the reliability of automation at 90%, they were much more dependent on automation (about 82%) that was 70% reliable than they would have been if they were calibrated to the reliability of the system.

Hypothesis 2

We also hypothesized that transparency would increase trust and dependence. We found that transparency increased trust for trust1, trust2, and final trust (see Figures 14, 15, and 16). For dependence, we found that transparency decreased dependence (see Figure 17). It is important to consider that, graphically, the effects of transparency on both trust and dependency appear to be present only at 70% reliability and not at 90% even as the interaction between transparency and reliability was not statistically significant.

We expected that the type of transparency present in Experiment 2 would result in decreased dependence with transparency, but higher trust. Transparency that indicates when automation errs should decrease dependence because this raw data indicates that something is wrong with the automation's prediction and therefore one should not depend on the forecast. However, the transparency being available to provide this insight would have increased trust in automation based on existing literature that transparency increases trust (Sargent et al., 2023), though trust may have decreased in its predictions alone.

Hypothesis 3

Found in Gegoff et al. (2024), we hypothesized that we would find an interaction between reliability and transparency for trust and dependence. For final trust, we found this interaction to be significant (see Figure 16), that transparency reduced the effect of reliability on

final trust. We did not find an interaction of reliability and transparency for dependence (see Figure 17). Though participants trusted transparency predictions more, they depended on them less. This could mean that participants may not have necessarily trusted the predictions of weather forecasts more when transparency was present, but they trusted the transparency cues more and the automation prediction less, resulting in lower dependence with transparency at low reliability. As stated before, to make up for the increased number of errors that automation made at 70% reliability, participants likely felt more skeptical of automation more frequently and disagreed with automation more when transparency cues were available.

We found sufficient evidence to conclude that transparency reduces the effect of reliability on trust, but not that there is a moderated effect of transparency and reliability on dependence, although the data in Figures 16 and 17 are trending in this direction.

Hypothesis 4

In accordance with findings by Boskemper and colleagues (2022), Bartlett & McCarley (2017, 2019, 2021, 2022), and Munoz Gomez Andrade and colleagues (2025) we found that at high reliability, participant accuracy was significantly lower than the automation (90%) for both transparency and non-transparency conditions (see Figure 18). At 70% reliability, participant accuracy was not statistically different from the automation. Participants seemed to not be using highly reliable automation as often as they should, to achieve its 90% accuracy, but they were about as accurate as the lower reliability automation. Presenting transparency did not allow participants to detect when the automation was going to err, in this case. This is likely due to participants not interpreting transparency cues the same way that they were designed because the cues were presented ambiguously. The lower dependence on low reliability for participants with transparency did not result in the detection of automation errors often enough that their accuracy

was overall better than the automation. Additionally, participants were not dependent on the automation enough to receive the benefit of high reliability automation.

Distance from the Decision Threshold

We hypothesized that there would be more underlying uncertainty with deciding to close the school or remain open when the forecasted snowfall was close to the decision threshold due to the lack of "room for error" as distance from the decision threshold decreased. Therefore, trust, dependence, and performance would all increase as the absolute value of predicted snowfall from the six-inch decision threshold increased. We found this effect for all three dependent variables.

Participants were especially less likely to trust automation, depend on automation, or make the correct school closure decision when the predicted amount of snowfall was 6 inches (see Figures 19, 20, and 21). Predictions that are on the cusp of a different decision are fragile in the sense that a slight deviation of the prediction would result in the opposite decision. The decision threshold in the current experiment is arbitrary and possibly limiting in the context of the experiment.

Automation operates based on algorithms and is quick to categorize suggested decisions to fit inside those categories. However, a human looking at a predicted decision takes in more information, such as the distance from the decision threshold. A prediction of 6 inches could be perceived to have little room for error by a user of automation. Automation is the most helpful when it performs better than a human could (Wickens et al., 2023), and when automation is less apt at detecting the lack of room for error, participants seem to trust it less, depend on it less, and be less accurate in use of automation because in this instance, users were better at detecting that there was little room for error in the forecasted predictions.

Additional Survey Questions

We found in the questions asked of participants at the end of the experiment that participants preferred the forecast type that included transparency (see Table 7) and, in most cases, found this type of prediction to be more accurate (see Table 8), though both transparency and non-transparency forecasters had equal reliability. Most participants had positive feelings of transparency (see Table 9). Participants estimated the error rate of 70% reliability well on average but overestimated the error rate of 905 reliable automation (see Figure 22). We also found that participants had a high baseline trust in weather forecasts (see Figure 23).

Though participants reported that they liked transparency more than no transparency, we did not always find that they trusted transparency conditions more and we found that they often depended on automation less when it had transparency. Participants may have reported that they liked the presence of transparency because it allowed them to depend strictly on the predictions of automation less. Though participants were able to detect the reliability of the automation in hindsight, their dependence tended to be higher than the reliability of the automation at 70% reliability. Participants also reported a fairly high baseline trust in automation, which was represented in their fairly high averages of trust regardless of the reliability of automation and presence of transparency.

GENERAL DISCUSSION

A summary of the findings in Experiment 1 and Experiment 2 can be found in Appendix B.

In the current experiments, we set out to determine the effects of high (90%) versus low (70%) reliability and transparency on trust, dependence, and accuracy. Between experiments, we were also interested in the effect of type of automation error on these variables. We proposed

three main hypotheses that were applicable to both experiments based on the existing literature regarding relationships between these variables. For H1, based on consistent findings in published studies, we hypothesized that increasing reliability would increase trust, dependence, and accuracy (Pharmer et al., 2021; Holthausen & Walker, 2022; Wickens et al., 2023; Patton & Wickens, 2024). Based on the meta-analysis of Sargent and colleagues (2023), which found that trust and dependence increased with the presence of transparency, we expected to find this effect of transparency, hypothesized in H2. In accordance with the work of Gegoff et al., (2024), we expected to find an interaction between reliability and transparency, such that transparency would increase trust and dependence more at low reliability (H3). Additionally, we expected to find differences in the outcome variables of trust, dependence, and accuracy from Experiment 1 (where automation errors resulted from the challenge of predicting a noisy weather environment) to Experiment 2 (where automation errors resulted from an algorithmic error by the automation and users could subsequently use transparency to determine if automation was going to make an error). Specifically, compared to Experiment 1, we expected higher trust in transparency in Experiment 2 (H4A), lower dependence with transparency in Experiment 2 (H4B), and higher accuracy with transparency in Experiment 2 (H4C).

We found support of hypothesis 1 for both experiments, that increasing reliability increased final trust and dependence (see Figures 6, 7, 16, and 17). In Experiment 2, trust1 (Figure 14) and trust2 (Figure 15) also had a significant main effect of reliability. In alignment with existing literature (Pharmer et al., 2021; Holthausen & Walker, 2022; Patton & Wickens, 2024), this finding suggests that users of automation are sensitive to automation reliability, and they trust and depend on high reliability automation more than lower reliability automation. Therefore, reliability is a key feature of automation that influences how users interact with it. To

optimize the relationship between users and automation, reliability should be as high as possible. When automation reliability is lower, users should be aware of this to shorten the period needed to calibrate to the reliability of automation because users otherwise tend to assume that automation is highly reliable.

We had mixed findings for hypothesis 2 across experiments, finding no main effect of transparency on final trust or dependence in Experiment 1, and that transparency increased final trust, but decreased dependence in Experiment 2 (see Figures 6, 7, 16, and 17). We found that in both experiments, transparency increased trust1 and trust2 (see Figures 4, 5, 14, and 15). When the transparency cues failed to provide insight to the user if the automation has made a prediction error (as was the case in Experiment 1), transparency had no effect on trust or dependence for users. However, when transparency cues provide insight into when automation will make a prediction error, transparency overall increased final trust and decreased dependence because this transparency provides a reason for users to disagree with the prediction of automation. Additionally, transparency seems to reliably increase trust on a trial by trial basis, but not reliably in hindsight when participants were asked to report final trust. Therefore, transparency does ont always increase dependence, but it may decrease dependence depending on the type of transparency used. Transparency is still a beneficial attribute of automation, specifically when the transparency indicates that automation may err. Transparency has the potential to improve the relationship between automation and the user, but only when transparency provides insight to automation that is otherwise unknown. In accordance with the meta-analysis by Patton & Wickens (2024) which found that trust and dependence are unequal, particularly in terms of shifts in reliability. The effect sizes of the varying trust measures for both reliability and transparency are greater than the effect sizes of dependence (see Appendix C). This finding

aligns with the Patton & Wickens (2024) meta-analysis and adds that the relationship between trust and dependence seem to not only differ due to shifts in reliability, but also transparency.

We also had mixed findings for hypothesis 3. There was no support in Experiment 1, but in Experiment 2, transparency diminished the increasing effect of reliability on final trust (see Figures 6 and 16), a finding consistent with that of Gegoff and colleagues (2024). Regarding dependence, in Experiment 1, the interaction between reliability and transparency was approaching significance in that we found that transparency trended to mitigate the increasing effect of reliability on dependence (see Figures 7 and 17). A significant interaction would suggest that transparency is less beneficial for users at high reliability and that the importance of transparency increases as automation reliability decreases. Wickens et al. (2015) suggested and validated a model for acting when automation errs. To override automation, an error must first be noticed, then interpreted, and subsequent action must be taken. Gegoff et al. (2024) further found and validated that more attention may be provided to transparency when automation is lower in reliability, emphasizing both the importance of user reliability calibration and clarity of transparency in indicating automation errors. Therefore, transparency is most beneficial for the relationship between the user and automation when reliability is lower and again when transparency indicates that automation will make an error. Transparency only complicates decision tasks at high reliability and does not need to be included.

In support of hypothesis 4 in both experiments, we found that users of automation performed worse than highly reliable automation alone with and without transparency (see Figures 8 and 18), in alignment with existing literature (Boskemper et al., 2022, Bartlett & McCarley, 2017, 2019, 2021, 2022). At low reliability, users performed worse than automation alone with transparency, but not without for Experiment 1. Users performed as well as low

reliability automation in Experiment 2 with and without transparency. Even with the presence of transparency at high reliability, users of automation performed worse than automation alone, so they could not calibrate to the reliability of the automation quickly enough to experience the full benefits of using highly reliable automation. Furthermore, in Experiment 1 where transparency did not inform users that automation was going to make an error, transparency hurt performance whereas a lack of transparency did not. In contrast, transparency indicative of automation errors (as used in Experiment 2) did not hurt performance. Transparency that does not functionally provide additional information about the errors of automation and the accuracy of its predictions should be avoided. When transparency indicates errors of automation, users calibrate to its low reliability better. For highly reliable automation, regardless of transparency or the type of transparency, performance remains worse than that of automation alone.

When comparing the results of Experiment 1 to the results of Experiment 2, we expected to find higher trust with transparency in Experiment 2, lower dependence with transparency in Experiment 2, and better accuracy with transparency in Experiment 2. We did not find that trust1 or trust2 differed between Experiment 1 and Experiment 2 (trust1 $p = .30$, trust2 $p = .51$). We found a difference between Experiment 1 and Experiment 2 for final trust ($p = .04$), but not that final trust was lower without transparency in Experiment 2, rather than with transparency as hypothesized. We found no difference between experiments regarding dependence ($p = .33$) or accuracy ($p = .07$). These findings indicate that the type of transparency in Experiment 2 does not necessarily decrease dependence or improve performance. The difference in accuracy between experiments does approach significance, indicating that participants seemed to be more accurate with transparency that indicated automation errors. However, for final trust, experiment

2 decreased trust, trending that no transparency for this experiment type decreased trust, though this interaction was not significant (p = .12).

We also found that as distance from the 6-inch decision threshold increased, all outcome measures (trust, dependence, and performance) (see Figures 10, 11, 12, 19, 20, and 21) increased in both experiments. There are an inherent distrust and disuse in automation when the prediction lands on the decision threshold. Automation may not account for the nuance associated with the lack of room for error when a prediction lands close to criteria required to make a decision. Possibly, in these situations, the human automation team would benefit from additional transparency or precision in providing its predictions to bridge the gap between the user's uncertainty resulting from the lack of room for error and the automation's rigid predictions.

From the questions that were asked at the end of the experiment, in both experiments we found that participants overwhelmingly preferred the presence of transparency (see Tables 4 and 7) and felt that transparency predictions were more accurate (see Tables 5 and 8), even though this was not the case. Participants estimated the error rate of automation close to what it was (suggesting good sensitivity to the reliability of automation) in both experiments for 70% reliability, but they overestimated the error rate for 90% reliability in both experiments (see Figures 12 and 22) and they had a high baseline trust in weather forecasting automation (see Figures 13 and 23). Most users of automation only needed 20 (10 trials for each block) total trials to calibrate to the reliability of automation. Regardless of whether we found that transparency increased trust, dependence, and performance, most users of automation liked automation to have transparency, such that they perceived automation to be more accurate when transparency was present. Additionally, users of weather forecasts typically have a high baseline trust in these weather forecasts and will subsequently likely trust and depend on these forecasts higher than in

their use of other automation. At high reliability automation participants experienced what is referred to as sluggish beta, that they were slow to calibrate to the reliability of automation when it was highly reliable (Wickens et al., 2021). Therefore, participants find a benefit in transparency being present, possibly to an extent that inflates the reliability of automation.

Limitations and Future Directions

One limitation of the current experiments was that the transparency cues were not often interpreted in the way they were designed. Multiple participants reported being confused by or failing to understand the transparency. Future research should examine different types of transparency and determine the best way to present transparency to participants. Another limitation was that stimuli added an additional layer of complexity to interpreting results, due to the finding that distance from the decision threshold had a significant effect on each outcome variable. The current research also only examined the effects of the intensity of errors that were consistent in magnitude (all errors were 2 inches). Future research could examine the effects of transparency and reliability when magnitude of error is manipulated, to build off the current research investigating type of automation error.

In accordance with existing literature, we found that high reliability automation was trusted more, depended on more, and resulted in higher accuracy in users. Reliability in automation should be maximized and disclosed to participants to shorten the calibration window. Additionally, transparency seems to aid in reliability calibration, but more so at low (70%) reliability and when it is not entirely redundant with automation's predictions but indicates its potential errors. Automation transparency is overall beneficial for the relationship between the user and automation, but its use is most effective and not detrimental when automation reliability is lower.

APPENDIX A

Package Name	**Version**	**Citation**
R Studio	2024 4.4.2	RStudio Team. (2020). RStudio: Integrated Development Environment for R [Computer software]. Boston, MA: RStudio, PBC. http://www.rstudio.com/
tidyverse	2.0.0	Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." _Journal of Open Source Software_, *4*(43), 1686. doi:10.21105/joss.01686 <https://doi.org/10.21105/joss.01686>.
lme4	1.1.36	Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.
ggeffects	2.2.1	Lüdecke D (2018). "ggeffects: Tidy Data Frames of Marginal Effects from Regression Models." _Journal of Open Source Software_, *3*(26), 772. doi:10.21105/joss.00772 <https://doi.org/10.21105/joss.00772>.
emmeans	1.11.0	Lenth R (2025). _emmeans: Estimated Marginal Means, aka Least-Squares Means_. R package version 1.11.0, <https://CRAN.R-project.org/package=emmeans>.

lmerTest	3.1.3	Kuznetsova A, Brockhoff PB, Christensen RHB (2017). “lmerTest Package: Tests in Linear Mixed Effects Models.” _Journal of Statistical Software_, *82*(13), 1-26. doi:10.18637/jss.v082.i13 <https://doi.org/10.18637/jss.v082.i13>.
MuMIn	1.48.11	Bartoń K (2025). _MuMIn: Multi-Model Inference_. R package version 1.48.11, <https://CRAN.R-project.org/package=MuMIn>.

APPENDIX B

Experiment 1: main effect of reliability

Measure	Trust1	Trust2	Final Trust	Dependence	Accuracy
t or z statistic	0.20	1.82	2.52	5.57	9.07
p-value	0.84	.07	.01	< .001	< .001

Experiment 1: main effect of transparency

Measure	Trust1	Trust2	Final Trust	Dependence	Accuracy
t or z statistic	9.26	7.10	-0.83	0.22	-3.71
p-value	< .001	< .001	.41	0.82	< .001

Experiment 1: interaction between transparency and reliability

Measure	Trust1	Trust2	Final Trust	Dependence	Accuracy
t or z statistic	2.93	1.15	1.47	-1.88	-0.42
p-value	.003	.25	.14	.06	.68

Experiment 2: main effect of reliability

Measure	Trust1	Trust2	Final Trust	Dependence	Accuracy
t or z statistic	2.08	2.93	5.57	5.27	8.91
p-value	.04	.003	< .001	< .001	<. 001

Experiment 2: main effect of transparency

Measure	Trust1	Trust2	Final Trust	Dependence	Accuracy
t or z statistic	13.11	10.02	3.74	-4.82	1.05
p-value	< .001	< .001	< .001	< .001	.29

Experiment 2: interaction between transparency and reliability

Measure	Trust1	Trust2	Final Trust	Dependence	Accuracy
t or z statistic	-.09	0.27	-2.43	1.11	-1.42
p-value	.93	.79	.02	.27	.16

APPENDIX C

Experiment 1 Effect Sizes (using R squared conditional or R2C)

Effect Sizes	Effect on Reliability	Effect on Transparency
Trust1	**0.28**	**0.32**
Trust2	**0.29**	**0.31**
Final Trust	**0.41**	**0.41**
Dependence	**0.14**	**0.13**

Experiment 2 Effect Sizes (using R squared conditional or R2C)

Effect Sizes	Effect on Reliability	Effect on Transparency
Trust1	**0.27**	**0.34**
Trust2	**0.31**	**0.35**
Final Trust	**0.40**	**0.42**
Dependence	**0.09**	**0.10**

REFERENCES

Bartlett, M. L., & McCarley, J. S. (2017). Benchmarking aided decision making in a signal detection task. *Human Factors*, *59*, 881–900. https://doi.org/ 10.1177/ 0018720817700258

Bartlett, M. L., & McCarley, J. S. (2019). No effect of cue format on automation dependence in an aided signal detection task. *Human Factors*, *61*, 169–190. https://doi.org/10.1177/ 0018720818802961

Bartlett, M. L., & McCarley, J. S. (2021). Ironic efficiency in automation-aided signal detection. *Ergonomics*, *64*(1), 103-112. https://doi.org/10.1080/00140139.2020.1809716

Boskemper, M. M., Bartlett, M. L., & McCarley, J. S. (2022). Measuring the efficiency of automation-aided performance in a simulated baggage screening task. *Human factors*, *64*(6), 945-961. https://doi.org/10.1177/0018720820983632

Burgeno, J. N., & Joslyn, S. L. (2020). The impact of weather forecast inconsistency on user trust. *Weather, climate, and society*, *12*(4), 679-694. https://doi.org/10.1175/WCAS-D-19-0074.1

Chung, H., & Yang, X. J. (2025). Understanding multi-referent trust in AI-supported evacuations: The role of transparency and altruism. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 69*(1), 214-217. https://doi.org/10.1177/10711813251358779

Davenport, R. B., & Bustamante, E. A. (2010). Effects of false-alarm vs. miss-prone automation and likelihood alarm technology on trust, reliance, and compliance in a miss-prone task. *Proceedings of the human factors and ergonomics society annual meeting 54*(19), 1513-1517. https://doi.org/10.1518/107118110X12829370088642

Elder, K., Xirasagar, S., Miller, N., Bowen, S. A., Glover, S., & Piper, C. (2007). African Americans' decisions not to evacuate New Orleans before Hurricane Katrina: A qualitative study. *American Journal of Public Health*, *97*(1), https://doi.org/10.2105/AJPH.2006.100867

Endsley, M. R. (2017). From here to autonomy: lessons learned from human–automation research. *Human Factors*, *59*(1), 5-27. https://doi.org/10.1177/0018720816681350

Eriksson, A., & Stanton, N. (2017). Takeover time in highly automated vehicles: Nontraditional transitions to and from manual control. *Human Factors*, *59*, 689–705. https://doi.org/10.1177/0018720816685832

Gegoff, I., Tatasciore, M., Bowden, V., McCarley, J., & Loft, S. (2024). Transparent automated advice to mitigate the impact of variation in automation reliability. *Human Factors*, *66*(8), 2008-2024. https://doi.org/10.1177/00187208231196738

Gegoff, I., Tatasciore, M., Bowden, V. K., & Loft, S. (2025). Deciphering automation transparency: do the benefits of transparency differ based on whether decision recommendations are provided?. *Human Factors*, https://doi.org/10.00187208251318465.

Guznov, S., Lyons, J., Pfahler, M., Heironimus, A., Woolley, M., Friedman, J., & Neimeier, A. (2020). Robot transparency and team orientation effects on human–robot teaming. *International Journal of Human–Computer Interaction*, *36*(7), 650-660. https://doi.org/10.1080/10447318.2019.1676519

Hancock, P. A., Billings, D. R., Oleson, K. E., Chen, J. Y., De Visser, E., & Parasuraman, R. (2011). A meta-analysis of factors influencing the development of human-robot trust. *Human Factors*, *53*, 517–727.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407-434. https://doi.org/10.1177/0018720814547570

Holthausen, B. E., Stuck, R. E., & Walker, B. N. (2022). Trust in automated vehicles. In *User experience design in the era of automated driving*. https://doi.org/10.1007/978-3-030-77726-5_2

Hussein, A., Elsawah, S., & Abbass, H. A. (2020). The reliability and transparency bases of trust in human-swarm interaction: Principles and implications. *Ergonomics*, *63*(9), 1116-1132. https://doi.org/10.1080/00140139.2020.1764112

Joslyn, S. L., & LeClerc, J.E. (2012). Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of experimental psychology: Applied*. *18*(1), https://doi.org/10.1037/a0025185

Kadous, K., Mercer, M., & Thayer, J. (2009). Is there safety in numbers? The effects of forecast accuracy and forecast boldness on financial analysts' credibility with investors. *Contemporary Accounting Research*, *26*(3), 933-968. https://doi.org/10.1506/car.26.3.12

Kahn, B. E., & Luce, M. F. (2003). Understanding high-stakes consumer decisions: Mammography adherence following false-alarm test results. *Marketing Science*, *22*(3), 393-410. https://doi.org/10.1287/mksc.22.3.393.17737

Kaltenbach1, E., & Dolgov, I. (2017). On the dual nature of transparency and reliability: Rethinking factors that shape trust in automation. In *Proceedings of the human factors and ergonomics society annual meeting 6*(1), 308-312. https://doi.org/10.1177/1541931213601558

Korbalich, C., Hättenschwiler, N., Ferris, T., & Lewandowsky, S. (2018). Automation in visual inspection tasks: X-ray luggage screening – Examining benefits and possible implementations of automated explosives detection. *Ergonomics*, *61*(5), 740–753. https://doi.org/10.1080/00140139.2018.1481231

Kunze, Alexander & Summerskill, Steve & Marshall, Russell & Filtness, Ashleigh. (2019). Function-Specific Uncertainty Communication in Automated Driving. *International Journal of Mobile Human Computer Interaction*. *11*. 75-97. https://doi.org/10.4018/IJMHCI.2019040105.

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors : The Journal of the Human Factors and Ergonomics Society.*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Losee, J. E., & Joslyn, S. (2018). The need to trust: How features of the forecasted weather influence forecast trust. *International Journal of Disaster Risk Reduction*, *30*, 95-104. https://doi.org/10.1016/j.ijdrr.2018.02.032

Luo, R., Du, N., & Yang, X. J. (2022). Evaluating effects of enhanced autonomy transparency on trust, dependence, and human-autonomy team performance over time. *International Journal of Human–Computer Interaction*, *38*(18). https://doi.org/10.1080/10447318.2022.2097602

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human factors*, *50*(2), 194-210. https://doi.org/10.1518/001872008X288574

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of experimental psychology: Applied*, *6*(1), 44-58. https://doi.org/10.1037/0278-7393.6.1.44

Munoz Gomez Andrade, F., Bartlett, M. L., Wickens, C. D., & McCarley, J. S. (2025). Leaving money on the table: As diagnostic aids become more useful, operators use them less efficiently. *Journal of experimental psychology. Applied*. https://doi.org/10.1037/xap0000549

Parasuraman, R., & Manzey, D. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, *52*, 381–410. https://doi.org/10.1177/0018720810376055

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

Patton, C. E. (2023). Automation use: The roles of self-confidence, trust, and workload. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 67*(1) 965-970 https://doi.org/10.1177/21695067231192708

Patton, C. E., & Wickens, C. D. (2024). The relationship of trust and dependence. *Ergonomics*, *67*(11), 1535-1552. https://doi.org/10.1080/00140139.2024.2342436

Patton, C. E., Wickens, C. D., Smith, C. A. P., Noble, K. M., & Clegg, B. A. (2023). Supporting detection of hostile intentions: automated assistance in a dynamic decision-making context. *Cognitive Research: Principles and Implications*, *8*(1), 69. https://doi.org/10.1186/s41235-023-00519-5

Pharmer, R. L., Wickens, C. D., & Clegg, B. A. (2025). Transparent systems, opaque results: a study on automation compliance and task performance. *Cognitive Research: Principles and Implications*, *10*(1), 8. https://doi.org/10.1186/s41235-025-00629-4

Pharmer, R., Wickens, C., Clegg, B., & Smith, C (2021). Effect of procedural elements on trust and compliance with an imperfect decision aid. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting.* https://doi.org/10.1177/1071181321651191

Rittenberg, B. S., Holland, C. W., Barnhart, G. E., Gaudreau, S. M., & Neyedli, H. F. (2024). Trust with increasing and decreasing reliability. *Human Factors*, *66*(12), 2569-2589. https://doi.org/10.1177/00187208241228636

Roth, G., Schulte, A., Schmitt, F., & Brand, Y. (2020). Transparency for a workload-adaptive cognitive agent in a manned-unmanned teaming application. *IEEE Transactions on Human-Machine Systems*, *50*(3), 225–233. https://doi.org/10.1109/THMS.2019.2914667

Sanchez, J., Rogers, W. A., Fisk, A. D., & Rovira, E. (2014). Understanding reliance on automation: effects of error type, error distribution, age and experience. *Theoretical issues in ergonomics science*, *15*(2), 134-160. https://doi.org/10.1080/1463922X.2011.611269.

Sargent, R., Walters., B., & Wickens, C. (2023). Meta-analysis qualifying and quantifying the benefits of automation transparency to enhance models of human performance. *International Conference on Human-Computer Interaction.* 243-261. https://doi.org/10.1007/978-3-031-35596-7_16

Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., & Vössing, M. (2022). A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society,* 617-626, https://doi.org/10.1145/3514094.3534128 ISBN: 9781450392471

Van de Merwe, K., Mallam, S., & Nazir, S. (2022) Agent transparency, situation aware-ness, mental workload and operator performance: A systematic literature review. *Human Factors*, 1-29. https://doi.org/10.1177/00187208221077804

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201-212. https://doi.org/10.1080/14639220500370105

Wickens, C. D., Sebok, A., Li, H., Sarter, N., & Gacy, A. M. (2015). Using modeling and simulation to predict operator performance and automation-induced complacency with robotic automation: A case study and empirical validation. *Human Factors*, *57*(6), 959–975. https://doi.org/10.1177/0018720814566454

Wickens, C. D., Fitzgerald, N. J., Clegg, B. A., Smith, C. A. P., Orth, D., & Kincaid, K. (2020). Decision aiding for nautical collision avoidance: Trust, dependence, and implicit understanding of the

decision algorithm. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 64*(1), 1950-1954. https://doi.org/10.1177/1071181320641470

Wickens, C. D., Helton, W. S., Hollands, J. G., & Banbury, S. (2021). *Engineering psychology and human performance*. https://doi.org/10.4324_9781003177616

Wickens, C., Sargent, R., Walters, B. (2023). An influence model of the human-automation team: Effects of workload and automation reliability, transparency, and degree. *Ergonomics International Journal.* https://doi.org/10.23880/eoij-1600031

Wischnewski, M., Krämer, N., & Müller, E. (April). Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* 1-16, https://doi.org/ 10.1145/3544548.3581197

Wright, J. L., Chen, J. Y. C., & Lakhmani, S. G. (2020). Agent transparency and reliability in human–robot interaction: The influence on user confidence and perceived reliability. *IEEE Transactions on Human-Machine Systems*, *50*(3), 254–263. https://doi.org/10.1109/THMS.2019.2925717

Xu, X. , Wickens, C. D. , & Rantanen, E. M. (2007). Effects of conflict alerting system reliability and task difficulty on pilots' conflict detection with cockpit display of traffic information. *Ergonomics*, 50, 112–130. https://doi.org/10.1080/00140130601002658

Yeh, M., Merlo, J. L., Wickens, C. D., & Brandenburg, D. L. (2003). Head up versus head down: The costs of imprecision, unreliability, and visual clutter on cue effectiveness for display signaling. *Human Factors*, *45*, 390–407. https://doi.org/10.1518/hfes.45.3.390.27249

Zhang, Q., Yang, X. J., & Robert Jr, L. P. (2025). Understanding explanation content for cognitive and affective trust in automated vehicles. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 69*(1), 1028-1033, https://doi.org/10.1177/10711813251366284