Trust between Humans and AI

Aakriti Kumar

> Ginny!" said Mr. Weasley, flabbergasted. "Haven't I taught you anything? What have I always told you? Never trust anything that can think for itself if you can't see where it keeps its brain?"
>
> J.K. Rowling, Harry Potter and the Chamber of Secrets

1 Introduction

The ubiquity of AI systems in our daily lives is undeniable: we talk to AI assistants, we let algorithms drive our cars, we seek their recommendations on what to buy, and so on. While we have made significant progress across varied domains in building fairly accurate and efficient AI systems, in most cases there still exists a need for human supervision and/or intervention.

This need for collaboration between humans and AI is due to many reasons. On one hand is the complementary nature of their abilities. While AI can look through vast amounts of data and make mathematically precise inferences, it still lacks the human ability to understand abstract concepts and generalise with much less data. On the other hand, a critical consideration that necessitates such human supervision, especially in high-stakes decision-making, is that algorithms are not infallible. There have already been instances that expose biases in algorithmic recommendations due to limited or biased training data. People have also reported cases of faulty recommendations by algorithms due to technical glitches [2]. To effectively leverage complementary abilities and to efficiently mitigate algorithmic errors, we need to design systems that are well understood and appropriately trusted by the human user. To this end, researchers have emphasised the importance of improving model interpretability and explainability. These efforts are focused on conveying the working and final recommendation of the model in a way that facilitates human understanding of the model. However, recent work by Lakkaraju and Bastani [26] and Bansal et al. [3] have shown that supplementing algorithmic decisions with more information or explanations doesn't necessarily help the human user make better decisions. One possible explanation for this observation is that humans are unable to build trust calibrated to the ability of the algorithm.

As Huang and Fox [21] suggest, decisions made in the real world are based on a mixture of rational calculations (within the limits of the information and mental resources available) and trust. While interpretablity efforts strive to make a model more understandable, they do not actively account for human reliance or trust in the model. In this review, we highlight the need to think about human trust when designing for effective collaboration between humans and machines. We review work on human-machine interaction with a focus on understanding how and when humans trust machines [1]. The paper is organized as follows. In Section 2, we briefly review the different ways trust is defined across fields. In Section 3, we review some popular models of trust between humans. In Section 4, we move to a discussion of trust between humans, the different factors that affect human trust in machines, and some models of trust between humans and machines. In Section 5, we conclude with some thoughts on the possible research directions.

2 What is Trust?

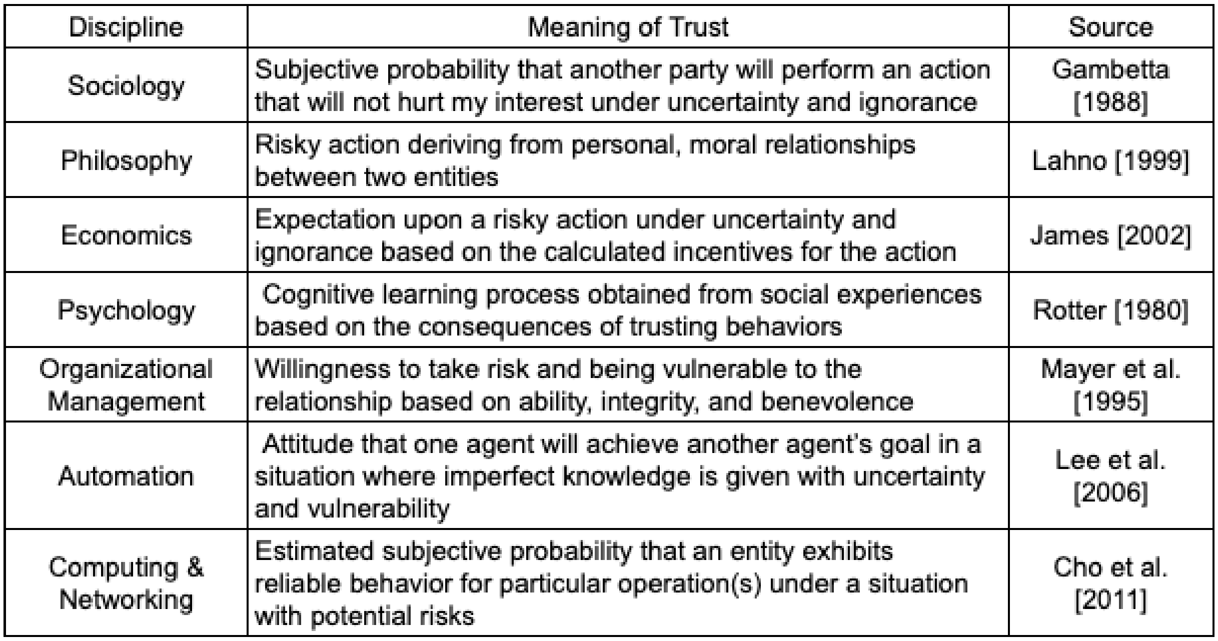
Trust is essential to the functioning of our society. Whether it is at our workplace, at home, or on the roads, everyday we implicitly trust others to more or less do what we expect them to do. We trust other drivers on the road to follow the rules, we trust our coworkers to work on tasks assigned to them, we trust our loved ones to look out for us, and so on. This trust is based on a combination of our past experiences and some assumptions about the world we inhabit. Without trust, the efficacy and efficiency of our day-to-day life would be severely impaired.

Researchers across domains have attempted to formalize trust. Cho, Chan, and Adali [11] provide a comprehensive survey of how trust is defined across disciplines. Based on the many definitions shown in Figure 1, the authors summarize trust as "the willingness of a trustor to take risk based on a subjective belief that a trustee will exhibit reliable behavior to maximize the trustor's interest under uncertainty of a given situation based on the cognitive assessment of past experience with the trustee".

House [20] was one of the first to capture the multidimensionalality of trust by defining it as a combination of three specific expectations: (1) a general expectation of the persistence of the natural (the expectation that natural physical laws are constant) and the moral social orders (we expect human life to survive, and mankind and computers to be good and decent); (2) a specific expectation of technical competence of the trustee; (3) an expectation that the trustee is responsible, will carry out their duties and in situations where it is needed the trustee will place others' interests before their own.

Rempel, Holmes, and Zanna[38] provide an alternate but important account of trust as a dynamic expectation that undergoes predictable changes as a result of experience in a relationship. Early in a

[1] We use the terms machine, AI, algorithm, and decision aid interchangeably to describe a black-box model.



Discipline	Meaning of Trust	Source
Sociology	Subjective probability that another party will perform an action that will not hurt my interest under uncertainty and ignorance	Gambetta [1988]
Philosophy	Risky action deriving from personal, moral relationships between two entities	Lahno [1999]
Economics	Expectation upon a risky action under uncertainty and ignorance based on the calculated incentives for the action	James [2002]
Psychology	Cognitive learning process obtained from social experiences based on the consequences of trusting behaviors	Rotter [1980]
Organizational Management	Willingness to take risk and being vulnerable to the relationship based on ability, integrity, and benevolence	Mayer et al. [1995]
Automation	Attitude that one agent will achieve another agent's goal in a situation where imperfect knowledge is given with uncertainty and vulnerability	Lee et al. [2006]
Computing & Networking	Estimated subjective probability that an entity exhibits reliable behavior for particular operation(s) under a situation with potential risks	Cho et al. [2011]

Figure 1: Multidisciplinary Definitions of Trust (Source: [11])

relationship, the trustor bases their trust upon the predictability of the trustee's behaviours. Later in a relationship, trust is based on the "attribution of a dependable disposition" or reliability of the trustee which is judged by the trustor through accumulated behavioural evidence. The final stage is marked by the development of faith, i.e, the confidence that the trustee will continue to remain dependable and predictable. Faith is strengthened by events which indicate intrinsic motivation of the trustee to remain in a relationship.

Another useful characterisation of trust provided by Marsh [32] separates trust into three distinct types: (1) Basic Trust - the disposition of a person to trust something new that is encountered. This trust is based on an individual's life experiences and is said to eventually become a stable personality characteristic. It is expected that individuals with a greater basic trust will be more trusting of another agent on initial contact when compared to someone with lower basic trust. (2) General Trust - the overall trust an agent places in another agent. This is not with respect to a specific situation or task. (3) Situational Trust - an agent's trust in another agent in relation to the context of the interaction.

While there are different ways trust plays into our everyday interactions, the key takeaways from the many definitions of trust are: First, trust is an expectation of or confidence placed in or reliance on the other, i.e, trust is always in relation to an 'other' - we trust in someone or something [34]. Second, trust is important to any form of collaborative work [11]. Third, trust implies there is risk associated with the task at hand and uncertainty associated with the trustee [32]. Fourth, trust is oriented towards future rewards, behaviours or events. Fifth, trust is dynamic - it is built over repeated interactions [32]; it grows with cooperation, and diminishes with betrayal.

3 Models of Trust between Humans

The previous section discussed a handful of the many perspectives on trust. We established that trust does not exist in a void, but requires two agents, a trustor and a trustee, to interact repeatedly to accomplish a task in the face of uncertainty and incomplete information. In this section, we review some models of trust between a human trustor and a human trustee.

A popular paradigm that has been used to study decision making between two humans with differentiated roles is the Judge-Advisor system (JAS) [39]. The JAS paradigm makes a distinction between one or more advising agents or advisors who provide recommendations and information, and a judge who makes the final decision. In this model, the judge is assumed to have lower expertise than the advisor. The judge is dependent on the advisor and hence trust in the advisor is of importance to the judge. Through a series of experiments, Sniezek and Van Swol[39] highlight the importance of cues such as high confidence ratings by Advisors on the Judges' ratings of trust and their tendency to follow advice. They also showed that the level of trust a judge has in an advisor is directly related to the degree to which the Judge took the advice into account and in the Judge's confidence in their decision.

Economists investigate behavioral manifestations of trust through games such as the prisoner's dilemma and the trust game. Berg, Dickhaut, and McCabe [4] proposed the trust game to measure trust between two agents. In this game, an agent (A) is given an initial sum of money x. In the first step of the game, A (the trustor) can choose to share a fraction c of the money with B (the trustee) or keep the money to themselves. The game ends if A chooses to not share any money with B. However, if A decides to share some amount with B, B receives three-fold the amount A transfers and the game continues. In the second step, B is given an option to reciprocate the gesture and share any amount with A. In a repeated game, it is in the interest of the two players to cooperate. Prisoner's Dilemma is another popular game which requires players to repeatedly decide whether to cooperate with their partner or to defect. It is different from the trust game in that the players make their decisions simultaneously and hence can only base this decision on their interaction history with the other player.

Kennedy and Krueger [24] use a version of Berg's trust game as a task to investigate trust and captured both - participants' behavioral data and their brain activations under separate Magnetic Resonance Imaging (MRI) scanners. They implemented a series of models using the ACT-R framework. They first implemented a "like-me" model in which the first player tries to infer what the second player would do by placing themselves in the other participant's position. This model consistently selected a non-trusting strategy for the first player and a defect strategy for the second player. This prediction did not match the observed data where participants cooperated most of the time. They next implemented a model of 'unconditional trust' where both participants always cooperate. This matched the strategy for 16 out of 22 pairs of participants. The authors hypothesise that allowing for some randomness or a tit-for-tat strategy might produce results closer to the observed data.

Zak, Kurzban, and Matzner [43] also conducted a study that used a one-shot trust game where participants were given a single intranasal dose of oxytocin or a placebo. They found that oxytocin helped humans overcome their natural aversion to uncertainty in the behaviour of other.

4 Trust between Humans and Machines

We now move to a discussion of the dynamics of trust between humans and machines. More often than not, humans work in teams of varying sizes to accomplish a wide variety of tasks. The industrial revolution greatly altered the structure of collaborative work by introducing machines in a previously human-dominated system. Now, AI is slowly permeating areas that were hitherto thought to be exclusively dependent on human subjectivity and expertise. From doctors who look towards binary classifiers to decide which patients to send to outpatient programs [23], to courts using risk assessment tools to estimate if criminal defendants will engage in unlawful behavior in the future [17], humans are increasingly reliant on complex algorithms to support their decision making and everyday workflow.

Collaboration between agents is a social process and human-machine teaming is no different. Hence, trust calibrated to the machine's ability is critical to effective collaboration between humans and machines. Muir [34] extends work by [20] and [38] on trust between humans and generalise it to trust between humans and machines. Muir proposed that trust in a decision aid is calibrated according to (1) predictability: how predictable are the aid's recommendations, (2) dependability: how dependable is the decision aid (which they expect can be inferred by a summary statistic of accumulated behavioural evidence), (3) faith: when working with AI, humans lack a complete understanding of the system's working but they still work with it because they appreciate the vastness of the problem and possible outcomes and realize that their own knowledge of the system is incomplete. These factors underlie 'a leap of faith' on the part of the human.

Hoff and Bashir [19] classify trust in an autonomous system into three categories: dispositional, situational, and learned. Dispositional trust is based on characteristics of the human. Merritt and Ilgen [33] suggest that humans have a general propensity to trust or distrust a machine just as they have have a general propensity to trust or distrust another person. Factors that influence dispositional trust do not vary greatly with time, but they impact human decision-making during interactions with an autonomous system. Situational trust is a result of a combination of factors that are external to the human (task difficulty, potential risks) and those that are internal to the human (self-confidence, expertise). Finally, learned trust is based upon a human's overall experience with the autonomous system.

A related but important specification of appropriate trust behavior is provided by Lee and See [27]. They describe mismatches between trust and the capabilities of automation in terms of (1) calibration: the correspondence between a person's trust in the automation and the automation's capabilities, (2) resolution: the precision with which a judgment of trust differentiates levels of automation, and (3) specificity: the degree to which trust is associated with a particular component or aspect of the automated system. We restrict our discussion in this paper to the calibration of trust.

4.1 Pitfalls and Biases

The advice-taking literature has shown evidence that humans discount advice from peers [5] and tend to rely more on their own judgment. Furthermore, an extensive literature on overconfidence repeatedly demonstrates that individuals report excessive unwarranted confidence in their own judgment relative to that of their peers [15]. Working with a machine is no different. Research has shown that humans are susceptible to a variety of misjudgements and biases when seeking advice from machines.

Parasuraman and Riley [36] describe inappropriate reliance on machines as misuse, disuse and abuse of automation. *Misuse* refers to failures that occur due to over-reliance on automation. *Disuse* refers to the failures that occur when humans rejects the help of automation when it could have been useful. *Abuse* refers to incorrect deployment of automation by the designers and managers - for example, using automation where human input is critical. Automation abuse can also increase misuse and disuse of automation by humans.

Researchers have identified two competing cognitive biases that humans are likely to display when working with machines: algorithm aversion and automation bias. Dietvorst, Simmons, and Massey [12] define *algorithm aversion* as the tendency of a human to disregard the recommendations of a machine after observing that it made a mistake. In contrast, *automation bias* is the tendency to over-rely on machine recommendations [16]. Both these biases lead to sub-optimal outcomes. Hence, calibrating human trust to match the algorithm's prediction accuracy and general ability is crucial for effective human-machine teamwork.

4.2 Factors that affect Trust

In this section, we identify and summarize factors that may affect a human's trust in a machine. These factors can be categorized as relating to properties of the different components of this collaboration:(1)the

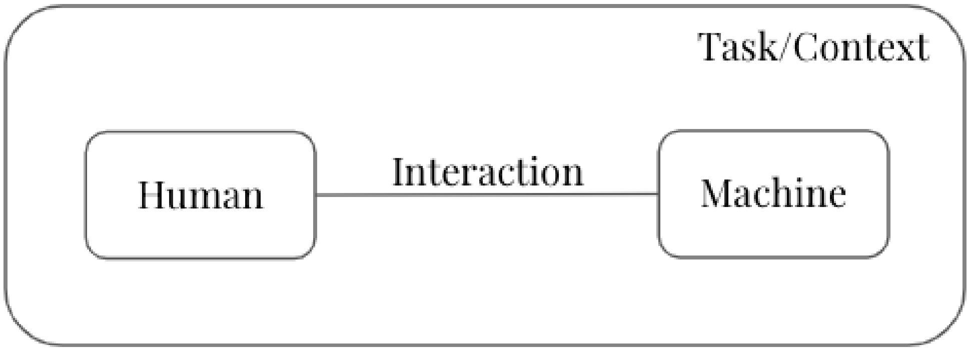


Figure 2: Components of a typical human-machine collaboration setting

human, (2)the machine, (3)the task or context in which the human and machine collaborate, and (4)the interaction between the human and the machine. Fig 2 shows the components of a typical human-machine collaboration setting.

4.2.1 Properties of the Human

Human Expertise Medical diagnostic decisions, recidivism judgements, child welfare, and fire risk assessment are a few examples of places where AI attempts to supplement human experts. However, there is evidence that experts incorporate any advice differently than non-experts. Jacobson et al. [22] found that expert attorneys are less likely than law students to give weight to advice in a verdict estimation task. found that human experts tend to dismiss algorithms. Logg, Minson, and Moore [30] second this finding by showing that experts heavily discounted advice from all sources including algorithms. In a recent paper, De-Arteaga, Fogliato, and Chouldechova [2] show that call-workers in a child-welfare center were able to override the risk assessment tool's decisions when the tool displayed mis-estimated scores. The authors hypothesize that the expertise of the call-workers had a role to play in the their ability to make predictions independent of the incorrect recommendations shown by the tool. However, in contrast to [31], the authors did not find the workers to be dismissive of the tool completely as there was an increase in overall performance of the workers.

Individual Differences It is well-established that individual differences affect the trust behavior of people. Merritt and Ilgen [33] show that individual differences in personality (extraversion) and the predisposition to trust machines play an important role in trust in automation use. People have different baseline propensity to trust machines and trust in general. The authors demonstrate the need to consider individual differences when discussing trust in machines. In their paper, they capture via trust ratings the difference in perception of the machine by 255 users while keeping the machine characteristic constant. There is some evidence that perceived self-confidence may influence a human's choice of using or discarding machine help especially when the human makes a mistake without the machine's help [29] [28]. There is also evidence that familiarity with mathematics and ML make people more likely to listen to algorithmic advice [40].

Demographics Studies have shown differences in trust behavior between people of different cultures and age groups. Akash, Jain, and Misu [1] worked with a geographically diverse set of particpiants to investigate dispositional trust or trust propensity of an individuals. They replicate a finding in the literature that Americans trust autonomous systems less than Mexicans and Indians, respectively. They also found that mistakes by the AI system induced stronger distrust in US participants than in Indian participants. In terms of differences across genders, the authors found that men trusted AI more than women and that women are more variable in their trust towards AI. However, other studies have shown mixed results for gender differences in trusting behavior towards AI. Haselhuhn et al. [18] found evidence that women show a smaller dip in trust in the AI when compared to men when they see algorithmic errors. There is no conclusive evidence that genders differ in their trust attitudes towards AI.

Training De-Arteaga, Fogliato, and Chouldechova [2] investigated how call-workers at a child welfare hotline service integrated recommendations from a risk assessment tool into their decision process. Call-workers decide whether a call concerning potential child neglect or maltreatment should be screened in for investigation. Around a year after deployment, it was discovered that a technical glitch in the tool had caused some of the scores shown to the workers to be mis-estimated. The data from this welfare center presents a unique opportunity to look at algorithm aversion, automation bias when experts work alongside algorithms in a natural setting. The tool was deployed to help workers identify high-risk cases when the information communicated in a call was inconclusive. It only provided a final risk score to the call-workers. Note that no explanations for these score were provided to the workers In a retrospective analysis of the data, authors found that call-workers almost always dismissed the machine's recommendation on instances where the risk score had been considerably underestimated. However, they did see a rise in the overall accuracy of the call-workers. Based on conversations with the call-workers, the authors hypothesize that training may have had a key role to play in their well-calibrated incorporation of the tool's recommendations. Call-workers received explicit instructions to treat the tool's recommendation as complementary information and not rely on the score as a replacement of their own judgement. We have already established that framing the decision to use an algorithm as an all-or-nothing decision is counterproductive and may lead to higher levels of algorithm aversion. The effect of framing a machine's contribution to the collaborative work, and communicating the machine's role and scope to the human (in the form of training) on human integration of machine recommendations is a promising avenue of further investigation.

4.2.2 Properties of the Machine

Performance A machine teammate is helpful if it decreases the human's workload, speeds up the task or increases the accuracy of the human in the task. We use performance as an umbrella term to capture different properties of a machine such as accuracy and predictability. Higher accuracy models are preferred to lower accuracy models. More importantly, more predictable models are preferred to less predictable models. Trust can develop when a systematic fault occurs for which a strategy can be developed [27]. The influence of a mistake on trust depends on both the magnitude of the error and the how unexpected or unpredictable it was. A small but unpredictable fault affects trust more than a large fault of constant error.

Interpretability and Transparency The literature on advice taking shows a robust effect of discounting advice from others because people don't have access to others' reasoning. Model interpretability is essential to establishing a useful working relationship between a machine and a human. However, supplying more information about the workings of the machine to the human has not shown very promising results. Bansal et al. [3] saw no improvement in team performance when they added explanations to model output. Suresh, Lao, and Liccardi [40] showed that participants over-relied on both correct and incorrect machine recommendations even when they were independently able to do that same task correctly. The authors also found that people were likely to accept a machine's recommendation despite being given information that points to very low confidence of the machine in it's recommendation. While increased transparency has been shown to improve human trust in the AI, it increases the workload of the human. Akash, Jain, and Misu [1] make a case for optimising transparency or using it sparingly. These results emphasise the need to integrate the humanness of stakeholders into model interpretability design considerations. As [27] point out, the objective is not to design systems to increase reliance or trust but to design for appropriate reliance and trust.

4.2.3 Properties of the Task or Context

Difficulty Gino and Moore [15] reiterate a robust finding in the advice taking literature that people put too little weight on advice from others when the task is easy and too much weight when the task is difficult. Complex tasks and higher workloads cause increased stress on cognitive capacity. While some studies have shown that humans may become overuse AI advice under increased workload, some others have found that increased trial difficulty improved performance. This suggests difficulty can motivate closer inspection of the task and decrease complacency [16].

Objectivity Castelo, Bos, and Lehmann [9] found that people clicked on ads for algorithm-based advice less than on ads for human-based advice when the task is subjective (dating advice), but not when the task is objective (financial advice). Logg [31] found that people seek algorithmic advice for objectives decisions and human advice for subjective decisions. This is in sync with the finding that people view machines and AI systems as more rational and objective than humans. Researchers have demonstrated that people exhibit algorithm aversion in subjective domains. Participants in work by Promberger and Baron [37] preferred a medical diagnosis from a doctor and reported feeling less responsible for the decision when taking the advice from the doctor. Tasks involving recommendations about books, movies, jokes also showed algorithm aversion [25]. Hence, designing for appropriate reliance requires thinking critically about the application's perceived difficulty and objectivity.

Risk Trust presupposes a situation of risk. Taking a risk reinforces trust that is there already if there is a favourable outcome of collaborating. In the event that an unfavourable outcome is observed, the risk associated with trusting is exposed, and trust decreases accordingly. If it was high initially, and the risk of rejection was great, then rejection causes a large loss of trust [6]. This prediction is in line with what Logg [31] show happens when people see a machine err early on in their interaction. Risk associated with a task can also be used as a proxy for how important the task is perceived to be.

4.2.4 Properties of the Interaction

Decision Autonomy Dietvorst, Simmons, and Massey [13] found that people are less likely to display algorithm aversion when working with an imperfect algorithm if they have some control over the final decision. In a series of experiments that allowed participants to modify the algorithm's forecast to different extents, the authors observed that people were more likely to positively weight and use the algorithm's recommendation as long as they were able to incorporate their own input and participate in the ultimate decision. The authors also highlight that participants were relatively insensitive to the amount by which they could modify the algorithm's forecasts.

Another configuration of decision making hierarchy is where the human is allowed a choice between taking advice from a machine or another human/ human expert. Logg [31] find that if available, humans prefer to take advice from human experts over algorithms. However, in some follow-up work [30], the authors found competing evidence that people trust predictions more when they believe that the predictions come from an algorithm as opposed to a human even in 'subjective' domains such as predicting music popularity and romantic matches. The authors observed that this preference for the algorithm was not very apparent when people were given the choice between using an algorithm's prediction and using their own prediction (as opposed to a prediction from another human).

Adaptive User Interfaces It is known that well designed interfaces can increase user acceptance and trust of the system. Content based image retrieval (CBIR) system proposed by Wan et al. [42] is one such tool. CBIR systems index and retrieve images based on automatically learned similarity metrics and are widely used to aid doctors. Doctors can use an image as a query for retrieving similar images from previously diagnosed patients. Cai et al. [8] investigated the use of CBIR by pathologists. They allowed pathologists to communicate what types of similarity are most important for each instance hence allowing for customised search based on the users need. Pathologists reported increased diagnostic utility of the images and higher trust in the algorithm. Another way that user custom user-interfaces can improve performance is by adapting to the levels of trust of the user. Estimates of a user's trust to can guide a system's decision to engage in trust dampening/enhancing actions [41]. Akash, Jain, and Misu [1] also demonstrate manipulating human's trust and workload dynamics by varying the automation's transparency - the amount of information provided to the human.

Interaction History Trust exists because we interact with others more than once. Algorithm aversion, a well-established finding in the literature, indicates a loss of trust after a human sees the algorithm make a mistake. Initial interaction and negative interactions have a greater impact on trust than interactions later in the exchange. Errors observed early on in the interaction result in substantial reliance reduction, whereas encountering an error later in the interaction affects reliance only temporarily. Dietvorst, Simmons, and Massey [12] and Logg [31] showed that people relied more on algorithms than themselves before they were given any performance feedback. However, the authors also observed that this effect was diluted when users were given more control over how to use the algorithm's predictions. Lee and See [27] too have emphasized displaying past performance of the machine to the user.

Feedback The only way to develop and adjust one's trust in another agent is to see the result of an exchange with the other agent. Feedback or reward realisation is critical to learn and calibrate expectations of the other. We know from [12] that seeing an algorithm err makes people less likely to rely on it compared to themselves or another human's advice even when they see the algorithm outperform the themselves/the other human.

The factors discussed above are not an exhaustive set of factors that affect human trust in machines. For example, we omit discussion on how anthropomorphizing machines affects human perception of and trust in the machine.

4.3 Models of Trust in Human-Machine Teams

We now move to a discussion of a few models of trust between a human and a machine. Here, the trustor is a human and a machine is the trustee and they interact repeatedly to accomplish a task in the face of uncertainty and incomplete information.

Muir [34] propose the following definition of trust based on Barber's characterisation of trust between human agents to trust between a human and a machine: "Trust (T) is the expectation (E), held by a member (i) of a system, of persistence (P) of the natural (n) and moral social (m) orders, and of technically competent performance (TCP), and of fiduciary responsibility (FR), from a member (j) of the system, and is related to, but not necessarily isomorphic with, objective measures of these qualities". They propose a linear additive model of trust which takes into account the three expectations held by the human (P, TCP, FR) and their interactions.

Chen et al. [10] propose a computational model to integrate trust into robot decision making. They model human trust as a latent variable in a partially observable markov decision process (POMDP). A POMDP is a framework to model sequential decision making under uncertainty. The proposed model allows the robot to infer trust of its human teammate, reason about the effect of its own actions on human trust and hence enables it to choose actions that improve team performance. their model does two things: evaluate trust dynamics of the human, i.e, evaluate how human trust evolves over repeated interaction with the robot and 2) how human trust maps to actions. While their model can accommodate a variety of trust dynamics and human decision models, they employ a data driven approach and learn these models from data which they collect through a table clearing task performed by humans and robots together. They start with an intuitive and well established assumption that human trust in a machine evolves based on the performance of the machine. The task involves a robot clearing a table by picking up and placing a few different items (specifically, 1 fish can, 1 wine glass and 3 plastic water bottles). The human can do one of two things: intervene or allow the robot to clear the table object by object; half the times robot picks random policy from all possible policies and the other half it selects a policy from a set of prespecified policies that were of interest to the authors. Continuous reports of human trust of the human in the robot is captured. The authors assume that humans follow the softmax rule to decide which action to take, they specify two competing behavioral models for the human: (1) Trust free behavioral model (TFBM) - where the human decides between the two actions probabilistically based on the expected reward of the action. This model assumes that the human's trust stays constant in the robot and is unaffected by their interactions (2) Trust based behavioral model (TBBM)

- where the human's belief about the robot's success changes over time and depends on the human's trust in the robot. TBBM showed higher team performance than TFBM supporting the author's intuition that incorporating trust considerations in choosing strategies improves team performance.

Visser et al. [41] propose a longitudinal approach to trust development and calibration in human robot teams. They attempt to define and measure 'relationship equity' that quantifies the 'goodwill' between the human and robot teammates. Their work was inspired by Gottman's work on calibrating trust in couples by analyzing moment-to-moment interactions over longer periods of time and identifying specific trust repair strategies to be used when trust is too low, and trust dampening strategies when trust is too high.

Logg [31] proposes a 'theory of machine' analogous to 'theory of mind' which posits that people build mental models of other people by interpreting their external actions. It also emphasises the importance of understanding the others' intentions as a key component to understanding others' minds. Logg [31] argues that as interaction between humans and artificially intelligent systems increases, there is a need to understand how human build mental models of machines. Similar to theory of mind, theory of machine requires people to think about the internal processes of an AI agent.

5 Future Directions

We highlight some work being done to improve human-machine collaboration and identify avenues for future research.

Delayed Feedback Most work on AI-assisted human decision-making has focused on characterising how people create and update their beliefs about the AI system based on immediate feedback/reward. However, in the real world, the rewards are often delayed. For example, a judge may use recidivism risk predictions made by an algorithm to inform parole decisions. Or, a doctor may take inputs from a binary classifier to make diagnostic decisions. In such scenarios, the reward or penalty is observed after a variable time period. There is a need to investigate the effect of delayed reward on the evolution of the human's trust in the AI.

Adaptive allocation Parasuraman, Mouloua, and Molloy [35] demonstrate that over-reliance can be reduced by adaptive task allocation. They advocate the need for active involvement rather than passive monitoring to obtain calibrated performance. In a series of experiments, a human interacted with an automated system where the control was returned to the human for a short while in the whole duration of the task. The authors observed increased accuracy in automation monitoring by humans after a period of full control. This is a promising direction for future work that remains under-explored. What are good strategies of task allocation to keep humans engaged and alert?

Adaptive explanations The conversation around the benefit of providing explanations about a machine's output to the human user is contentious. Some studies report improvements due to explanations only when the AI outperforms both the human and the best human-AI team. In contrast, Bansal et al. [3] failed to improve team performance by using adaptive explanation strategies. They observed that explanations increased reliance on recommendations even when they were incorrect. Zhang, Liao, and Bellamy [44] found that displaying confidence scores help calibrate people's trust in the machine but displaying explanations didn't have much of an effect. Similar to adaptive task allocation discussed above, there is a need to further investigate when, what and how much information should be supplied to the human to facilitate calibrated trust in the machine.

Tasks and the real-world Doshi-Velez and Kim [14] propose a taxonomy of evaluation approaches for interpretability research: 1) Application-grounded evaluation involving real humans and real tasks, 2) Human-grounded evaluation involving real humans and simplified tasks, and 3) Application-grounded evaluation with no humans and proxy tasks. The top two levels of this hierarchy are of particular interest to us. Human-grounded evaluation calls for real humans to be paired with simplified versions of actual real-world applications or proxy tasks. This is in-line with most research that is done in Cognitive Science. However, Buçinca et al. [7] demonstrate via a proxy task and an actual decision making task that evaluations using proxy tasks did not predict human performance on actual decision making tasks. The authors argue that tasks must be designed in a way that keeps decision-making the focus instead of forcing participants to pay attention to the AI and it's explanations. They argue that such a design is closer to a realistic decision making task where the primary concern is the task at hand and one can decide when how much to attend to an AI's input. Designing good tasks to gain insight about human-behavior in real-world decision scenarios is crucial to the design of decision aids. There is a need to think critically about the design of proxy tasks designed to capture real-world behaviors.

References

[1] Kumar Akash, Neera Jain, and Teruhisa Misu. "Toward Adaptive Trust Calibration for Level 2 Driving Automation". In: *Proceedings of the 2020 International Conference on Multimodal Interaction* (Oct. 2020). DOI: 10.1145/3382507.3418885. URL: http://dx.doi.org/10.1145/3382507.3418885.

[2] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. "A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–12.

[3] Gagan Bansal et al. "Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance". In: *arXiv preprint arXiv:2006.14779* (2020).

[4] Joyce Berg, John Dickhaut, and Kevin McCabe. "Trust, reciprocity, and social history". In: *Games and economic behavior* 10.1 (1995), pp. 122–142.

[5] Silvia Bonaccio and Reeshad S Dalal. "Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences". In: *Organizational behavior and human decision processes* 101.2 (2006), pp. 127–151.

[6] Sd Boon and J. Holmes. "The dynamics of interpersonal trust: uncertainty in the face of risk". In: 1991.

[7] Zana Buçinca et al. "Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems". In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 2020, pp. 454–464.

[8] Carrie J Cai et al. "Human-centered tools for coping with imperfect algorithms during medical decision-making". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–14.

[9] Noah Castelo, Maarten W Bos, and Donald R Lehmann. "Task-dependent algorithm aversion". In: *Journal of Marketing Research* 56.5 (2019), pp. 809–825.

[10] Min Chen et al. "Planning with trust for human-robot collaboration". In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 2018, pp. 307–315.

[11] Jin-Hee Cho, Kevin Chan, and Sibel Adali. "A survey on trust modeling". In: *ACM Computing Surveys (CSUR)* 48.2 (2015), pp. 1–40.

[12] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. "Algorithm aversion: People erroneously avoid algorithms after seeing them err." In: *Journal of Experimental Psychology: General* 144.1 (2015), p. 114.

[13] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them". In: *Management Science* 64.3 (2018), pp. 1155–1170.

[14] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608* (2017).

[15] Francesca Gino and Don A Moore. "Effects of task difficulty on use of advice". In: *Journal of Behavioral Decision Making* 20.1 (2007), pp. 21–35.

[16] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. "Automation bias: a systematic review of frequency, effect mediators, and mitigators". In: *Journal of the American Medical Informatics Association* 19.1 (2012), pp. 121–127.

[17] Ben Green and Yiling Chen. "Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency.* 2019, pp. 90–99.

[18] Michael P Haselhuhn et al. "Gender differences in trust dynamics: Women trust more than men following a trust violation". In: *Journal of Experimental Social Psychology* 56 (2015), pp. 104–109.

[19] Kevin Anthony Hoff and Masooda Bashir. "Trust in automation: Integrating empirical evidence on factors that influence trust". In: *Human factors* 57.3 (2015), pp. 407–434.

[20] James S House. *The Logic and Limits of Trust. By Bernard Barber. Rutgers University Press, 1983. 190 pp. Cloth, 27.50;paper, 9.95.* 1985.

[21] Jingwei Huang and Mark Fox. "An ontology of trust - Formal semantics and transitivity". In: *Proceedings of the ACM Conference on Electronic Commerce* (Jan. 2006), pp. 259–270. DOI: 10.1145/1151454.1151499.

[22] Jonas Jacobson et al. "Predicting civil jury verdicts: How attorneys use (and misuse) a second opinion". In: *Journal of Empirical Legal Studies* 8 (2011), pp. 99–119.

[23] Ece Kamar. "Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence." In: *IJCAI* (2016), pp. 4070–4073.

[24] William G Kennedy and Frank Krueger. "Building a cognitive model of social trust within ACT-R". In: *2013 AAAI Spring Symposium Series.* 2013.

[25] Jon Kleinberg et al. "Human decisions and machine predictions". In: *The quarterly journal of economics* 133.1 (2018), pp. 237–293.

[26] Himabindu Lakkaraju and Osbert Bastani. "" How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* 2020, pp. 79–85.

[27] John D Lee and Katrina A See. “Trust in automation: Designing for appropriate reliance”. In: *Human factors* 46.1 (2004), pp. 50–80.

[28] John Lee and Neville Moray. “Trust, control strategies and allocation of function in human-machine systems”. In: *Ergonomics* 35.10 (1992), pp. 1243–1270.

[29] Stephan Lewandowsky, Michael Mundy, and Gerard Tan. “The dynamics of trust: Comparing humans to automation.” In: *Journal of Experimental Psychology: Applied* 6.2 (2000), p. 104.

[30] Jennifer M Logg, Julia A Minson, and Don A Moore. “Algorithm appreciation: People prefer algorithmic to human judgment”. In: *Organizational Behavior and Human Decision Processes* 151 (2019), pp. 90–103.

[31] Jennifer Marie Logg. “Theory of machine: When do people rely on algorithms?” In: *Harvard Business School working paper series# 17-086* (2017).

[32] Stephen Paul Marsh. *Formalising trust as a computational concept.* 1994.

[33] Stephanie M Merritt and Daniel R Ilgen. “Not all trust is created equal: Dispositional and history-based trust in human-automation interactions”. In: *Human Factors* 50.2 (2008), pp. 194–210.

[34] Bonnie M Muir. “Trust between humans and machines, and the design of decision aids”. In: *International journal of man-machine studies* 27.5-6 (1987), pp. 527–539.

[35] Raja Parasuraman, Mustapha Mouloua, and Robert Molloy. “Effects of adaptive task allocation on monitoring of automated systems”. In: *Human factors* 38.4 (1996), pp. 665–679.

[36] Raja Parasuraman and Victor Riley. “Humans and automation: Use, misuse, disuse, abuse”. In: *Human factors* 39.2 (1997), pp. 230–253.

[37] Marianne Promberger and Jonathan Baron. “Do patients trust computers?” In: *Journal of Behavioral Decision Making* 19.5 (2006), pp. 455–468.

[38] John K Rempel, John G Holmes, and Mark P Zanna. “Trust in close relationships.” In: *Journal of personality and social psychology* 49.1 (1985), p. 95.

[39] Janet A Sniezek and Lyn M Van Swol. “Trust, confidence, and expertise in a judge-advisor system”. In: *Organizational behavior and human decision processes* 84.2 (2001), pp. 288–307.

[40] Harini Suresh, Natalie Lao, and Ilaria Liccardi. “Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making”. In: *arXiv preprint arXiv:2005.10960* (2020).

[41] Ewart J de Visser et al. “Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams”. In: *International journal of social robotics* 12.2 (2020), pp. 459–478.

[42] Ji Wan et al. “Deep learning for content-based image retrieval: A comprehensive study”. In: *Proceedings of the 22nd ACM international conference on Multimedia*. 2014, pp. 157–166.

[43] Paul J Zak, Robert Kurzban, and William T Matzner. “The neurobiology of trust”. In: *Annals of the New York Academy of Sciences* 1032.1 (2004), pp. 224–227.

[44] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. “Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 295–305.