The Human – AI Trust Relationship: Error Accuracy in Artificial Intelligence Advice and the Effect on the Human Trust Trajectory

Lucas R. Siecker

Department of Industrial Engineering and Innovation Sciences

0BEPP0: Final project bachelor Psychology & Technology

Supervisors: Gerrit Rooks & Patricia K. Kahr

June 26, 2022

Abstract

In the last couple of years, Artificial Intelligence (AI) is used more extensively. A highly capable AI can be better than humans for difficult tasks, but trusting advice from a computer seems difficult. Benevolence, competence, and integrity are key factors that determine trust in an AI system. Trust over time is a trajectory in which you can gain trust at a slow pace and lose trust very quickly. In this report, the effect of error accuracy on the trust in the AI advisory system and the long-term effects are investigated. Participants showed more trust in a system that was more reliable than in a less reliable system. Over time a low-accurate system shows a decrease in trust and a highly accurate system a small increase in trust. This trust is measured by a self-assessment of the participants but also calculated based on the Weight of Advice. The result of this study can show how Artificial Intelligence advice can be made more trustful and how it can be used more in difficult tasks.

Keywords: trust, algorithms, artificial intelligence, accuracy, reliability, legal advice, recommendation

Introduction

Artificial Intelligence (AI) is becoming a more critical technology every day. It is well-known for determining the content you will consume on the internet. Big technology companies like Meta, use AI and consumer data to determine which advertisements or product suggestions they must show (Bell et al., 2021). It makes the life of a lot of people easier. Information is always at hand with Artificial Intelligence agents on our smartphones, like Siri or Google assistant. But do humans always trust these agents? What determines when to trust the information given by an AI system?

AI is progressively advancing to the point that it outperforms human intelligence in certain areas. It started with outperforming humans in the most difficult games that exist, such as the game of Go. Due to numerous tricky situations on the board and a challenging number of moves and positions, it was difficult for an AI system to play. Eventually, also this game was mastered by an AI and it could beat humans (Silver et al., 2016). Since AI technology can become more intelligent than humans, it should be possible to assist humans that work in, for example, high-risk environments. There is already substantial progress in diagnostic medical imaging, in which the diagnostics of doctors can be assisted or reviewed by a computer system (Nakata, 2019). But why is AI not used in more applications, when it can have a great benefit?

For Artificial Intelligence systems to be used in all different kinds of applications, they should be reliable. If an AI gives recommendations, the users of this system should trust the system. If this is not the case, the recommendation is not heard and the system will not be used, while it is possibly even better than humans. This raises the question of what the effects of accuracy in an Artificial Intelligence recommendation are.

Artificial Intelligence systems can be very versatile. That is because this field of computer science is very broad, and the capabilities of computers are growing. This raises the question of what new development in AI comes next. Gathering and interpreting information using patterns can become extremely useful (Ferràs-Hernández, 2018). The results and outcomes of such a system can be more precise than humans are capable of, as shown earlier.

The use of an AI system in a high-risk environment was attempted at the end of the last century. An Artificial Intelligence system works on a set of predefined rules and thus it was thought that jurisdiction could be a beneficial use case. The system that was used, was based on rules gathered from legal experts and judges. However, this system showed that jurisdiction is more than just applying the rules (Noordegraaf et al., 2019). After years of development in

the whole field of Artificial Intelligence, there are now more possibilities for AI in the legal system. AI can function in three separate ways: Analysis of documentation, fast and efficient searching of documents, and assisting and deciding. The first two use cases involve text processing and are already used since 2019 by the police when a police report is filed. Thus, such a use case can be extended with existing technology and is indirectly used in the current legal system. A third and more interesting function of AI in the jurisdiction is the function of making decisions or assisting judges in making verdicts. To achieve this, firstly everything should be made digitally, and secondly, the field of jurisdiction should be open to the use of such an AI system (Noordegraaf et al., 2019).

For the use of AI in our daily lives, and this case more specifically jurisdiction, humans need to trust the technology. Trust is an essential feeling you have with a person, or in this case an embedded AI. With increased responsibility for AI, the trust humans have in AI will be more critical than before. Especially when life-changing decisions can become part of the AI advisors' portfolio. This vulnerability of humans while using an AI is part of trust. This human-AI trust relationship is emphasized more since AI behavior is complex and nondeterministic (Glikson & Woolley, 2020).

In this research, the effects of accuracy in Artificial Intelligence advice are being investigated in jurisdiction and how trust in an AI system develops. This leads to the following research question: How does the accuracy of Artificial Intelligence decision-making advice influence the human-AI trust trajectory over time? Results can lead to conclusions about accuracy in AI advice, taking into consideration the performance reliability and the machine capability of the decision-making AI advisor.

Theory

To gain the trust of a human, Mayer et al. describe three factors that are involved in perceiving trustworthiness between humans. The first one is the ability of a trustee, someone or something that is being trusted, in a specific domain. Only in this domain, the trustee can be trusted since the competence of the trustee is in just one technical area (Mayer et al., 1995). The second factor for trusting is intention. This is the extent to which is believed the trustee wants to do good and not be egocentric (Cook & Wall, 1980; Mayer et al., 1995). The last factor of the three is integrity and this means that a trustee's set of principles must be found acceptable by the trustor. If this is not the case the trustworthiness is lowered. However, the

three factors together define the perceived trust of a trustee, not just one of these factors (Mayer et al., 1995; Moorman et al., 2018).

Besides the factors of Mayer et al., it is considered that there are more key features. Multiple features of trust are described in several other research papers. They vary between features that indicate a level of performance, reliability, or intentions of a system or the process of interacting (Butler & Cantrell, 1984; Lee & See, 2004; Li et al., 2008).

For technology to gain trust, these do play a role. However, this depends on the implementation of the technology and how it appears to the user. A system that is not made to interact like a human, will gain trust more difficultly. In systems that only provide recommendations or calculations, the accuracy will play a significantly bigger role in gaining trust (Li et al., 2008). Research on trust between humans and technology is often in the form of recommendation systems. Empirical research showed that there are three significant factors (benevolence, competence, and integrity) for trusting, and therefore technology acceptance (Benbasat & Wang, 2005).

Embedded AI is a system that is not visually represented or has no virtual identity. Such a system has per definition no tangibility and users of such an AI may not even be aware they are using an assistant. This can limit the human-AI-trust relationship. However, showing the user that there is an AI involved, can raise ethical questions, and it can have an impact on the trust users have in the system they are using (Glikson & Woolley, 2020). Transparency could lead to an increase in trust, even after losing trust by revealing the AI. If there is a good explanation of how an AI system works, the trust in this system is significantly higher and as is the reliance on the system. On the other hand, it could have a positive effect to limit the level of transparency to a certain level. Research showed that additional information leads to a decrease in trust if the expectations of the system did not match (Kizilcec, 2016). Errors or uncertainties expressed by the AI could lead to distrust. Thus the reliability of a system determines how much a human will rely on the advice. If this reliability is low, trust is significantly decreased, and regaining this trust is a difficult and time-consuming task.

Task characteristics and immediacy behaviors also determine trust. It is shown that in tasks involving social skills, the trust in humans is higher than in an AI system (Glikson & Woolley, 2020). However, the user's self-confidence also plays a key role in trusting such a system. If they perceive that they can perform better, the trust is decreased, and the user will rely less on the AI system (Lee & Moray, 1994). A system should match a user's personality

and the level of interaction. If a system is highly intelligent, it can be beneficial for the trust level, to make the personality of the system appear highly intelligent (Nass & Moon, 2000).

During an interaction with an AI, the level of trust can be seen as a trajectory. The reliability at the beginning of this trajectory determines the overall level of trust and therefore the trajectory. If there are errors at the beginning of the interaction the trust level trajectory stays significantly lower than when no error is made at the beginning. The trajectory was less influenced by errors that occurred later during the interaction. When the trust was lost initially, research showed that this trust level could be recovered. However, this takes time and is a challenging task (Tolmeijer et al., 2021). Because of the big influence of an error at the beginning of an interaction, the order of errors is in this research randomized throughout the experiment. In this way, the trust trajectory can be investigated over time, which depends on the error's accuracy and not on the error's place.

During an interaction with a machine, trust is developed continuously. Between humans, trust increases over time due to interactions. However, humans tend to blame technology for mistakes made, and therefore, the general trust in technology during an interaction is decreasing (Madhavan & Wiegmann, 2007). Other research showed that users have an initial trust level, which is adjusted during an interaction, depending on the system's performance. In general, trust acquisition is becoming slower and slower over time. Trust loss is accelerated if there are multiple consecutive errors. There is some evidence for general decreasing or increasing trends, but this depends on the reliability of a system (Yu et al., 2017).

Something Yu et al. did not include in their research is the accuracy of a single error. In their research, binary values were presented during several trials, and only the accuracy of this full set of trials was determined. The accuracy of individual errors and their effect on the trust trajectory can be more interesting. An error can have different accuracies if they are not binary values. In jurisdiction, this can be done by varying the jail time sentences. One error might be more severe than another and this effect of accuracy on trust should be investigated more. The findings of Yu et al. resulted in the first hypothesis for this study:

Hypothesis 1: When a recommendation system is more accurate, the trust in this system will be higher.

An embedded AI system in this research can be in a range from high to low reliability. This can affect the trust trajectory of the human-AI trust relationship. High reliability does not per definition mean that there are no errors. However, the effects of accuracy are not yet

researched in the context of the human-AI trust relationship. Trust development takes time and therefore accuracy could also affect the trust trajectories.

Some research shows that for AI-assisted decision making, the effect of showing information about the accuracy of a system can influence the trust participants have in the system and the recommendation. Participants are more willing to follow the recommendation if this information is displayed. This only includes information about the accuracy, and not the experienced accuracy of the system (Zhang et al., 2020). However, Zhang et al, showed that displaying accuracy information does depend on the accuracy of the model. Only high confidence levels were enhancing the trust significantly.

The accuracy that is stated by the system in a recommendation does have a significant effect on the trust people have, regardless of its observed accuracy (Yin et al., 2019). Due to the findings of Zang et al., and Yin et al., the information about the accuracy, confidence level, or any other information about the system's performance is not displayed to the participants in our research.

When participants learn about the correct outcome, their trust in the system is adjusted. Therefore, it is important to measure the trust before and after learning about the correct outcome. Using both measurements it can be determined how the trust is changed by experiencing the accuracy of the system. Yin et al. also researched the effect on trust after the resolution in combination with the observed accuracy. It is shown that participants increased their trust if the observed accuracy was high and lowered their trust in the system if the observed accuracy was low. Together with the information from Tolmeijer et al., the second hypothesis in this research is the following:

Hypothesis 2: While using a highly accurate recommendation system, the trust is increased over time.

There is previous research done in decision-making with a judge advisory system. In these researches, the extent to which a recommendation is being used is measured by calculating the Weight of Advice (WOA). Participants can be asked to provide an initial estimation, followed by showing a recommendation and then asking for a final estimation. The difference between the initial and final estimation is the basis for the WOA. If the WOA is below 0, the advice is not followed at all and the final estimation is even further away from a correct answer than the initial estimation. A WOA above 2 is the result of a final estimation,

where the advice is followed, but this estimation is equal to or even worse than the initial estimation (Gino & Moore, 2006).

Another way to provide insight into how advice is taken into account is the Belief Adjustment Model (Sniezek & Buckley, 1995). This model takes into account the prior beliefs of a human. To what extent the advice will be taken into account depends on the sensitivity of the participant to new information. Additionally, the intentions and process of interaction influence decision-making. Previous research in an advisory system for judges determined how advice is taken into account by the WOA. Therefore, in this study, the Weight of Advice is used to determine the trust participants have in the system (Hogarth & Einhorn, 1992).

Method

The hypotheses were tested with a between-subject experiment. In this experiment, participants were presented with 20 law cases. In each individual case, they had to determine the prison sentence the suspect should get. After their initial estimation, an AI decision-making assistant will give a recommendation (advice) to the participant. Based on this advice, participants can change their initial prison sentence estimation, but this is not obligatory. After their second estimation, the case results are shown and participants can learn about the correct verdict that was given by a judge. Trust levels were compared between participants to determine what effect accuracy has on the trust level. Additionally, to explore the trust trajectory, the trust levels within a participant were investigated.

Participants

For this research, the required sample size for the experiment was 176 participants. This was based on the comparison between two independent means. The desired power was 95%, where the p-value is 0.05 and the research contains two conditions, high and low accuracy. The expected effect size was 0.5 because 50% and 90% accuracy differ a lot and effects were shown in research where smaller differences were used (Yu et al., 2017). 176 participants were recruited correctly via Prolific. 125 were female, and 51 were male. They had an average age of 37.0 years (SD = 12.7). None of the participants declined to indicate their gender. To increase the reliability of the answers the participants gave, they were selected based on the criterium that they were familiar with the current jurisdiction. Due to this criterium, participation was limited. During the research, this criterium was released to recruit more participants. Additional criteria were that all participants had at least the age of 18 and were citizens of the United Kingdom.

Design

The experiment has a 2x2 between-subject design. The dependent variables were the indicated trust in the system by the participant and the effect of the advice on the participant (WOA). Accuracy and perceived intelligence of the recommendation system were the two independent variables. Each participant was assigned randomly to one of the four conditions. Perceived intelligence is an addition, due to collaboration with other researchers while conducting the same experiment.

Accuracy was calculated using the correct sentence that was given by a judge. Within an accuracy condition, the accuracy was calculated using a normal distribution. The normal distribution was either within 10% of the correct jail sentence or 50%. In this way, an average accuracy of 90% and 50% respectively was achieved. Using a normal distribution, some recommendations can be experienced as outliers, which could give a more natural feeling while interacting with the system. The recommendation was calculated using the correct jail sentence, normal distribution, and a random number between 0 and 1 to determine the probability within the normal distribution.

The second factor is the perceived intelligence of the recommendation system. Each participant interacted with a system that was either displayed as highly intelligent or low intelligent. This was done by the wording and explanation of how the AI works and how it was built and trained. Also, during the cases, the font display and presentation of the analysis were changed according to this condition. For example, the high intelligence AI advice explanation was given as followed: "I **analyzed the case** and found (and categorized) the following (grouped) **keywords**: *stabbed*, "f*atal consequences" (potential)*, "*attempted manslaughter"."*. In comparison, the low intelligence explanation looked like this: "4 keyword matches: *stabbed*, *attempted, manslaughter, fatal".*

All participants received the law cases in random order. This was to prevent the effect of the first case and other potential influences on trust because of the order (Tolmeijer et al., 2021).

Materials and setting

The data for the 20 law cases that were used came from the "de Rechtspraak" database, which can be accessed via a Dutch website (Rechtspraak.Nl - Zoeken in Uitspraken). All 20 law cases were picked by hand from the database, to ensure multiple difficulty levels and no extremely shocking cases were included. An additional case was picked, to be used as a training

case, which was shown during the introduction phase of the experiment. This case was not added to the 20 law cases which were used for the experiment.

The participants did the experiment from home. Therefore the setting in which the experiment was done, could not be regulated by the researchers. The appearance of the system was done in LimeSurvey and the distribution via Prolific. All texts were translated and rewritten to make the cases and recommendations accessible for everyone.

Procedure

All participants were registered at the Prolific platform and taken to the experiment website after accepting the research. They were first presented with a consent form and asked to give consent. Following, participants were assigned randomly to one of the four conditions. To begin the experiment participants got a general introduction to the experiment and what was expected from them. After this general introduction, an introduction to the Artificial Intelligence system was given. At this step, the first difference was made between the groups. Two groups received an introduction in which the perceived system intelligence was high, and the two other groups received a low intelligence system introduction.

Followed by an introduction, a training case was presented which was also altered for the perceived intelligence. Since the sentence in this specific case was just one month, there was no difference in accuracy. Then the 20 trials were presented to the participant. Each case started with a general explanation of the case, on which the participant could estimate what the sentence should be. This estimation was filled in and after a loading screen, the AI recommendation was displayed. The loading screen was visible for a variable time between 2 and 5 seconds. The AI recommendation had either a high (90%) or a low (50%) accuracy. In addition, an explanation was added to the advice. Based on this advice, participants were asked to change their estimation.

Following, a resolution page was shown, including the correct jail sentence, their second estimation, and the AI advice. Finally, the participants were asked for their trust in the system. After 20 law cases, a post-experimental questionnaire was completed by the participants. This questionnaire included demographics, level of experience with jurisdiction, and personality traits. After the whole experiment, which was about 30 minutes, participants received a $7 payment.

Results

During the research, 197 participants started with the survey. 20 participants are excluded because they did not reach the end of the experiment and 1 participant is under the minimum age of 18, and clearly unreliable. After a closer investigation of the data, it can be seen that one participant followed the advice in all cases and one participant is not influenced at all by the recommendation and always filled in her first estimate. This estimate is in a lot of cases an extensive amount of months, which is no reliable data. This results in 176 participants who did complete the experiment correctly and reliably. 123 participants are female and 51 are male. They were randomly assigned to a condition (see Table 1). Participants in conditions 1 and 3 received low-accurate recommendations, while conditions 2 and 4 received high-accurate recommendations. The descriptive statistics about the age and jurisdictional knowledge of the participants can be found in Table 2.

Table 1
Descriptive statistics of the random assignment to conditions by gender

Condition	Female	Male	Total
1 (High intelligence, Low accuracy)	31	11	42
2 (High intelligence, High accuracy)	41	17	58
3 (Low intelligence, Low accuracy)	29	12	41
4 (Low intelligence, High accuracy)	24	11	35
Total	125	51	176

The distribution of the random assignment to a condition. It shows that the distribution varies and that more participants have interacted with a highly accurate system.

Table 2
Descriptive statistics about the age and jurisdictional knowledge

Variable	Accuracy	N	Mean	Standard Deviation	Minimum value	Maximum value
Age	Total	176	37.05	12.75	19	72
	High	93	36.76	13.41	19	72
	Low	83	37.50	11.86	21	68
Affinity with Jurisdiction	Total	176	4.21	2.51	1	10
	High	93	4.24	2.59	1	10
	Low	83	4.18	2.40	1	9

Here the differences in Age and the affinity with jurisdiction are shown. The distribution over the two conditions is quite even., for all the minimum, maximum, and mean values. The mean affinity with jurisdiction is quite low for both groups.

To measure trust during the interaction with the system, the Weight of Advice (WOA) is used (Sniezek & Buckley, 1995). The calculation that is done for every individual case:

$$WOA = \frac{\text{second estimation} - \text{initial estimation}}{\text{correct sentence} - \text{initial estimation}}$$

Out of the 3443 individual cases, 1147 cases result in a WOA = 0 and therefore the participants in these cases did not change their initial estimate after seeing the recommendation from the system. In 408 individual cases, the initial estimate of the participant was changed to the recommendation that the system gave, which results in a WOA = 1. As mentioned earlier, one participant did follow the recommendations all the time with a mean WOA of 1 and a standard deviation of 0. Another participant did not follow any recommendations at all and stayed throughout the whole experiment with his or her initial estimate. This results in a mean WOA of 0 and a standard deviation of 0. Both of these participants are removed from the data and the remainder of the participants give us 3404 cases to work with.

Based on the WOA values, 125 cases in which the WOA is higher than 2 or lower than 0, are removed. The values for these cases are deemed not reasonable during the experiment since the estimations are not using the recommendation or divert even further from the recommendation and do not indicate any reasonable trust value. The result is that 3355 cases can be used for the multi-level regression.

To perform a multi-level regression, a few assumptions need to be tested. Both of the dependent variables trust and WOA are tested for normality with a Shapiro-Wilk test and the

Skewness and Kurtosis test and in both tests, normality is rejected, which can be seen in figure 1. The histogram does show the residuals are not normally distributed. However due to the sensitivity of these tests when working with a large sample of cases, the analysis will continue.

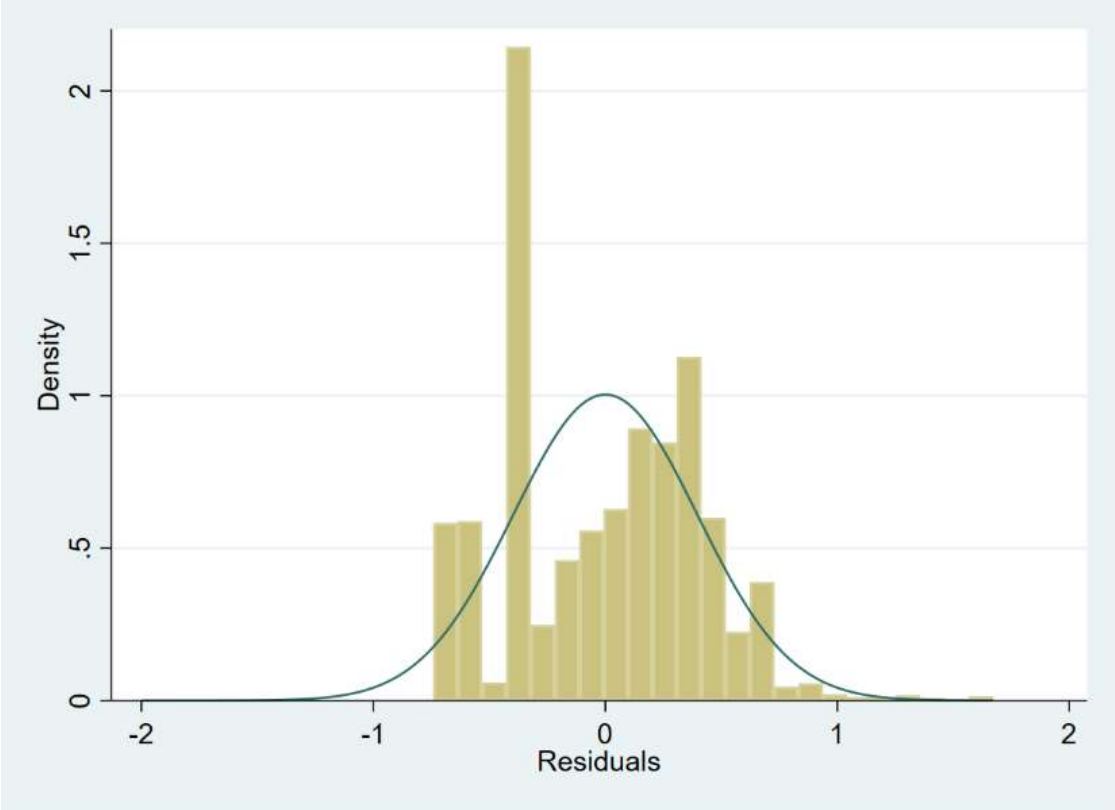

Figure 1: Histogram of the residuals with a normal distribution curve plotted on top. It can be seen that the distribution of the residuals is not normal.

Following these tests is a regression test for only the dependent variable, with id as cluster and no predictor variables. This results in a rho = 0.555 ($p < 0.10$) for Trust and a rho = 0.299 ($p < 0.10$) for WOA. For both variables, this shows that the data is indeed clustered and that a multi-level regression is needed. Multi collinearity between the predictor variables (accuracy, intelligence, and the interaction between accuracy and intelligence (ia)), is rejected with a 2.72 on average (accuracy = 2.33, intelligence = 2.12, ia = 3.71)

Higher accuracy increases trust

To test the first hypothesis, a multi-level regression is done including the variables accuracy, intelligence, and the interaction between accuracy and intelligence (ia). The variables accuracy and intelligence represent the high condition with a 1 and the low condition with a 0.

For the model with WOA as the dependent variable, the only significant variable is accuracy ($p = 0.00$). The intelligence has no significant effect on the WOA ($p = 0.704$) and neither does the interaction between accuracy and intelligence ($p = 0.154$). The results for this model can be found in table 3. In the model with trust as the dependent variable, the effect of accuracy is significant ($p = 0.000$). Just as in the model for WOA, neither the interaction ($p = 0.375$) nor intelligence ($p = 0.796$) are significant. The results for this model can be found in table 4.

The variables age, gender, and indicated knowledge about the jurisdiction are added to the model, to see if any of these variables has an effect on the trust in the system. Only age results in a significant effect ($p = 0.002$) on the WOA. This effect is one-hundredth of the effect of accuracy and therefore all these variables are not used during further analysis.

The desired effect size of 0.5 with a power of 95%, required two groups of 88 participants. However, after dropping some participants the low accuracy condition has 83 participants and a high accuracy condition has 93 participants. A post hoc computation of the achieved power showed that the tests still have a power of 95% using this model with an effect size of 0.5.

Table 3
Results from the multi-level regression with dependent variable WOA

WOA	Coefficient	Std. err.	z	P>|z|
ia	.0874482	.0613504	1.43	0.154
accuracy	.2158404	.0460313	4.69	0.000
intelligence	.0167794	.0442078	0.38	0.704

The results show a significant effect for accuracy, indicating that if accuracy is high, there is a positive effect on the Weight of Advice.

Table 4
Results from the multi-level regression with dependent variable Trust

Trust	Coefficient	Std. err.	z	P>|z|
ia	.4221806	.4763752	0.89	0.375
accuracy	2.204553	.3573782	6.17	0.000
intelligence	-.0887592	.3431781	-0.26	0.796

The results show a significant effect for accuracy, indicating that if accuracy is high, there is a positive effect on the Trust level.

Trust trajectories based on the accuracy

To be able to explore the data for hypothesis 2, a trajectory is made. During the experiment, the order of all the cases was randomized. For every single participant, the order was registered for all cases. Due to an error in registering this, cases 11, 12 & 13 have to be excluded for every participant. These three cases all registered the number in the order where case 14 was presented. This results in gaps within the trajectories for each participant. However, due to the great number of individual cases, there are enough cases to determine a reliable means for trust and WOA to investigate.

After summarizing the WOA and trust values by order, two plots are created for WOA and trust, with the means for all the cases in order. The trajectories can be seen in figure 2

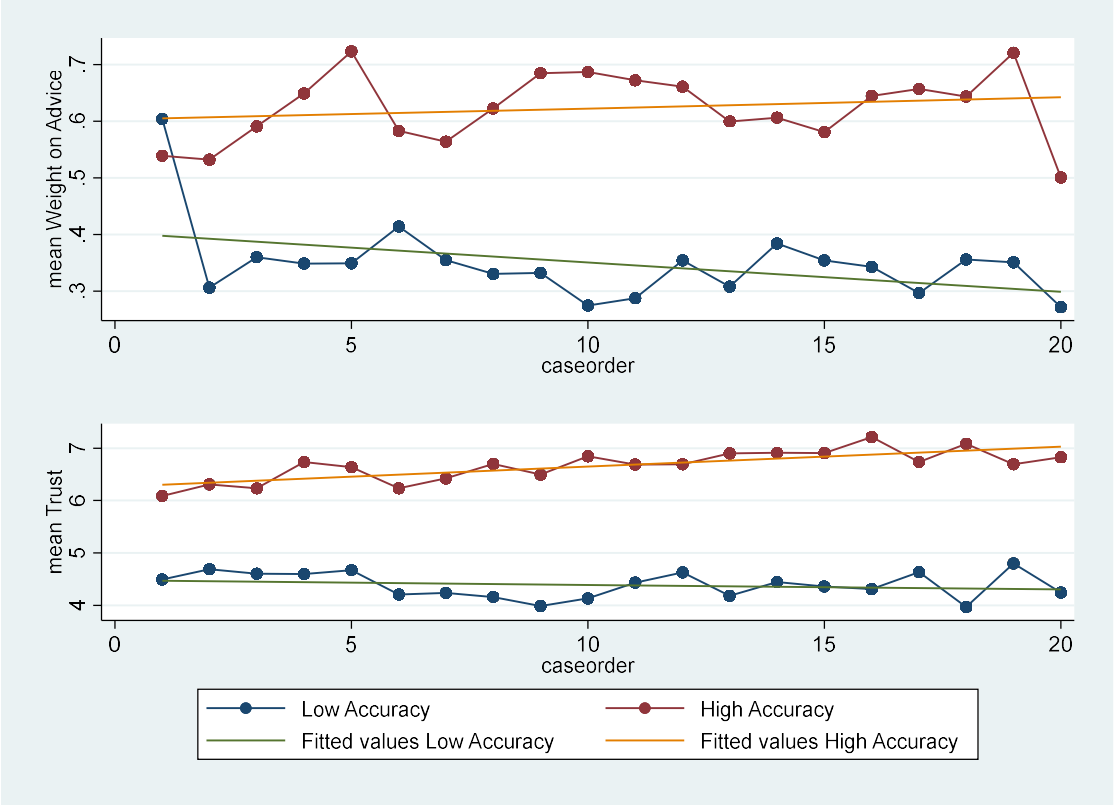

Figure 2: Plot of trust trajectories, where an increasing trend can be seen for high accuracy and a decreasing trend for the low accuracy condition. The WOA for the first case, where the participant has not learned anything about the accuracy, is roughly the same for both conditions.

Discussion

The focus of this research is to look at the effect accuracy has on the human-AI-trust relationship in decision-making advice. To test if the effect is significant, there are two

dependent variables. Trust is a variable that represents the indication of trust by the participant after each case during the interaction with the Artificial intelligence recommendation system. The Weight of Advice is the variable that is measured before the participant learns the correct verdict. Therefore there is one dependent variable during the interaction in each case and one variable at the end when there is already a change in trust.

In general, it can be seen from the results that the higher the accuracy the higher the trust in the system is. In both tests, for trust and WOA, it is clear that accuracy is significant and has a coefficient that is relatively big for the range of WOA (0-1) or trust (0-10) (table 3 and table 4). In all tests, intelligence does not have a significant effect nor does the interaction. Therefore, neither these nor other variables are taken into account when answering the research question:

Research Question: How does the accuracy of Artificial Intelligence decision-making advice influence the human-AI trust trajectory over time?

When interpreting the WOA, a WOA of 1 means full trust in the recommendation, and 0 means that there is no trust at all. If a WOA < 0, the participant does not only distrust the system, but the second estimate is further away from the recommendation than the first. If the WOA is greater than 1, the participant trusts the advice and goes even further than the advice, and therefore overshoots his or her second estimation. If the WOA is even bigger (WOA > 2), the participant overshoots their second estimation and this estimation is even further away from the recommendation than the first one. Thus in this research, all individual cases in which the WOA > 2 or the WOA < 0 are considered not reasonable and dropped from the analysis. While answering the research question and considering the valuable cases where the WOA is reasonable, the following hypothesis is accepted:

Hypothesis 1: When a recommendation system is more accurate, the trust in this system will be higher.

As can be seen in tables 3 and 4, accuracy has a significant effect on both WOA and trust. Trust is measured on a 10-point Likert scale and if the accuracy is high, the trust in the system is increased by 2.2 points. For the WOA, which ranges from 0 to 1, this is 0.2. Therefore it can be concluded that the effect of accuracy is present during the case while filling in the estimates, and after the case when learning about the correct verdict. This result is what could be expected from research that was done earlier (Yu et al., 2017).

In addition to answering the research question, the development of trust over time can be very interesting and therefore a closer look at the trajectory of trust is needed. More specifically, the trajectories of WOA and trust are increasing when accuracy is high and decrease when the accuracy is low. Therefore, the following hypothesis is accepted:

Hypothesis 2: While using a highly accurate recommendation system, the trust is increased over time.

To explore this hypothesis more, the trajectories in figure 2 can be investigated more closely. In the first case, participants have still no idea whether the recommendation is accurate or not. This effect is shown in the graph for WOA. The mean values of WOA for all the first cases are very close for high and low accuracy. After participants have seen the resolution and the correct verdict, the WOA stays either around the same level or drops significantly for the low accuracy condition. This was expected, since gaining trust is slow and losing trust is much easier (Yu et al., 2017). The WOA trajectory is not as smooth as might be expected. This can be influenced by the difficulties of the cases. In general, participants tend to take advice more seriously for difficult cases than for easier ones (Gino & Moore, 2006). However, this is not researched in a recommendation given by an Artificial Intelligence agent.

The indicated trust by participants had to be filled in after they have seen the correct jail time. In the graph, it can be seen that participants in the low accuracy condition already have less trust after seeing the resolution of case 1, compared to the trust level indicated by the WOA. For high accuracy, the trust in the system is high after the first case. In the trust trajectory, A second observation is a slight decrease over time for the low-accurate recommendation system and an increase of trust in the system of almost 1 point that gives high accuracy advice.

In the research, there are also limitations. Jurisdiction is not only about the accuracy of the final verdict, but in a jurisdiction, there are two types of errors. The first one is a severity error, which means that a suspect gets a verdict that is more severe than necessary. The second type of error is a lenience error. This error involves a lower sentence for a suspect than suitable. These errors are the result of one of the two patterns that can be made. An error can be caused by an application error, in which the relevant law and facts are correct, but the sentence does not fit the stated facts. A second pattern that can cause errors is omission. This occurs when not all the facts and relevant laws are stated when the sentence is given to the suspect (Kress Weisbord & Thomas, 2016). These errors are part of the high-risk environment in which the

research is done. However, the focus is on determining the trust trajectory of a human during an interaction with an embedded AI. In this interaction, both the AI and the participant have the same data at hand and thus these errors in a jurisdiction are not considered for this research.

In a high-risk jurisdiction environment, jurisprudential is a particularly important keystone. This means that a verdict of a judge is always based on past verdicts. Since every law case is unique, the subjective opinion of a judge is especially important. For comparing with other law cases in the past, but also for the comparison with current laws and regulations. This is a complex system on its own and that is the reason there is no objective way to determine what is right or wrong (Stephen Russell et al., 2017). The combination of using real-life cases in which a professional judge weighed all the facts properly, and participants with relatively low knowledge about the jurisdiction, is an additional limitation.

Furthermore, the participants are recruited via Prolific, where the criterium was that they have experience in law. This was to ensure a good first estimation could be given and the advice and resolution were comprehensive. An indication by Prolific shows that there are enough potential participants with this criterium and this is the reason that only citizens from the United Kingdom could participate. However, due to limited participation by people with this criterium. This criterium is removed and everyone who was 18 years or older could participate.

The data that is used came from a Dutch database and therefore is connected to the laws, regulations, and sentences in the Netherlands (Rechtspraak.Nl - Zoeken in Uitspraken). This results in a mismatch between citizens from the United Kingdom and the Dutch jurisdiction. In the United Kingdom, common law is used as the legal system, while in the Netherlands this is civil law. The mismatch between these two legal systems is a limitation of this study (Tetley, 1999; The Dutch Court System | Administration of Justice and Dispute Settlement | Government.Nl).

In terms of future research, it would be useful to extend the current findings by examining trust levels while interacting with a system that is improving/training itself or deteriorating during the interaction. In this way, it can become clear what the influence is of accuracy in an evolving system, in combination with the knowledge about the effect of the first case on trust (Tolmeijer et al., 2021) and the accuracy of a system.

Conclusion

Despite the limitations, this research has effectively shown how the accuracy of Artificial Intelligence decision-making advice influences the human-AI trust trajectory over time. A higher accuracy will result in more trust in the recommendation system. Secondly, a high accuracy recommendation will result in an increase in trust over multiple interactions with the system. A low accuracy recommendation will result in reduced trust over time.

References

Bell, S., Berg, T., Grace, A., Zhang, N., & Whaley, J. (2021, June 22). *Advancing AI to make shopping easier for everyone*. Meta AI, ML Applications | Computer Vision.

Benbasat, I., & Wang, W. (2005). Trust In and Adoption of Online Recommendation Agents. *Journal of the Association for Information Systems*, *6*(3). https://doi.org/10.17705/1jais.00065

Butler, J. K., & Cantrell, R. S. (1984). A Behavioral Decision Theory Approach to Modeling Dyadic Trust in Superiors and Subordinates. *Psychological Reports*, *55*(1), 19–28. https://doi.org/10.2466/PR0.1984.55.1.19

Cook, J., & Wall, T. (1980). New work attitude measures of trust, organizational commitment and personal need non-fulfilment. *Journal of Occupational Psychology*, *53*(1), 39–52. https://doi.org/10.1111/J.2044-8325.1980.TB00005.X

Ferràs-Hernández, X. (2018). The Future of Management in a World of Electronic Brains. *Journal of Management Inquiry*, *27*(2), 260–263. https://doi.org/10.1177/1056492617724973

Gino, F., & Moore, D. A. (2006). *Weighing Advice: Effects of Task Difficulty on Use of Advice*.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1–55. https://doi.org/10.1016/0010-0285(92)90002-J

Kizilcec, R. F. (2016). *How Much Information? Effects of Transparency on Trust in an Algorithmic Interface*. https://doi.org/10.1145/2858036.2858402

Kress Weisbord, R., & Thomas, G. C. (2016). *JUDICIAL SENTENCING ERROR AND THE CONSTITUTION*.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*(1), 153–184. https://doi.org/10.1006/IJHC.1994.1007

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/HFES.46.1.50_30392

Li, X., Hess, T. J., & Valacich, J. S. (2008). Why do we trust new technology? A study of initial trust formation with organizational information systems. *The Journal of Strategic Information Systems*, *17*(1), 39–71. https://doi.org/10.1016/J.JSIS.2008.01.001

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Mayer, R. C., Davis, J. H., & David Schoorman, F. (1995). *An Integrative Model of Organizational Trust* (Vol. 20, Issue 3). https://www.jstor.org/stable/258792?seq=1&cid=pdf-

Moorman, C., Deshpandé, R., & Zaltman, G. (2018). Factors Affecting Trust in Market Research Relationships: *Https://Doi.Org/10.1177/002224299305700106*, *57*(1), 81–101. https://doi.org/10.1177/002224299305700106

Nakata, N. (2019). Recent technical development of artificial intelligence for diagnostic medical imaging. In *Japanese Journal of Radiology*. Springer Tokyo. https://doi.org/10.1007/s11604-018-0804-6

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Noordegraaf, M., du Perron, E., & Verberk, S. (2019). *Algoritmes in de rechtspraak. Wat artificiële intelligentie kan betekenen voor de rechtspraak*. www.sdu.nl/service

Rechtspraak.nl - Zoeken in uitspraken. (n.d.). Retrieved June 22, 2022, from https://uitspraken.rechtspraak.nl/

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes*, *62*(2), 159–174. https://doi.org/10.1006/OBHD.1995.1040

Stephen Russell, Ira S. Moskowitz, & drienne Raglin. (2017). *Human Information Interaction, Artificial Intelligence, and Errors*.

Tetley, W. (1999). *Mixed Jurisdictions: Common Law v. Civil Law (Codified and Uncodified)*.

The Dutch court system | Administration of justice and dispute settlement | Government.nl. (n.d.). Retrieved June 22, 2022, from https://www.government.nl/topics/administration-of-justice-and-dispute-settlement/the-dutch-court-system

Tolmeijer, S., Gadiraju, U., Ghantasala, R., Gupta, A., & Bernstein, A. (2021). Second chance for a first impression? Trust development in intelligent system interaction. *UMAP 2021 - Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, 77–87. https://doi.org/10.1145/3450613.3456817

Yin, M., Vaughan, J. W., & Wallach, H. (2019, May 2). Understanding the effect of accuracy on trust in machine learning models. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300509

Yu, K., Conway, D., Berkovsky, S., Zhou, J., Taib, R., & Chen, F. (2017). User trust dynamics: An investigation driven by differences in system performance. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 307–317. https://doi.org/10.1145/3025171.3025219

Zhang, Y., Vera Liao, Q., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305. https://doi.org/10.1145/3351095.3372852