The Impact of Transitivity, Similarity and Implicit Bias on Trust in AI

Adam Coscia
Georgia Institute of Technology
acoscia6@gatech.edu

Sarah Mathew
Georgia Institute of Technology
smathew64@gatech.edu

Marina Vemmou
Georgia Institute of Technology
mvemmou@gatech.edu

***Abstract*—Human civilization has been built upon the necessary foundation of trust between humans. With the rapid increase in deployment of Artificial Intelligence (AI) systems in real-world scenarios, the question of how to establish trust between humans and AI systems is of increased interest and urgency. In this work, we examine three aspects that have been studied as playing a role in trust establishment between humans: transitivity, similarity and implicit bias. We conduct a crowd-sourced study centered on a series of decision-making tasks and accompanying AI-generated decision suggestions in order to determine whether the aforementioned factors can inspire trust in the tasks' AI. We enhance our task-based study with surveys that further capture general attitudes towards AI and trust in it. Our results show that transitivity, similarity and implicit bias do not appear to play a significant role in establishing trust between humans and AI, but illuminate other important factors such as transparency and human agency. This work is a small but important step towards understanding the complex relationship between humans and AI, one that will remain prominent and we believe is of great importance in maximizing AI's contributions to society.**

***Index Terms*—trust, artificial intelligence, transitivity, implicit bias, similarity, recommendation systems**

I. Introduction

Trust is what enables relationships between people, systems, and society to function. It is required for cooperation, and for complex goals to be achieved. We take it for granted during various day-to-day activities, such as when we learn about a new topic, drive on the road, or see a health professional. In today's world of growing reliance on AI technology in these day-to-day activities, trust continues to impact the decisions we make when using AI. Thus, we seek to investigate how trust functions in human-AI relationships.

One aspect of trust we explore in this work is **transitivity**. This phenomena is when individuals are trusting towards an entity *given* that the people they know, or who are in their social circle, trust it too [21]. One example of transitivity in action is the referral process. If someone in an organization recommends an applicant, a hiring manager may take them more seriously than if they were a complete stranger. Another example of transitivity is product reviews. If someone in your social circle recommends a business or tool, it can help narrow the selection process. This also happens online, on platforms such as Amazon, where people rely on reviews to better understand how useful or reliable a product will be. Given that many people rely on such reviews to make decisions, our question is: *"how can trust be relayed when the reviewed entity is an artificial intelligence?"*

A different aspect we seek to explore is **similarity**. The concept of similarity is fundamental to cognitive science, as the basis of analogical thinking [12]. Humans find similarities between objects, situations and other humans and use them to decide about the best action based on past, similar experiences. Similarity as a factor of trust between humans has been mainly studied from the aspect of how much person A identifies with person B, and whether this can make A trust B more. When it comes to placing trust in products or tools, a positive association between an existing, trusted object and an unknown one can create the positive relationship required for trust. If a new movie comes out, and is described as as emotionally engaging as your favorite film, you will be inclined to see that movie. These kinds of comparisons are commonplace in everyday life, and we want to explore if they can be utilized to inspire trust in an AI-powered system.

More specifically, we are curious about if and how these phenomena play a role in in the context of **collaborating** with an AI. This kind of technology is being used more and more in society, and in more obvious ways, such as in recommendation systems and voice assistants. People and AI can augment each other's abilities, and it has the potential for adding immense value to businesses along dimensions of "flexibility, speed, scale, decision making, and personalization" [26]. How do people trust to use such a system, and can this trust be influenced by what others say about it, or what it is compared to? What would be needed for people to adopt and use these systems?

In this work, we present a user study that evaluates the effects of transitivity and similarity on trust in AI systems used to assist in decision making. Most users of AI are unlikely to have the technical knowledge required to evaluate it and judge its trustworthiness, and the "black box" nature of AI systems makes creating simple and easily understandable "trust" metrics difficult, especially for systems that are relatively new and not well-understood yet. We attempt to examine how two already existing cognitive mechanisms for establishing trust in humans and tools, transitivity and cognition, can be used in the context of AI. We examine how and when users presented with a decision-making task interact with a provided AI-assistant, and whether reviews or positive associations with other technologies affects their behavior in terms of trust. To further enrich our

insights into the matter, we supplement the task-based study with questions regarding general attitudes towards AI, and more specifically trust in AI and what affects it.

II. Related Work

A. Human-Centered AI (HCAI)

HCAI is a field where the goal is to understand how people interact with AI technologies, and how to use that interaction to improve the AI. Recently, Amershi et al. developed a set of 18 interaction design guidelines to consider while interacting with an AI-enabled system [1]. Some of the more relevant guidelines include learning from user behavior (G13) and showing why the system made a particular decision (G11). The latter guideline is an effort to improve explainability, which is a current problem in AI systems that will be covered in more detail in Section II-B.

[26] argue that human-AI collaboration, where both sides work in a complementary way, is the way AI should be used in the future as opposed to replacement; they share examples in industry where these partnerships exist.

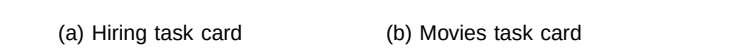
Adjacent to designing AI technologies is AI literacy, where the focus is on understanding and evaluating AI systems. Long and Magerko outline 16 competencies and 15 design considerations in AI literacy, and define it as "a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace" [15]. How well a person understands what AI is, and where it is, can affect how they approach a new technology.

Overall, for efficient interactions, the AI needs to have social intelligence [13] to understand what humans expect from interaction, which needs skills such as cooperation and theory of mind [3].

B. Explainable AI (XAI) and Bias

Our work is also related to trust in AI in general, especially when AI can be a "black box" from the perspective of the user, which may not be interpretable and/or explainable [9]. An AI is characterized as *interpretable* when humans can understand how the system made its decisions, but not why the criteria used were sensible. An *explainable* AI will provide an explanation as to why it arrived to a behavior or decision, but not how exactly it arrived there. Biran and McKeown show that when an AI system offers explanations for its stock suggestions, users are more likely to use it and benefit from it [4]. Ribeiro et al. [20] make an interesting distinction between *trusting a prediction* (whether a user trusts a given AI suggestion enough to follow it or take it into consideration), and *trusting a model* (whether a user trusts an AI to be deployed and produce a reasonable behavior). Although a user may gain confidence on a single prediction if an explanation is provided, trust on a model may require multiple prediction-explanation pairs to establish trust in deployment for real-world scenarios.

Although AI is already used in a plethora of domains, it can have harmful effects because of bias [23]. *Survival of the Best Fit* is an interactive application created to illuminate how bias can arise in the candidate selection process for a specific job. [6]. Users pick out candidates and at the same time, their choices are being learned by an AI. Since the dataset is small and each user has a specific strategy for hiring (affected by both conscious and unconscious factors), the AI makes similarly biased decisions that reject qualified candidates. This issue is relevant because it can be difficult to develop trust in a system that could be biased.



(a) Hiring task card (b) Movies task card

Fig. 1: Task cards which users selected during multiple rounds of each task.

C. Online Reviews, Recommendations

Trust for agents can be defined as a mental attitude, towards or a set of beliefs about another agent concerning a behavior to reach some goal [5]. Recommendations, or referrals, are ways for one agent to practice theory of mind and convince another agent to trust a third party. Referrals in the application process for example tend help applicants get into interviews and be hired [7].

To develop a better mental model of something, maybe a product, users read online consumer reviews (OCRs). This is to develop trust that the product will behave well or as expected, but it also requires that the reviews are trustworthy as well.

Consumers have looked at different kinds of features to assess the quality (usefulness) of reviews. These include contextual information (argument quality, user rating, if it's about core services or interpersonal aspects, valence and order of valence [24], length), intrinsic information (reviewer anonymity and pronoun use) and representational (grammar, spelling) [17]. Other aspects of the reviewer a consumer may measure are positivity, involvement, experience, reputation, competence and sociability [2] as well as perceived background similarity when the goal is low-stakes [19]. Over the set of reviews, quantity of reviews and overall rating is also considered [16]. Nusumasondjaja et al. found that positive reviews contribute to initial impressions of trust [14]. Maslowska et al. identifies a "too good to be true" phenomena where products with high reviews (4.5-5 / 5) are seen as not trustworthy [16].

To our knowledge, interactions between people and reviews on AI technologies have not been studied.

III. Methodology

In order to study the effects of transitivity, similarity and implicit bias on decision-making in the presence of AI, we conducted a crowd-sourced study centered on a series of decision-making tasks and accompanying decision suggestions. In this section, we describe the two tasks used to asses the

Fig. 2: Sample movies domain task given to participants, including an optional choice of utilizing an AI technology to help make the decision.

attitude of users towards utilizing an AI assistant (Section III-A) and the implementation of an interface to engage users with decision-making (Section III-B).

A. Tasks and Datasets

The general look of the cards and options for each task are based on the card design from the Survival of the Best Fit bias game. We decided to create two tasks with two different domains, hiring and movie selection. Our goal was to minimize domain effect (users exhibiting a behavior that is domain-related that they would not apply to other domains), and to examine one domain that has higher stakes (who gets hired or not) and requires the user to show higher responsibility, and another that has lesser consequences (picking movies for profit). Another design choice we made was that we wanted each task to have non-obvious answers. This was to encourage users to 1) consider asking for the AI's suggestion and 2) consider using the AI's suggestion.

1. Hiring

Task. The hiring task domain was adapted from the bias game. Each user was presented with a number of candidate portfolios, that included a set of parameters relevant to this candidate, and was asked to select a candidate to move forward in the interview process. The parameters used on the cards were inspired from the ones available in the bias game; these include school prestige, skill level, work experience, and ambition. We re-named ambition to confidence/ambition to make it clearer, and added two additional binary features: if the applicant was referred, and if they had applied before to the company. We added these because they seemed like features that a hiring manager would consider. Referrals also tend help applicants get into interviews and be hired.

Dataset. The applicant cards were made in Canva [18], with an example shown in Figure 1a. The level for each feature was determined with a random number generator set between 1 and 10. This was done to avoid biases in selecting random numbers manually for each category. The "referral" and "applied previously" categories were determined with a random binary number generator. After the feature values were all generated, they were grouped together in a pseudo-random way. They were first labeled by how many features had better than a 50% level. So Figure 1a for example would be labeled in a "3" group because Work Experience, Skill, and Ambition/Confidence were all over 50%. Cards were grouped this way to prevent situations where the answer seemed obvious, i.e. some applicants were more qualified than others by having more stronger features. There are three to four applicant cards put together in one task group. Once all groups were made, another random number generator was used to determine what choice the bot would suggest. We went over the random choices to make sure the bot didn't suggest a candidate that objectively had worse parameters than another choice in the same task (we define "objectively" as having a lower score in all parameters that the other candidates).

2. Movies

Task. The general design of the movies task was also based on the hiring game. Each user was now presented with a number of movie cards that included several characteristics related to a movie, and was asked to select the movie that would be most profitable if projected at a local cinema. The characteristics we used include category information such as genre, as well as ranges for critic's rating and cost to acquire. There is also binary information on audience appropriateness, A-list actors/directors, and if the movie is a sequel or part of

a franchise. In the first iteration of this task, we also included audience rating for the task card, but after a walk-through we realized it was too strong of an indicator for the right choice, so we decided to remove it.

Dataset. The movie cards were made in Canva [18], with an example shown in Figure 1b. All the features were manually randomized for this task, along with the bot's suggestions. The task cards were also grouped randomly together. A pass was done to make sure the bot did not make a suggestion that was objectively worse than other choices, similarly to the hiring task.

Since randomization processes were performed manually for this task, we did a walk-through to make sure there wasn't a card group with an obvious optimal selection, and that each group was sufficiently open-ended.

B. System

Overview. We developed a minimalist, interactive point-and-click system to support decision-making and promote reflection on participant choices. The primary view (Fig 2) consists of multiple dynamic components:

(a) a counter to track the participant's progress in completing the tasks,
(b) a **Recommend** button to receive a recommendation from an AI regarding the current task,
(c) a **Save** button to submit the currently selected candidate / movie card and advance to the next round,
(d) the AI's recommendation, which conditionally appears if either I. the **Recommend** button is pressed before the **Save** button is pressed, or II. if the **Save** button is pressed before the **Recommend** button, and
(e) a set of three or four cards which describe the candidates / movies that the participant will select from during each round.

Not shown is condition II., which replaces (b) and (c) with a **Continue** button that moves the participant to the next task. Clicking on a candidate / movie card puts a green border around the card, indicating that the card has been selected. Only one card can have a green border at any given time. The design of the interface is consistent across both the hiring and movies tasks, with only the text (e.g. candidates / movies) changing to match the task. Each task is prefaced by an instruction page that describes the goal and directions to accomplish it.

AI Recommendations. Clicking on the **Recommend** button displays a message (Fig 2d) that conveys the candidate / movies that the AI suggests the participant should pick. This message is intentionally fabricated by the researchers and the suggested candidate / movie is unique for each task. In other words, no actual AI was used to produce the suggestions; the only algorithm that determined the message contents was a deliberation between researchers about what the AI should say. The use of AI brings connotations about performance, reliability, transparency — all key components of trust [2]. Thus, we chose to fabricate the AI responses in order to control for the responses of the AI and more closely study the effects of transitivity and implicit bias between subjects. The recommendations were also simple, they only mention which applicant or movie to pick by ID, with no other information alongside it. This means the bot provided no justification for its recommendation. We recognized that this would make the system less explainable, but we wanted to control for the effect of the reviews (i.e. instead of the trusting the bot because it seems trustworthy in its behavior, trusting it because someone else did).

Reviews. To assess the effects of transitivity and similarity on trust in AI, we produced two versions of the task instructions page: a Control version (CTRL) of the interface, and an intervention version (WTHN, BTWN, BOTH) of the interface, which was modified to include fabricated positive reviews of the AI technology (Section VIII-A). In the WTHN condition, participants were shown eponymous reviews written by other humans, describing their personal experience with the AI. These reviews were written to capture transitivity within human-to-human trust. In the BTWN condition, participants were instead shown anonymous reviews of the AI technology from fictional news outlets in the technology domain. These reviews compared the AI in the study to other, well-known AI technologies such as Siri and Hulu in order to convey the capabilities of the technology.

In the BOTH condition, participants were shown reviews from both the WTHN and BTWN conditions. All the reviews were positive, and had as a target to motivate the user to use and trust the AI.

Frameworks, Libraries and Packages. We developed the tool using the Angular 11 framework [10] for the web interface. We developed a server in Python 3 and leveraged Socket.IO [11] for real-time, bidirectional communication with the web interface to facilitate the collection of interaction data such as clicks.

C. Surveys

We created background, pre-study and post-study questionnaires for the users to fill out. The focus of the pre-survey was the gauge participants' familiarity with AI as a concept as well as on a technical level. We also collected AI-interaction data from them, such as how AI seems to assists them day-to-day, and if they purposefully try to use it. We collected this information to understand possible implicit biases that may play a role in the interaction with the AI. For example, people who may not understand AI on a technical level or who do not work with it may exhibit the ELIZA effect, where more intelligence is attributed to an AI than it is actually doing [25]. In the post-study, our goal was to understand participants task strategies, what they thought of the bot's suggestions and the reviews of the bot systems (if they were in that condition). We included questions about attitudes on AI and trust, some of these questions were adapted from a national survey on AI attitudes [8]. The pre-study and post-study questions can be found in the appendix.

IV. Evaluation

A. Procedure

This study utilized a design which manipulated whether participants were exposed to positive reviews of the AI technology before undertaking the task. Participants in the user study were randomly assigned to one of four conditions: CTRL, WTHN, BTWN, and BOTH. After providing informed consent, participants completed a background questionnaire. Participants were then given the opportunity to practice by choosing to the best of their ability the best grocery store out of a set of 3 options. This practice scenario was not included in the final data analysis. Before completing any tasks, participants took a pre-survey to convey their understanding of AI technologies and any implicit biases they may have with regards to using AI. Participants then completed the first task (either hiring or movies) followed by the second (movies or hiring), with the order counterbalanced between subjects. For each task, participants chose a single candidate / movie from a collection of three or four options across six rounds of selection. Before submitting their choice each round, they could optionally receive a recommendation from the AI technology. Then, participants submitted their decision. If the participant chose not to receive a recommendation before the submission, the interface would show the user what the AI would have recommended to them before allowing the participant to move on to the next round. Those who saw the AI recommendation beforesubmission could change their choice to align with the AI, while those who saw the AI recommendation after submission could only use the recommendation to reflect afterwards, without impacting any decisions during that round. Meanwhile, every participant's interactions were recorded (i.e. card and button clicks). Finally, after completing all tasks, participants took a post-survey to describe their experience with the interface and AI technology present, as well their thoughts and feelings about trust in AI in general.

B. Participants

Due to constraints on time and the limitations of not being able to pay participants, we were able to recruit 26 participants by sending a web link out to family, friends, and peers. We rejected none of the submissions, leaving us with data from 26 participants who were randomly assigned to one of four conditions (7CTRL, 6 WTHN, 7 BTWN, 6 BOTH). By gender, participants identified as female (15), male (10), and prefer not to say (1). By race, participants predominately identified as White (15) and Asian (9), with 1 Black or African American participant and 1 Puerto Rican participant. Participants were between 21 and 52 years old with a median age of 25 (1 preferred not to say). They had a wide range of educational backgrounds (associate's degree (1), bachelor's degree (5), master's degree (9), doctoral degree (11)) and a variety of fields of work, including e.g. international relations, financial services, marketing, pharmaceuticals, software/biomedical/electrical engineering, research physics, AI, robotics and computer science, among others.

C. Hypotheses

Our hypotheses for this experiment are as follows. We organize our hypotheses according to those regarding Transitivity (TN), Similarity (SM), Implicit Bias (IM), Trust (TU) and Usability (US).

- **TN** WTHN and BOTH participants that see the positive reviews before engaging with the task will be more likely to use the AI technology than CTRL participants.
- **SM** BTWN and BOTH participants that see the positive reviews before engaging with the task will be more likely to use the AI technology than CTRL participants.
- **IM** The level of understanding that users have about AI will affect their attitudes towards it.
- **TU1** Participants will be more likely to use the AI technology during the movies task than during the hiring task, due to the difference in the perceived stakes.
- **TU2** BOTH participants will be more likely to use the AI technology than CTRL, WTHN and BTWN participants, as they are presented with more, and more diverse, information.
- **US** WTHN, BTWN, and BOTH participants will find the reviews more useful than not.

Our analyses are both quantitative and qualitative. We used One-way ANOVA for all significance measures when comparing values across groups. Complete analyses are available in supplementary materials.

D. Transitivity and Similarity

With regards to transitivity, we observed the following quantitative interactions with the system and qualitative responses in the surveys.

Interactions. Across both tasks we observed small differences between conditions in the average number of times participants requested a recommendation from the AI before submitting their selection, as shown in Figure 4a. In particular, we found CTRL and BTWN participants used the recommendation feature less than WTHN and BOTH participants. However, we found these differences to be insignificant between both CONTROL (CTRL) and INTERVENTION (WTHN, BTWN, BOTH) ($m_{CONTROL} = 1:36$, $m_{INTERVENTION} = 2:03$; $p = 0:3038$) and all conditions ($m_{CTRL} = 1:36$, $m_{WTHN} = 2:33$, $m_{BTWN} = 1:36$, $m_{BOTH} = 2:50$; $p = 0:3341$). This result provides little support for hypotheses TN and SM.

Responses.

"Is there any other information that the reviews could have mentioned that would have helped you make your decisions? Please elaborate and specify if any specific reviews impacted your decision-making process."

Overall, the reviews about the bot seemed to have little direct effect on the users. Participants generally commented that the reviews were positive but lacking in detail. One user wrote that they ignored the reviews because they were all just positive (I08). Another user (I03) said that the lack of detail even made them seem unreliable, creating distrust. One participant (I16) started with distrust regardless of what the reviews said.

Fig. 3: Number of participants that indicated how often they think AI assists them versus how often they purposefully use AI to assist them.

Other participants mention that trust is more case-by-case, writing "The AI's usefulness is something I rather discover through use" (I10) and "I read the reviews but wanted to make my own opinion of the AI and its reliability" (I03).

One issue we noticed is that this question was misinterpreted. Some users thought it was about the bot's suggestion and not the reviews about the bot ($n = 4$)

E. Implicit Bias

With regards to implicit bias, we observed the following quantitative interactions with the system and qualitative responses in the surveys.

Interactions. We observed that around half of the participants ($n = 14$) reported a working knowledge of the main concepts of AI (YES), whereas the rest of the participants reported having little to no knowledge (NO). In particular, those who reported a working knowledge used the AI technology slightly more than those who did not. However, we found this difference to be insignificant across all tasks and conditions ($m_{YES} = 1.96$, $m_{NO} = 1.71$; $p = 0.6598$). This result provides little support for hypothesis IM.

Responses. In the pre-study questionnaire, participants were asked how often they believed AI technologies assisted them during common tasks, followed by how often they purposefully used AI technologies to assist them with tasks. We found an inverse relationship existed in the frequency of responses between each question, as shown in Figure 3. What this tells us is that most interactions that participants think about that involves AI are not sought out; this means that the AI is just there, even possibly being intrusive, and participants cannot opt out of interacting with it.

In the open response for these questions, most participants explain that they encounter AI through recommendations and ads they see online ($n = 17$). I03 commented on how these tools are not intrusive, saying "I don't usually have to think about them". Some participants mentioned domains such as "Lyft route" or "filtering spam" (I03), "Google translate" (I05), "driving, taking pictures" (I06), "chatbots" (I07), and "customer service" (I08). Some of these same participants, along with those who did not mention specific domains or applications generally write that AI is probably in a lot of places they interact with day-to-day, and that it is more pervasive than they think ($n = 9$). Another mentioned that they "try to avoid 'AI'" in general (I02). Only one participant said that they didn't think AI is involved in their day-to-day online activities, such as "social media or streaming platforms" (I10).

When asked about how often users purposefully use AI, many answers overlapped with the previous question about where AI could be working. Many participants again mentioned recommendation algorithms ($n = 7$). I13 specifically offers which recommendations they would consider more thoughtfully, specifically music and videos over purchases. This introduces a "gradient" of trustworthy domains, which we explore later. Some new domains users mentioned here include troubleshooting problems with a computer (I15), using a home speaker (I22, I03), social media (I19), and investing (I21). Several participants share that they don't actively use an AI's suggestions. I16 points out that they "wouldn't say [AI] necessarily informs my decisions", while I11 writes "I'd rather make my own choices with limited guidance from algorithms." There was also a theme of AI awareness in answers; several users point out that the technologies they use are passive ($n = 4$). I23 wrote that technologies they use don't seem like AI, that "it feels like most of my queries could be done in a non-AI way." I09 offers a counter, saying that they "don't even know if I'm using AI in my life or not."

There were also participants that actively avoided these technologies and interactions ($n = 4$). I21 mentions they avoid data-gathering technologies, writing "i try to give corporations as little data as possible" and I12 writes "I avoid using [AI] when possible as I find it creepy".

I07 was the only participant who mentions ethics for this question, saying "I don't want to get away by blame the AI for a bad-decision".

F. Trust

With regards to trust, we observed the following quantitative interactions with the system and qualitative responses in the surveys.

Interactions. Across all conditions, we observed a small difference in the average number of times participants requested a recommendation from the AI before submitting their selection. In particular, the AI technology was used slightly more during the movies task than during the hiring task. However, we found this difference to be insignificant between the two domains ($m_{hiring} = 1.69$, $m_{movies} = 2.00$; $p = 0.5955$). This result provides no support for hypothesis TU1.

Based on the observed differences by condition in the number of times recommendations were requested across all rounds, as explained in Section IV-D, we further compared by condition the percentage of participants that requested a recommendation each round (Fig 4a) with the percentage of participants that

Fig. 6: Percentage of participants that indicated the level of impact on their trust in AI for different entities.

Fig. 4: Comparing requests for recommendations from the AI each round with whether participant's selections matched the AI's recommendation each round.

Fig. 7: Percentage of participants that chose each option of the stated questions in the pre- and post- study questionnaires.

Fig. 5: Distribution of number of rounds during a single task that participants achieved a particular combination of electing to receive a recommendation or not compared with whether their saved selection matched the AI's recommendation or not

aligned their selection with the AI's recommendation each round (Fig 4b). Henceforth these comparisons across each round will be referred to as *trajectories*. We found that comparing the trajectories of participant requests for recommendations and matches between participant selections and AI recommendations revealed no correlations. Furthermore, we examined whether participants that requested a recommendation ended up matching their answer to it, across all conditions and rounds (Fig 5). No such behavior was observed. These results disconfirm hypothesis TU2.

However, we did find a significant effect in the first round where participants were more likely to utilize the AI technology versus during any other round. We hypothesize that this may be an "exploratory" interaction for the participant to gauge how the AI performs, and may be used to influence their decision-making after that round.

Responses. In the post-study questionnaire, participants reported the level of impact that family, friends, co-workers, professionals, and anonymous persons (e.g. an online review with no name attached) had on their trust in AI technologies. We found that, in general, friends, co-workers and professionals were most likely to have an impact on a participant's trust in AI, as shown in Figure 6. Participants had the least amount of trust in anonymous reviews.

Participants were also asked to select from a variety of choices the *technologies* (Fig 7a), *factors* (Fig 7b), *entities* (Fig 7c) and *tasks* (Fig 7d) that were most relevant to them with regards to their trust in AI. For technologies regularly used, it was interesting to note that while everyone used online banking/email, only one participant mentioned how AI is used in this domain (spam filtering.) Most people talk about recommendations as their idea of AI online For social media, while many people use it (92.31%), no participants mentioned how tools such as facial recognition,

Fig. 8: Summary of usefulness and frequency scores as reported by participants in the post-study questionnaire.

content moderation, and facial filters are also used these platforms. We are not sure if participants didn't realize these tools were AI or maybe don't use them, if they forgot, or if our task primed them since it was recommendation-based. It seems like their mental models of AI is a technology that is more active or seen

From here, it is conflicting to see that "AI Literacy" is rated the lowest on factors that are important for trust in AI, especially since most people missed others ways AI is used. Participants valued factors with results, such as risk mitigation and fairness, along with human factors such transparency in the process and oversight. Only half of participants thought accountability was an important factor.

Participants strongly prefer that non-profit organizations regulate AI, slightly ahead of academic organizations and far ahead of government organizations. This is interesting to note, especially with an increase in AI regulations and legislation across the world [22].

For kinds of tasks participants would trust an AI to assist with, the top three tasks are driving directions, shopping and movie watching while the bottom two are medical decisions and hiring decisions. It is evident that there exists a difference in trust due to the different stakes and impact these tasks have, as well as possibly the amount of specialized knowledge one has to posses to make such decisions. Decisions on tasks like shopping and movie watching have low impact, are easily reversible, and whether they are successful relies on personal preferences and the ability of the AI to learn them. Medical decisions on they other hand are mostly irreversible, have high impact and almost no margin for errors and require specialized domain knowledge and complex reasoning.

G. Usability

With regards to usability, we observed the following qualitative responses in the surveys.

Responses. During the post-study questionnaire, we assessed the usefulness of both the AI recommendations and the reviews of the AI, as well as how often the recommendations of the AI affected the participant's decision-making process. We observed very little difference between conditions and tasks for all of the questions posed. For instance, Figure 8a shows the usefulness scores for the AI recommendations during the hiring task. For the hiring scenario, we find that the average usefulness of recommendations ($m_{CTRL} = 2.71$, $m_{WTHN} = 3.00$, $m_{BTWN} = 2.14$, $m_{BOTH} = 2.17$; $p = 0.4807$), frequency of recommendations affecting decisions ($m_{CTRL} = 2.29$, $m_{WTHN} = 1.67$, $m_{BTWN} = 1.57$, $m_{BOTH} = 1.83$; $p = 0.7152$), and usefulness of reviews ($m_{WTHN} = 1.67$, $m_{BTWN} = 2.00$, $m_{BOTH} = 1.67$; $p = 0.7683$) were not significantly different between conditions. Similarly, for the movies scenario, we find that the average usefulness of recommendations ($m_{CTRL} = 3.14$, $m_{WTHN} = 2.50$, $m_{BTWN} = 2.71$, $m_{BOTH} = 2.33$; $p = 0.6627$), frequency of recommendations affecting decisions ($m_{CTRL} = 3.00$, $m_{WTHN} = 1.67$, $m_{BTWN} = 2.43$, $m_{BOTH} = 2.17$; $p = 0.3235$), and usefulness of reviews ($m_{WTHN} = 2.00$, $m_{BTWN} = 2.43$, $m_{BOTH} = 1.67$; $p = 0.5810$) were also not significantly different between conditions. This result disconfirms hypothesis US

V. Discussion

A. Pre-Survey

We discuss the implications of the pre-survey responses in the context of common questions that encompass the thoughts and feelings of the participants.

"Please describe what artificial intelligence (AI) means to you." There are four kinds of answers here that participants gave on understanding what an artificial intelligence is. The first kind is data and process-centric, where users talked about inputs and outputs, mappings, and automatic decision making ($n = 8$). I04 writes that AI "make[s] a complex mapping between input and output to give a desired result." The second kind is concerned about intelligence, either to replicate it or to behave in an intelligent way, both without an explicit mention of humans ($n = 4$). Two users write that AI tries to "mimic" (I24) or "emulate" (I17) intelligence. The third kind is human-centric, meaning users wrote that AI is built to do tasks people do, to support them, respond to them, etc. ($n = 11$). Participants made comments about AI, such that it is a "system... [that] assist[s] humans with some task", it "make[s] decisions resembling those of a rational person" (I03), or "that can respond to human stimuli in different, reasoned ways" (I10). The general idea is that AI helps people with day-to-day tasks, matches their performance on tasks, or tries to surpass them. The last kind of answer is related to advanced-programming; this means they talked about AI as sophisticated computer programs ($n = 5$). Users described AI as "smart if-elifs" (I07), "machines that utilize more than simple heuristics" (I16), and that "[to] determine the next outcome... what makes it Intelligent is there is no set selection of parameters" (I23). One person (I19) used imagery to answer this question, writing " I honestly picture robotic arms or the lasers that perform surgery, but I know that it can be much smaller and simpler than that, I just have no idea what." Another user, who did not complete the full study, mentioned an example from media, writing that AI for

them meant "For humors sake, the Terminator films." One person (I05) mentioned whole domains, "Robotics, Medicine, Neuroscience, Language and Vision." while another (I06) used low-level concepts such as "translation, object detection". No other users gave examples of AI in explaining what it meant to them, so it was interesting to see these users use case-based reasoning for their answer.

These impressions of AI offer a view into possible biases that users may have during the collaborative task.

B. Post-Survey

We discuss the implications of the post-survey responses in the context of common questions that encompass the thoughts and feelings of the participants.

1. Strategies

"Please describe your strategy for selecting applicants in the hiring scenario" Most people wrote that they relied on the skills feature to select applicants ($n = 16$), followed by work experience and then confidence. It was common for people to use a pair, or mix of categories to go by during decision making. A few people ($n = 3$) emphasized that they wanted a "well-rounded" or "balanced" candidate.

Every participant except for one mentioned at least one feature they used to complete the hiring task. Several participants ($n = 8$) explicitly said they relied on school prestige the least, with only person saying they considered it while the rest didn't mention it at all in their responses.

Most users did not mention the referral feature in their explanations. For the ones who did ($n = 8$), referral was an important feature in their decision making. One user (I19) said "If a person had a referral I automatically assumed they are at the top of the list despite any other attributes," similarly I25 wrote "Also if they were referred I viewed that as more of a positive." This assumption means they put value in what a referral meant- another human suggested the person. Another user (I24) qualified this reasoning, writing "referrals only mattered if the person had skill." A different user (I18) dismissed referrals for this reason, saying "Referral ... didn't play a significant role in my final decision due to the fact that they are parameters that define the candidate's worth, indirectly."

In relation to the AI, most participants did not mention it. The two who did mention it wrote that they did not find its suggestions useful. One participant (I03) wrote "I almost never changed my answer to the AI recommendation because it seemed to me that the AI over valued school prestige", while another (I16) said "I felt [the suggestions] were less useful."

"Please describe your strategy for selecting movies in the movies scenario" Most participants cited the cost-to-critics'-rating ratio as their main or only strategy when selecting a movie ($n = 17$). Other factors specifically mentioned where the amount of boxes each movie had checked, as well as whether it featured A-list actors in particular. Some participants ($n = 3$) wrote that they made their choices based on personal preference in terms of movie genre. One participant (I18) tried to maximize the target audience of a film, while one participant (I07) mentioned not having a specific strategy.

A number ($n = 4$) of participants explicitly mentioned the AI while explaining their strategies. I14 utilized the AI's recommendation if no other factor seemed to satisfy their criteria, stating "I trust the algorithm more than pure chance in such a case". Interestingly, participant I03 specifically mentioned that they preferred critics' score over an AI recommendation as "the critics' rating is more valuable than the AI recommendation because it represented the feedback of human experts". This is inline with the general trend of mistrust towards AI identified in this study. Most importantly, it showcases how humans value human cognition more than machine cognition, as they think their opinion on a matter (here, a movie) will align better with that of a fellow human than with an AI's. The same participant mentioned that they were satisfied when the AI seemed to mostly agree with their choices, and completely ignored its recommendations when they disagreed.

Two participants specifically mentioned that the AI played a role in their strategy/affected their thinking process. I10 wrote that they started taking the "Part of Franchise" attribute into consideration after the AI seemed to show preference with movies that fulfilled this attribute. I12 described a different approach: while initially taking only cost and critics' ratings into consideration and planning to ignore the AI, they then opted to to scrutinize the recommendations and try to reason about them. This made them "mentally engage with the task to a greater degree and ultimately remember that there may be other features I should try to balance out". This demonstrates the high value of having explainable and interpretable AI not only for establishing trust, but also as a tool to further enhance a human's cognitive process by illuminating factors that could play an important role in decisions and humans could overlook.

"Did the AI ever recommend something unexpected? Please elaborate and specify which tasks the recommendation(s) occurred in." Participants made three kinds of statements about the AI's suggestions. These are surprise, meaning that they found the AI's choices to be random or without pattern, explainable, as in they thought it had preferences for certain features or a particular strategy, or agreement, meaning that the bot made suggestions the user agreed with or considered as a good option.

For the hiring task, several participants felt that the bot favored applicants with school prestige ($n = 5$). Four participants thought the bot made choices they agreed with.

One participant (I16) commented that even though it seemed like the bot was random, "it made me think". Another participant (I07) was frustrated with the bot, expecting it to be more sophisticated- they wrote it should have "asked for my preferences up-front or at least learnt on-the-fly".

Overall, users wanted to understand the bot's decision making strategy during the task; without this, and with too-positive reviews, trust did not seem to form properly.

2. General AI Attitudes

We discuss general attitudes of participants towards AI and AI technologies.

Fig. 9: Reviews shown during the hiring task for the BOTH condition.

"Do you think there are domains where AI shouldn't be trusted? Please elaborate./ Do you trust AI to make decisions without human supervision?" In general, participants agreed that AI should be supervised in domains with some stakes involved or substantial risk ($n = 5$); essentially, the AI should be in an assisting role. Several participants selected medicine ($n = 8$), finances ($n = 5$), and hiring ($n = 5$) as domains where AI should not be trusted. Some other domains that they mentioned include ones dealing with "human emotions" (I25) or "mental health" (I02), campaign ads/politics (I18, I14), and education (I04). One participant (I24) considered that if an AI was left to decide, that it "should always be checked in some regular basis for quality control."

"Overall, with regards to technology, AI, and the future, are you: A pessimist, An optimist, or It's complicated"

For participants who selected "It's complicated" ($n = 5$), the main points was that AI development needs to happen responsibly.

From these questions, it's clear that domains establish a prior in trust. In which case, different kinds of information in a review about the system in a particular domain may be necessary.

VI. Limitations and Future Work

A. Methodology and Evaluation

Due to limitations during recruitment, we were not able to properly assess the effects between study conditions. These limitations included concentrated age distribution around 25 years of age, concentrated racial background around White and Asian participants, concentrated education background around doctoral and master's degrees and a generally small number of participants for each condition. A proper crowd-sourced study with accompanying power analysis would provide a more diverse and representative sample of participants from the population as a whole. In particular, we are interested to how people with no technical understanding of AI and limited familiarity with technology in general approach collaboration with one.

Our goal in this experiment was to understand how opinions of an AI system, or reviews, shape how people interact with it. We provided simple reviews at the beginning of each task, but several participants found it unhelpful by being only positive and lacking in detail. If we were to do this experiment again, we would make the reviews sound more detailed and realistic so participants would be less skeptical of them. Keren found that negative views generated more trust in people [12]. It also seemed like participants reviewed it too quickly or ignored it, so it could be emphasized more in a different study. There exists the fundamental challenge of trying to ensure participants take into consideration utilizing the AI without directly or implicitly enforcing it. One implementation idea was to present participants with negative feedback on their choices in the middle of each task, so as to encourage them to question their strategies and see if it affects their interaction with the AI. This idea was not used as we deemed it too invasive, and wanted to evaluate how participants acted without any external simuli. An interesting way this balance between encouraging but not enforcing could be achieved is by introducing time and load of work as parameters in the task. AI is often being deployed in decision-making when the pool of decisions is vast in order to accelerate the process. By introducing participants to a task that requires them going over much more possible options (hiring portofolios / movie cards) per round, and/or have to make quick decisions under some time limit, we could ensure that participants are more likely to utilize the AI, without giving away any information about whether their existing strategy is good or not.

The participants also thought the hiring bot skewed towards school prestige, which made some of them ignore the bot. The bot's choices were random and earlier tasks happened to emphasize school prestige. A simple user-test before rolling out the study could have caught this so the participants didn't think the bot was focused on a single variable. In Figure 6, most participants indicated that they would not trust an AI to assist them with hiring; we wonder if this is because the bot seemed skewed or if the domain did seem more impactful/serious.

We also wonder if the participants misunderstood the hiring task as a final choice; we wrote that selected candidates would go on to more interviews, but in the post-survey many participants thought AI should not be involved in a hiring domain. The difference between viewing it as a final decision vs. as a step in the process means viewing the AI as having a greater effect vs offering helpful options.

One experiment design we considered was to precisely determine if the bot changed the user's mind during a task. The difference would be that the participant would select an initial answer, then view the AI's suggestion and have the opportunity to change their selection before before a final submission. We also noticed an initial exploratory behavior between the user and AI. From this, another potential experiment is understanding how users develop trust in initial interactions between AI and other humans (i.e., does trust form or develop differently between these two kinds of agents?).

B. Surveys

One problem we noticed in the survey was that questions about the bot reviews were misunderstood. In the post survey, there is an open-ended question (q8) that asked about what content the initial reviews could have had to make them more useful for the users. Several of the participants wrote that the bot could have had a more explainable suggestion. One user wrote that "the AI could at least tell _why_ it suggested something" (I07). Another user disclosed that they "never looked at the AI before making a decision" (I04).

On a design side, we would re-format the questions about user strategy (q3,q4) in the post-survey to include an automatic ranking system for the different parameters of the task. For the hiring task, all users except for one explained their strategy by mentioning at least two features in their explanation. Re-designing the question to ask for ranking would help clarify preferences better instead of taking a best guess based on the written answer.

We would also add a question about the different ways AI manifests in a particular domain; this would help us understand what users can identify in a given system (and what they are biased to identify).

VII. Conclusion

In this paper we explored the impact that transitivity, similarity, and implicit bias has on whether people trust an AI technology to help them make decisions. To study this effect, we presented study participants with two task scenarios consisting of multiple rounds of decision-making alongside an AI technology that could optionally be asked for recommendations. We find that transitivity plays a very minor role when it comes to trust, and in some cases is detrimental to trust when the source of that secondary information is not trustworthy itself. Further, we found that implicit bias has a very strong impact on pre- decision-making trust in AI, leading to a "judge a book by its cover" effect that precludes development of trust regardless of an AI's implementation. Overall, we find that the trajectory, or timeline, of when users decide to trust and/or engage with AI is dependent on how much trust a user has in AI already.

Acknowledgements

The authors wish to thank their friends, families and peers for volunteering for the study.

References

[1] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, et al. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–13, 2019.

[2] S. Banerjee, S. Bhattacharyya, and I. Bose. Whose online reviews to trust? understanding reviewer trustworthiness and its impact on business. *Decision Support Systems*, 96:17–26, 2017. doi: 10.1016/j.dss.2017.01.006

[3] S. Baron-Cohen. Evolution of a theory of mind? *The Descent of Mind: Psychological Perspectives on Hominid Evolution*, 01 2012. doi: 10.1093/acprof:oso/9780192632593.003.0013

[4] O. Biran and K. R. McKeown. Human-centric justification of machine learning predictions. In *IJCAI*, vol. 2017, pp. 1461–1467, 2017.

[5] C. Castelfranchi and R. Falcone. Social trust: cognitive anatomy, social importance, quantification and dynamics. August 1998.

[6] G. Csapo, J. Kim, M. Klasinc, and A. ElKattan. Survival of the best fit, 2019.

[7] R. M. Fernandez and N. Weinberg. Sifting and sorting: Personal contacts and hiring in a retail bank. pp. 883–902. American Sociological Review, Dec 1997.

[8] N. Gillespie, S. Lockey, and C. Curtis. Trust in artificial intelligence: Australian insights, 2020.

[9] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89. IEEE, 2018.

[10] Angular, n.d.

[11] Socket.io, n.d.

[12] G. Keren. Framing, intentions, and trust–choice incompatibility. *Organizational Behavior and Human Decision Processes*, 103(2):238–255, 2007. doi: 10.1016/j.obhdp.2007.02.002

[13] M. Kleiman-Weiner, Y. Zhao, and J. Tenenbaum. Cooperative social intelligence: Understanding and acting with other. In *CogSci*, 2017.

[14] S. Kusumasondjaja, T. Shanka, and C. Marchegiani. Credibility of online reviews and initial trust the roles of reviewer's identity and review valence. *Journal of Vacation Marketing*, 18:185–195, 07 2012. doi: 10.1177/1356766712449365

[15] D. Long and B. Magerko. What is ai literacy? competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2020.

[16] E. Maslowska, E. C. Malthouse, and S. F. Bernritter. Too good to be true: the role of online reviews' features in probability to buy. *International Journal of Advertising*, 36(1):142–163, 2017. doi: 10.1080/02650487.2016.1195622

[17] D. W. McCloskey. An examination of the data quality of online reviews: Who do consumers trust? IGI Global, Jan 1970.

[18] Canva, n.d.

[19] P. Racherla, M. Mandviwalla, and D. J. Connolly. Factors affecting consumers' trust in online product reviews. *Journal of Consumer Behaviour*, 11(2):94–104, 2012. doi: 10.1002/cb.385

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

[21] O. Richters and T. P. Peixoto. Trust transitivity in social networks. *PloS one*, 6(4):e18384, 2011.

[22] A. Satariano. Europe proposes strict rules for artificial intelligence, Apr 2021.

[23] C. S. Smith. Dealing with bias in artificial intelligence, 2019.

[24] B. A. Sparks and V. Browning. The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management*, 32(6):1310–1323, 2011. doi: 10.1016/j.tourman.2010.12.011

[25] N. Wardrip-Fruin. Three play effects: Eliza, tale-spin, and simcity. pp. 1–2. Digital Humanities, 2007.

[26] H. J. Wilson and P. R. Daugherty. Collaborative intelligence: Humans and ai are joining forces. Harvard Business Review, Nov 2019.

VIII. Appendix

A. Reviews

1) Hiring:

WTHN **Mary M**, Hiring Manager, 10/15/2020: "I couldn't imagine doing my job without it."

WTHN **Kaley R**, Software Developer, 09/23/2020: "I'm impressed by how well it can predict great candidates. Very powerful!"

WTHN **Sofie A**, Hiring Manager, 01/13/2021: "It was really helpful when I was stuck between certain choices ... Definitely saved me some time."

BTWN **'AI Monthly' Magazine**, Article, 02/05/2020: "This AI was built using similar technologies behind Siri and Watson."

BTWN **ai.powered.io**, Blog Post, 03/14/2021: "This AI is ... a complete upgrade to similar tools on the market."

BTWN **businessnewsonthedaily.com**, Article, 06/17/2020: "This is a great cognitive tool like the Watson Jeapordy player AI ... it understands what people look for."

2) Movies:

WTHN **Nick V**, Movie Theater Owner, 05/10/2020: "Saved my business a lot of unnecessary purchases ... can't imagine operating without it."

WTHN **Helen T**, Investor, 11/12/2020: "I didn't expect it to be so effective!"

WTHN **Jim D**, Movie Theater Operator, 03/30/2020: "It helped me pick some hidden movie gems I would have otherwise missed ... definitely improved our final catalog."

BTWN **aigeekweek.com**, Blog Post, 03/08/2020: "An excellent tool ... like a spiritual successor to Netflix's recommendation technology."

BTWN **'Tech Now' E-Zine**, Newsletter, 08/31/2020: "This 'assistant', like Google Maps, really shows the great things AI tools are capable of."

BTWN **bizreview.org**, Product Review, 01/23/2020: "Sometimes it feels like this AI knows movies better than people do ... watch out, Hulu!"

B. Questionnaires

1) Background:

q2 With which gender identity do you most identify?

q3 What is your age?

q4 Are you of Hispanic, Latino, or of Spanish origin?

q5 How would you describe yourself?

q6 What is the highest level of education you have either completed or are currently pursuing?

q7 What is your major field of study (if you are a student) or field of occupation?

2) Pre-Study:

q2 Do you know what an artificial intelligence (AI) is?

q3 Please describe what artificial intelligence (AI) means to you.

q4 Do you understand the main concepts of AI such as machine learning, neural networks, deep learning, etc?

q5 What kinds of technology have you and/or do you use regularly? Please select all that apply.

q6-7 How often do you think AI assists you with tasks, e.g., making purchases/decisions, enabling/troubleshooting technology, etc.? Please elaborate on your answer above.

q8-9 How often do you purposefully use AI to assist you with tasks, e.g., making purchases/decisions, enabling/troubleshooting technology, etc.? Please elaborate on your answer above.

q10 Does your work include any component of interacting with AI, e.g., developing/researching AI, using AI to accomplish tasks, writing/talking about AI, etc.?

3) Post-Study:

q2 Please describe your strategy for selecting applicants in the hiring scenario.

q3 Please describe your strategy for selecting movies in the movies scenario.

q4 How useful did you find the AI recommendations to be in each task?

q5 How often did the AI recommendations affect your decision making in each task?

q6 Did the AI ever recommend something unexpected? Please elaborate and specify which tasks the recommendation(s) occurred in.

q7 How useful did you find the reviews of the AI to be in each task?

q8 Is there any other information that the reviews could have mentioned that would have helped you make your decisions? Please elaborate and specify if any specific reviews impacted your decision-making process.

q9 Which of the following decision-making processes would you trust AI technologies to assist you in? Please select all that apply.

q10 Do you think there are domains where AI shouldn't be trusted? Please elaborate.

q11 Do you trust AI to make decisions without human supervision?

q13 What factors are most important for you to trust an AI technology? Please select all that apply.

q14 Which of the following entities do you trust to regulate AI technologies in the best interest of the public? Please select all the apply.

q15 How impactful have the following entities been on your trust in AI technologies?

q16 Overall, with regards to technology, AI, and the future, are you: