Human Factors in Design, Engineering, and Computing, Vol. 159, 2024, 294–304
https://doi.org/10.54941/ahfe1005594

Trust Us: A Simple Model for Understanding Trust in AI

Kyra Wisniewski, Christina Ting, and Laura Matzen

Sandia National Laboratories, Albuquerque, NM 87123, USA

ABSTRACT

We address the dissonance in the formal study of trust in artificial intelligence (AI) by presenting a simple trust model that connects the two most-commonly cited definitions of trust. This dissonance can be largely attributed to the fact that expressions of trust are familiar to us, but the abstract concepts we formally study are not. To illustrate, consider what it means to trust your car navigation system. You might say that you trust your navigation system's ability to recommend the best route during rush hour. However, when it comes down to it, you may opt to stay on your standard route. Your words express trust as an attitude, your actions express trust as an intention. While we can easily differentiate the expressions of trust in everyday life, overloading of the term "trust" to mean both an attitude and an intention has led to a lack of precision and confusion in its formal study. We analyze the two papers most frequently cited by the community for their definitions of trust. One paper defines trust as an attitude (Lee and See, 2004), while the other defines trust as an intention (Mayer, Davis, & Schoorman, 1995). We develop a simple trust model that clearly articulates the relationship between these definitions. Simply put, trust as an attitude is weighed against perceived risk to determine trust as an intention. We also use the model to define appropriate trust in AI. A major goal of this work is to enable the design of trust experiments that manipulate and measure components of a shared model, allowing for comparison across research efforts and the building up of a consistent body of trust research. A practical implication of understanding trust is to strengthen the relationship between humans and technology.

Keywords: Trust in AI, Trust in automation, Trust measures, Trust, Appropriate trust, Human-machine teaming, Decision making

INTRODUCTION

Trust is expressed in two ways. We *say* we trust and we *show* we trust. Sometimes our words and our actions can conflict. To illustrate, consider what it means to trust your car, your navigation system, or your home security system. You might say that you trust your car's ability to drive across the country, your navigation system's ability to recommend the best route during rush hour, or your home video security system's ability to detect intruders while you are out of town for a month. However, when it comes down to it, your actions may suggest otherwise. Even though you said you trust your car, you may rent a car instead. Even though you said you trust your navigation system, you may opt to stay on your standard route. Even though you said

you trust your home security system, you may ask your friend to stop by and check that everything is secure.

In the literature, trust is defined as both an attitude and an intention. We express our attitudes and intentions through our words and actions, respectively. While we can easily differentiate these expressions of trust in everyday life, overloading of the term "trust" to mean both an attitude and an intention has led to dissonance in its formal study. To resolve this dissonance we:

1. Connect the definitions of trust as an attitude and trust as an intention.
2. Present a trust model that clearly incorporates the two definitions of trust as separate concepts.
3. Define appropriate trust in artificial intelligence (AI) within the context of the model.

Our goal is to provide researchers studying trust in AI with common ground.

TWO DEFINITIONS OF TRUST

There are two definitions of trust frequently cited by the trust in AI literature: trust as an attitude (Lee & See, 2004) and trust as an intention (Mayer, Davis, & Schoorman, 1995). To formally study trust, it is critical to understand the distinction and relationship between these two definitions of trust.

Lee and See (2004) define trust as:

> *The attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability.*

This definition describes trust as an attitude. Importantly, the uncertain situation may result in gains or losses as a result of trusting the agent in question;[1] trust as an attitude can be thought of as an evaluation of this potential. Mapping the navigation system example to this definition, trust in your navigation system is the attitude that the route your navigation system suggests is likely to help you reach your destination on time.

On the other hand, Mayer et al. (1995) define trust as:

> *The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party.*

Consistent with Lee and See (2004)'s understanding that "an attitude is an affective evaluation of beliefs that guides people to adopt a particular intention", Mayer et al. (1995)'s definition describes trust as an intention, or a willingness to be vulnerable to the trustee, based on an attitude, or expectation, about the trustee. Therefore, despite defining the word "trust" to mean different concepts, the two authors agree that trust as an attitude

[1]Trust is not a relevant attitude when the ability of the agent to help you achieve your goals is certain; that is, all gains and no losses.

guides trust as an intention. Mapping the navigation system example to this relationship, trust as an intention is the willingness to follow the navigation system's suggested route based on the attitude that the navigation system is likely to help you reach your destination on time.

To avoid using the word "trust" to refer to both an attitude and an intention, we note that both authors indirectly refer to trust as an attitude as "perceived trustworthiness." Mayer et al. (1995) refer to the perceived characteristics of the trustee that guide the development of trust as an intention as perceived trustworthiness. Lee and See (2004) connect trust as an attitude with perceived trustworthiness by referring to an automation's capabilities as its trustworthiness and defining trust as an attitude in terms of perceptions about these capabilities. Except to emphasize the distinction between attitudes and intentions, we will also refer to trust as an attitude as "perceived trustworthiness", leaving the word "trust" to refer to trust as an intention. Said succinctly, perceived trustworthiness guides trust.

The two authors additionally agree that other attitudes beyond perceived trustworthiness guide trust. For Lee and See (2004):

> *Trust [perceived trustworthiness] combines with other attitudes (e.g., subjective workload, effort to engage, perceived risk, and self-confidence) to form the intention to rely [trust] on the automation.*

and for Mayer et al. (1995):

> *We propose that the level of trust [perceived trustworthiness] is compared to the level of perceived risk in a situation. If the level of trust surpasses the threshold of perceived risk, then the trustor will engage in the risk taking in relationship [an expression of trust].*

Mayer et al. (1995) define perceived risk to involve "the trustor's belief about likelihoods of gains or losses outside of considerations that involve the relationship with the particular trustee". Similarly, the "other attitudes" listed by Lee and See (2004) are about gains or losses that do not involve the trustee. Therefore, trust is an evaluation of attitudes about the potential for gains or losses involving the trustee against those not involving the trustee. In what follows, we will refer to all attitudes not involving the trustee as "perceived risk". Mapping the navigation system example to this relationship, trust as an intention is the willingness to follow the navigation system's suggested route based on the attitude that you are more likely to reach your destination on time if you use the navigation system than if you do not. Said succinctly, trust is an evaluation of perceived trustworthiness against perceived risk.

In summary, Lee and See (2004) argue that perceived trustworthiness combines with perceived risk to form trust; Mayer et al. (1995) argue that trust is the willingness to be vulnerable based on perceptions of trustworthiness and risk. The agreement between these two perspectives is obscured when the word "trust" is used to refer to both attitudes and intentions.

A SIMPLE TRUST MODEL

Our analysis of the two definitions of trust leads us to a simple model; see Figure 1. In this section, we explore a theoretical framework for understanding how different factors influence perceived trustworthiness and risk and how these attitudes are evaluated to form an intention to trust. Our model is described in three layers: the actual trustworthiness and risk, the individual's perceptions of trustworthiness and risk, and the individual's decision to form an intention to trust.

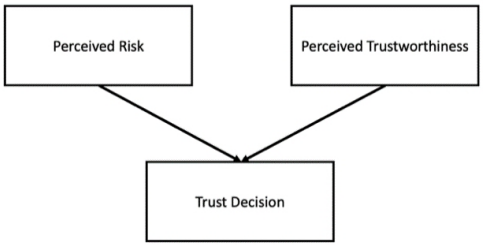

Figure 1: A simple trust model.

Actual Risk and Trustworthiness

Perceived trustworthiness and risk were previously introduced as an individual's attitudes about the potential for gains or losses in a situation with and without the AI's help in achieving a goal. Actual trustworthiness and risk can therefore be defined in terms of the true potential for gains or losses when the AI is and is not used. A natural way to represent the potential for gains or losses is as a probability distribution over the set of outcomes, where outcomes with positive consequences are gains and outcomes with negative consequences are losses. The set of outcomes and their consequences are specified by the goal and the probability of those outcomes is determined by the situation. We visualize examples of these distributions as violin plots, where the vertical position denotes the consequence of the outcome and the width at that position represents the probability.[2]

Figure 2A shows the actual risk and trustworthiness of two different situations. In both situations, you are driving to work on your normal interstate route during standard rush hour traffic, and your goal is to get to work on time. The consequences of possible outcomes, as dictated by the goal of getting to work on time,[3] range from arriving early with more time to prepare for the day (gains) to arriving late with less time to prepare for

[2]Although the distributions are represented continuously for theoretical discussion, in application this level of granularity is rarely required. For example, some goals have discrete outcomes like correct or incorrect.

[3]Examples of changes in goals that would affect the consequences of possible outcomes include getting to work for a very important meeting (consequences of being early are very positive whereas being late are very negative) or getting to work the day before a long weekend (consequences of being early, on time, or late are about the same and neutral).

the day (losses). The probability of these gains or losses is determined by the situation.

In **Situation 1**, you have allowed sufficient time for your usual commute. However, your outdated navigation system suggests an alternative route using local roads. Its recommendation is based on the current traffic conditions, so by the time you reach that route it may be backed up and no longer optimal. The actual risk of staying on the interstate is that you will most likely arrive slightly early. The actual trustworthiness of complying with your navigation system is that you may arrive early or you may arrive late, depending on how many other people also take the alternative route. It is slightly more likely that you will arrive early, as shown by the width of the violin plot.

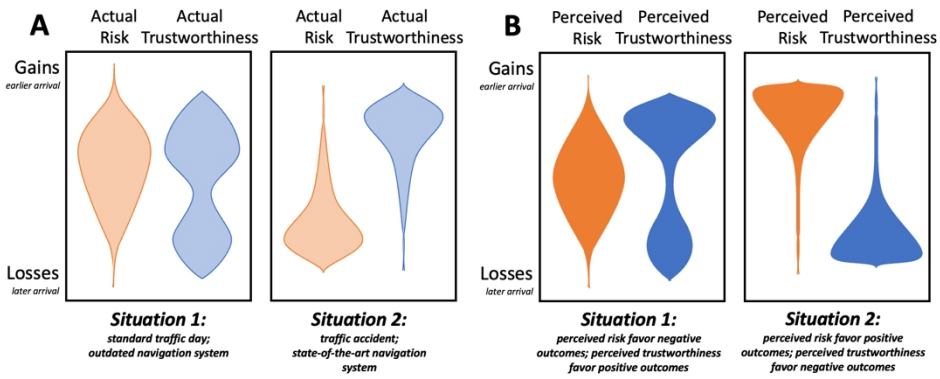

Figure 2: Visualizations of the probability for gains or losses as violin plots. A) The actual risk and trustworthiness and B) the perceived risk and trustworthiness for the same two situations. See the text for a description of the factors influencing the situations and perceptions.

In **Situation 2**, a traffic accident has just occurred, and all lanes are blocked. This time you have a state-of-the-art navigation system that can forecast traffic conditions in the future. Your navigation system suggests that you take the next exit. The actual risk of staying on the interstate is that you will be stuck in traffic and very likely to arrive late. The actual trustworthiness of the navigation system is high, so complying with it will successfully route you around the accident and get you to work early.

Perceived Risk and Trustworthiness

As with actual risk and trustworthiness, perceived risk and trustworthiness can also be represented by a probability distribution over outcomes;[4] here the outcomes are perceived. Factors that affect an individual's perceptions include individual differences, the situation, and past experiences. Lee and See (2004) include many of these factors in their trust model: reputation, gossip, interface features, organizational structure, cultural differences, predisposition to trust, workload, exploratory behavior, effort to engage,

[4]This representation is useful for theoretical explanation, but not necessarily a realistic representation of perceptions. We do not suggest researchers try to fully characterize these distributions.

self-confidence, time constraints, and configuration errors. Some of these factors may affect both perceptions of risk and trustworthiness, while others may affect only one. For example, Lee and Moray (1992) explicitly list process, performance, and purpose of the AI as factors that affect perceived trustworthiness.

Figure 2B shows possible perceptions for **Situation 1** and **Situation 2**. The perceived risk may favor more positive outcomes (gains) if you are confident in your own assessment of the traffic conditions or recently made it through traffic quickly. Alternatively, it may favor more negative outcomes (losses) if you heard from a friend that it is a particularly bad traffic day or have a low risk tolerance. The perceived trustworthiness may favor more negative outcomes (losses) if you have a low risk tolerance or a prior bad experience with your navigation system. Alternatively, it may favor more positive outcomes (gains) if the navigation system explains how it calculates the best possible route or you have a high trustor's propensity. As mentioned above, some factors such as risk tolerance can affect both perceived risk and trustworthiness. Note that the perceived risk and trustworthiness do not match the actual risk and trustworthiness in either situation. In general, the actual risk and trustworthiness are unknown to the individual.

Trust Decision

A trust decision is made from an individual's evaluation of perceived risk and trustworthiness. In contrast to perceptions, the trust decision is binary: you are either willing to be vulnerable to the AI in a given situation or you are not. The decision making strategy an individual uses is complex and, as with perceptions, depends on individual differences, the situation, and past experiences. The large body of literature studying the complexities of human decision making should be referenced to understand this in depth (e.g., Tversky & Kahneman, 1974; Tversky, Kahneman, & Slovic, 1982).

Sometimes the decision making strategy is apparent. For example, consider **Situation 2** in Figure 2B. Because you perceive a high probability that you will arrive early on your standard route to work (gains in perceived risk) and late if you comply with your navigation system (losses in perceived trustworthiness), you decide not to trust your navigation system. Other times the decision making strategy may be more complex. Consider now **Situation 1** in Figure 2B. Whether or not you decide to trust the navigation system depends on if you prioritize the probability that you could arrive early (gains in perceived trustworthiness) or the probability that you could arrive late (losses in perceived trustworthiness) with the help of the navigation system.

APPLYING THE SIMPLE TRUST MODEL

In this section, we explain how researchers can apply the simple trust model to evaluate the appropriateness of a trust decision and to assess the perceptions that led experiment participants or real-world system users to a trust decision.

Evaluating the Appropriateness of the Trust Decision

A comparison between actual risk and trustworthiness determines whether or not the AI should be used. Researchers who are studying trust in AI can design experiments where the actual risk and trustworthiness are controlled. When studying real-world systems, researchers can estimate the actual risk and trustworthiness through observation or through modelling and simulation. To measure actual risk and trustworthiness, the true probability distributions of potential outcomes and their consequences are determined. For theoretical discussion these distributions were presented as continuous, but in the real-world consequences are often discrete. Specifically, they can be ordinal and therefore ranked; for example as good, better, best or simply as good vs. bad. Additionally, while a probability distribution is preferred because it allows for a richer comparison, summary statistics (e.g., accuracy) can be used to estimate the actual risk and trustworthiness.

Trust is expressed through actions, so the trust decision can be measured through use, such as compliance with or reliance on an AI. It is important to emphasize that trust is an intention which is distinct from use. Frequently, there are external factors such as workload and configuration errors (Lee & See, 2004) that can prevent an individual who intends to trust from using a system. For use to be a reliable measure of trust, these external factors must be controlled for in experiments and accounted for in real-world systems.

We can now define appropriate trust in AI. The actual risk and trustworthiness determine the trust decision that *should* occur. The individual's use determines the trust decision that *does* occur. Using these terms, appropriateness is when the trust decision that *should* occur, *does* occur. The four fundamental types of trust decisions are shown in Figure 3. They are appropriate trust, appropriate distrust, inappropriate trust, and inappropriate distrust.

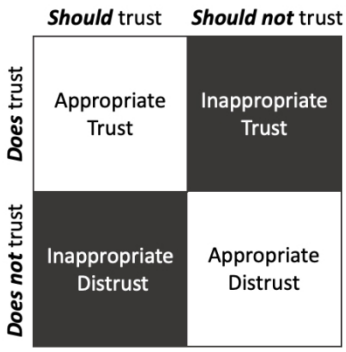

Figure 3: The four fundamental types of trust decisions.

Alternative definitions of appropriateness based on actual and perceived trustworthiness have been proposed (de Visser et al., 2019; Lee & See, 2004). However, these definitions do not account for perceived risk or for the decision making process that leads to the trust decision. Figure 4 shows

why this is problematic. Using **Situation 1** as the example, a comparison of the most likely outcomes of the actual risk and trustworthiness indicates that the navigation system should not be trusted. However, an inappropriate decision to trust the navigation system can occur even with accurate perceptions of the navigation system's trustworthiness because of (A) an inaccurate perception of risk due to lower self-confidence in your navigation abilities or (B) a different decision making strategy due to risk aversion.

Understanding How the Trust Decision was Made

We have shown that evaluating the appropriateness of a trust decision is straightforward using our model. A more challenging application of the model is understanding *how* a trust decision was made. Trust is an evaluation of perceived risk and perceived trustworthiness. Therefore, we need to measure the individual's perceptions of risk and trustworthiness and gather information about how they are weighing these two factors in their trust decision.

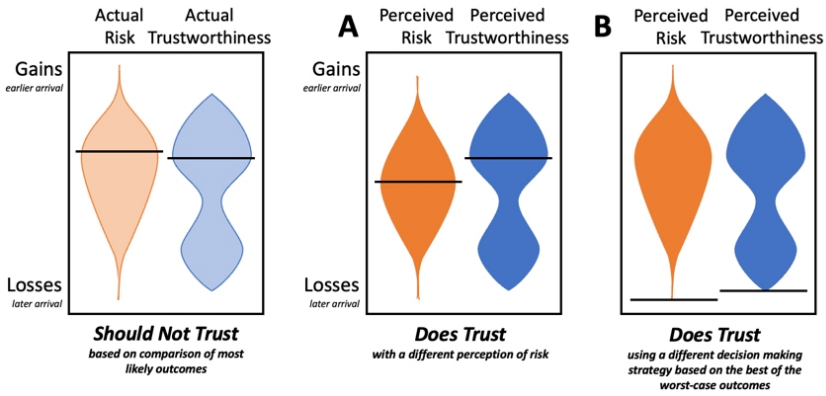

Figure 4: Even with accurate perceptions of trustworthiness, different perceptions of risk (A) or different decision making strategies (B) can tip the decision to inappropriate trust. The horizontal black lines indicate the portion of the distribution used to evaluate risk against trustworthiness.

While we have described human perceptions of risk and trustworthiness in terms of distributions of potential outcomes and their consequences, we do not suggest researchers try to fully characterize these distributions. In practice, researchers can instead make inferences about how participants perceive risk and trustworthiness through self-report measures. There are a large number of existing self-report measures in the trust literature; see Kohn and colleagues (2021) for a survey. However, overloading of the term "trust" in the literature means that a careful interpretation of these measures is necessary. In particular, self-report measures of trust asking people to express an attitude (e.g., "To what degree do you trust the AI?") capture perceived trustworthiness rather than trust (Lee & Moray, 1994). Additionally, many measures of perceived trustworthiness contain a mix of questions that ask about trust as an attitude (e.g., "The algorithm helps me achieve my goals")

and questions about specific aspects of a model (e.g., "I understand what the algorithm should do") that may contribute to trust as an attitude (Wojton, Porter, Lane, Bieber & Madhavan, 2020). Perceptions of specific aspects of a model's process, performance, or purpose (Lee & Moray, 1992) will influence perceptions of its overall trustworthiness, but questions about these aspects of the model may not directly capture trust as an attitude. When selecting or designing self-report measures, researchers should be attentive to how the terms "trust" and "trustworthiness" are defined.

From these self-report measures, researchers can try to infer how an individual's perceived risk and trustworthiness are used to make a trust decision. However, self-report data is imperfect and provides only limited insight into perceptions and their evaluations. Small changes in perceptions or decision making strategies may not be captured through self-report but can tip a decision from trust to distrust. As a result, self-reported measures of perceptions are not necessarily consistent with the trust decision; words and actions can (and frequently do) conflict. To understand how a decision to trust, and ultimately use, an AI is made, researchers need to make additional inferences about perceptions through careful experimental design and measures of individual differences in cognition. There are numerous measures of individual differences in cognition that may relate to perceptions of risk and trustworthiness, as well as decision making strategies; for example, trustor's propensity (Singh, Molloy, & Parasuraman, 1993), risk tolerance (Dohmen et al., 2011), need for cognition (Cacioppo & Petty, 1982), and perceived workload (Monfort, Graybeal, Harwood, McKnight, & Shaw, 2018). Incorporating measures of individual differences in cognition can help researchers to understand the patterns of decisions made by the participants in their experiments or the users of real-world AI systems.

CONCLUSION

Trust in AI is simpler than it seems. In the literature, trust is defined as both an attitude and an intention. Overloading the term "trust" to mean both an attitude and an intention has hidden the distinctions between the two concepts and made it challenging to formally study trust. We clarify these two definitions and show their relationship through a simple trust model. If the actual risk of the situation and trustworthiness of using the AI is known (whether pre-determined by an experimenter or characterized through observation or modelling and simulation), a person's use of the AI can be categorized as reflecting appropriate trust, inappropriate trust, appropriate distrust, or inappropriate distrust.

The challenge for researchers arises in determining how that person perceived the risk and the trustworthiness to arrive at their use of the AI. Understanding these perceptions is important for advancing the research on trust in AI and for supporting appropriate trust in real-world systems. However, these perceptions, and how people evaluate them, can be difficult to measure. Additionally, much of the existing literature focuses on perceived trustworthiness without paying much attention to perceived risk or to the decision making process. This gap, in addition to the conflation of terms like

"trust" and "trustworthiness" has led to a muddle of conflicting findings and claims.

Our goal in this paper is to highlight these discrepancies and gaps, in hopes of helping researchers to find common ground. In order to make progress as a field, we must work from a common understanding of what trust in AI means. Furthermore, we must be clear about what we are measuring when we are assessing self-report data and use. Clearly defining actual risk and trustworthiness, perceived risk and trustworthiness, and trust as an intention to use the AI will clarify conflicting findings in the existing research and support better experimental designs and replicability in future research. This clarity is necessary for developing coherent theories about human-AI interaction and for designing systems that support appropriate trust in real-world applications. Moving forward, we hope that the model presented in this paper will provide common ground to help researchers navigate the trust literature and design experiments to study trust in AI.

ACKNOWLEDGMENT

The authors would like to acknowledge Anna Polski at the University of Illinois Urbana-Champaign and Katherine Goode, Marieke Sorge, Mallory Stites, and Marie Tuft at Sandia National Laboratories. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

REFERENCES

Cacioppo, J. T. and Petty, R. E., 1982. The need for cognition. *Journal of personality and social psychology*, *42*(1), p. 116.

De Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R. and Neerincx, M. A., 2020. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics*, *12*(2), pp. 459–478.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J. and Wagner, G. G., 2011. Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the european economic association*, *9*(3), pp. 522–550.

Kohn, S. C., De Visser, E. J., Wiese, E., Lee, Y. C. and Shaw, T. H., 2021. Measurement of trust in automation: A narrative review and reference guide. *Frontiers in psychology*, *12*, p. 604977.

Lee, J. and Moray, N., 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), pp. 1243–1270.

Lee, J. D. and Moray, N., 1994. Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies*, *40*(1), pp. 153–184.

Lee, J. D. and See, K. A., 2004. Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), pp. 50–80.

Mayer, R. C., Davis, J. H. and Schoorman, F. D., 1995. An integrative model of organizational trust. *Academy of management review*, *20*(3), pp. 709–734.

Monfort, S. S., Graybeal, J. J., Harwood, A. E., McKnight, P. E. and Shaw, T. H., 2018. A single-item assessment for remaining mental resources: development and validation of the Gas Tank Questionnaire (GTQ). *Theoretical Issues in Ergonomics Science*, *19*(5), pp. 530–552.

Singh, I. L., Molloy, R. and Parasuraman, R., 1993. Automation-induced" complacency": Development of the complacency-potential rating scale. *The International Journal of Aviation Psychology*, *3*(2), pp. 111–122.

Tversky, A. and Kahneman, D., 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, *185*(4157), pp. 1124–1131.

Tversky, A., Kahneman, D. and Slovic, P., 1982. *Judgment under uncertainty: Heuristics and biases* (pp. 3–20).

Wojton, H. M., Porter, D., T. Lane, S., Bieber, C. and Madhavan, P., 2020. Initial validation of the trust of automated systems test (TOAST). *The Journal of social psychology*, *160*(6), pp. 735–750.