Human Computer Trust

[redacted]

PHYS 410
Senior Seminar
Goshen College

[redacted]

Thesis

Human-computer trust is not only determined by the degree of competence proved by the algorithm, but it is highly influenced by human's own bias and beliefs such as perceived risk, perceived computer's intentions and computer manifestation of humanness.

Outline

1. **Introduction**
2. **What constitutes trust?**
3. **Autonomous Vehicles**
4. **Perceived Risk**
5. **Algorithm Aversion**
6. **Ethics**
7. **Conclusion**

1. Introduction

With technology getting better year over year, computers have become a valuable asset for decision-making and problem-solving, however, at the end of the day it is up to humans to decide whether or not to trust the output of the algorithm. We tend to believe that we rely too much on computers, and that makes us feel a bit uncomfortable when having a direct interaction with them, even up to the point where we are completely skeptical of their output or capabilities often with essentially no reason, i.e. algorithm aversion (Prahl & Van Swol, 2017). At the same time, on many other occasions we discount human advice more than computer-generated advice (Logg et al., 2018). What are the features and characteristics of our relationship with these intelligent machines that play a determining role in trust?

There is an extensive list of published papers that study just one aspect that might influence this relationship, consequently, in this paper I seek to explore not one, but the most important factors that affect human-computer trust. My analysis indicates that human-computer trust is not only determined by the degree of competence proved by the algorithm, but it is highly influenced by human's own bias and beliefs such as perceived risk, perceived computer's intentions and computer manifestation of humanness.

2. What constitutes trust?

Since it is quite hard to give a formal definition of trust, I will discuss a few characteristics that constitute it and are relevant for my following discussion. When thinking about this concept, one generally considers the degree of comfortableness and reliance he/she has on other people. Therefore, in past literature it has been related to the idea of perceived “goodwill” we see in others (Baier, 1994). The more friendly and empathic we see someone, the easier it is going to be to build trust. Since trust arises from the fact that humans live socially and the necessity to cooperate, if we see a warmth factor, i.e. willingness to do good, unselfishness, politeness, from another individual, then the level of trust we might have increases (Judd et al., 2005).

However, it seems obvious that just the observed goodwill that someone might present is not enough in order to fully build trust. For example, a doctor might be one of the kindest persons one has met, nevertheless if the doctor does not show the ability to perform an operation or to correctly diagnose a disease, the level of trust is automatically reduced in spite of his/her evident warmth factor. Therefore, it is clear the need to introduce another characteristic that constitutes trust to account for these types of situations: the competence factor (Philip & Stephan, 2018). The trustee has to explicitly show her ability to carry out a desired task, in order for the truster to feel comfortable relying on her.

Then, there are two important characteristics that constitute trust: perception of positive intentions and competence. But, how are these two relevant for this paper? How do they connect with the relationship we might have with computers?

In a society where computers are taking up more and more jobs and are constantly increasing their presence in the world, we are getting used to encountering them in many daily life situations. Since they are replacing a job that a human could be doing, it is normal that we tend to treat them as social agents (Reeves & Nass, 1996), that is in a similar way that we would treat a person that was performing that same job.

Therefore, it seems reasonable to be judging, in terms of trust, a computer using the same criteria we use when interacting with humans. Consequently, just as we are more likely to trust people whose perceived behavioral intentions are related to empathy, kindness and friendliness; a study suggests that if computers present their advice in a “polite and friendly manner,” users are more willing to trust their advice and results (Parasuraman & Miller, 2004).

Computers striving to be as human as possible seems to play a big role in human-computer trust. Nevertheless, it is not just the verbal aspect that helps in that direction but “anthropomorphic computerized non-verbal signals such as smiling, eye-contact, and immediacy cues, thus can play an important role for the perceived warmth” (Philip & Stephan, 2018). A picture is worth a thousand words, and so is a facial expression, in fact a person can infer someone’s trustworthiness (as well as attractiveness, likeability, competence, etc) in just a matter of 100 milliseconds (Willis & Todorov, 2006).

3. Autonomous vehicles

In a world where automation is a priority in order to increase efficacy and efficiency, self driving cars clearly fall into this category. Data shows that car accidents have been among the top ten leading causes of death worldwide in 2020, according to the World Health Organization. Humans are still showing a sense of irresponsibility and lack of seriousness while driving; common behaviours include speeding, distracted driving, drowsiness, driving while being under the influences of alcohol, etc. Despite all skepticism about self-driving cars, these facts make us question whether or not humans really are more reliable than autonomous vehicles.

As for every revolutionary worthy invention there is always the transition that goes from what we are used to, to the new innovation. In self driving cars, this is highly dependent on the rate of acceptance which is directly proportional to the trust associated with them. Given the potential benefits of implementing autonomous vehicles for humanity, designers need to carefully consider what aspects are crucial for improving the passengers' trust for this technology.

A study looked into how the engineers that built a self driving car would rate the car's reliability across a variety of different scenarios and compare their responses to the rating done by randomly picked passengers (Walker, 2020). The results show discrepancies between the engineers' rates and the passengers' rates, which seems obvious since the engineers knew exactly how the car would behave in each scenario given their knowledge of the algorithm. On the other hand, the passengers did not have a way to check how the autonomous vehicle was going to react, therefore leading to gut feeling guesses based on their level of skepticism of the algorithm. Furthermore, they did not have a way to evaluate the competence factor, in addition to not having any sort of feedback that boosted the warmth factor described earlier. In essence, the algorithm was similar to a black box. So, how do you expect someone to trust their life to an unfamiliar technology that provides no real-time feedback whatsoever?

In the end of this study, recommendations were given on how to improve human computer trust in this field based on the passengers' ratings. For example, it was suggested that a if the autonomous car presented a screen where real-time feedback was provided, so that the passenger could see what the car is detecting (other cars, traffic lights, stop signs, etc) and the consequent path that it will take, if these two match with what the passenger might see and do in a this situation, then the algorithm has succeeded in showing the competence factor (Philip & Stephan, 2018), which in turn reduces their level of anxiety, leading to an increase of human-computer trust.

As previously stated, the transition from what we are used to to the innovative technology might be psychologically challenging, therefore just by explicitly showing the competence factor might not be enough, and consequently the use and implementation of the warmth factor is demanded. It has been shown that “participants trusted a virtual driver more when the driver's computer-generated face was based on their own face, thus increasing perceived similarity” (Philip & Stephan, 2018, Verbena et al., 2015). Thus, by implementing a more humane feature to the self-driving car and keeping everything else the same, participants of the study unconsciously increased their trust in the autonomous vehicle.

Furthermore, another way to express the warmth factor together with the competence factor is by having a conversational user interface (CUI), as it was done in a study where anthropomorphism was used “to shape the interaction, by applying Gricean Maxims[1] (i.e. guidelines for effective conversation)” (Ruijten et.al., 2018). The CUI was presented in a way where the algorithm justified its actions by giving brief and informative arguments about the reasoning behind its decisions and own limitations, as well as showing what the algorithm was seeing (i.e. detecting on the road). This method of transparency between the passenger and the automated car, made the passengers rate the CUI higher than the simple graphical user interface (same as the CUI but without the conversation abilities) in all studied areas: trust, perceived intelligence, anthropomorphism and likability. Therefore, we see that by adding a degree of humanness and making the real-time feedback more obvious (i.e. implementing conversation) in a way that helps to better communicate the behavioral decisions of the autonomous vehicle is shown to have a positive impact on trust and acceptance.

4. Perceived Risk

Even though people might find self-driving vehicles exciting and potentially beneficial, most are still hesitant to relinquish control of the vehicle. On the other hand, people are just fine with using a smartwatch to measure their heart rate while working out, even though its output is shown to be inconsistent. In both cases, an algorithm is used to perform a required task, then, what is the difference between one and the other? The answer is perceived risk.

A Harvard study used a number of experiments to show how a person would react against the advice coming from an algorithm or coming from another person. Their results show that people tend to discount advice from other people and rely more on algorithmic advice when the stakes are low; for example, when estimating the weight of a person (Logg et al., 2018). Here the possibility of losing something important for the person who is deciding whether to take the advice from the algorithm or the human is nonexistent, therefore the perceived risk is very low.

On the other hand, in fields like e-commerce where users need to be more cautious about their actions since they are aware of the possibility of getting hacked and consequently losing money, the perceived risk is higher, which “decreases risk-taking propensity, and in turn, risk-taking propensity significantly impacts behavior intention" (Hansen, 2017), meaning that the user is scared away and will not buy a product through such online platform. Nevertheless, if this perceived risk is reduced by increasing the perceived trust, which, for example, can be achieved by making the platform easier to use, risk-taking propensity increases and, consequently, behavioral intention as well. (Hansen, 2017).

As stated before, skepticism about using self-driving cars is quite high, in fact a survey of 499 participants suggests that autonomous vehicles should be from four to five times safer than a human driven vehicle in order for the participants to be comfortable using self-driving cars (Liu, et al., 2019). It seems already quite evident that bias plays a big role in these surveys since a big part of the participants are likely to have not experienced at least an autonomous vehicle ride, but yet they seem to be pretty confident with their opinion about whether or not a self-driving car is trustworthy. In literature, this mentality seems to be associated with high levels of perceived risk due to the newness of this technology in contrast to what we have been accustomed to so far (Brell, 2019). This process of adaptation to get to know the unknown comes with uncertainty and skepticism and, even though perceived risk is initially present, depending on the ongoing experience this level increases or decreases (Brell, 2019).

Furthermore, an article provided by Pennsylvania State University shows that participants were more comfortable sharing their credit card information with a computerized travel agent rather than with a human travel agent (Swayne, 2019). This suggests that while participants did not have a specific knowledge about the functioning of the machine, they felt like the human travel agent could behave in an unlawful way taking advantage of their private information. So in this case, it was not a matter of demonstrating the competence or warmth factors, but it was immediately determined by the participants’ own biases. Similarly, another study suggests that humans trust a machine-generated password more than a human expert-generated password in order to keep their private online information safe (Alqahtani, 2019). These results suggest that participants might have thought that if a human can come up with such a password, then another human expert can as well, however, since the machine is unpredictable, then in this case it might be a wiser choice to go for the machine-generated password. This can also be seen in a study where identical resumes of candidates called “John” and “Jennifer” were evaluated by scientists across the US to see how gender could affect their selection process for a STEM work position. Each scientist was given either the fictitious John or fictitious Jeniffer resume. It turns out that even by having the same content, Jeniffer was perceived as significantly less competent. So it seems like in a case for evaluating a resume, where the subjectivity and bias of the evaluator can be detrimental in some cases, a computer would do a much better/objective job in advising which candidate is more competent for a determined job position. All of these studies show that the amount of trust we could have in a machine, it might be quite influenced by the assumptions (i.e. unlawful behaviour, subjectivity, vulnerability, emotional behaviour, etc.) we have about the other choice, in this case a human, rather than direct knowledge of the algorithm.

5. Algorithm Aversion

As discussed before, our own bias about the available options when accounting for advice is inevitably influential. Even though humans want to make an informed decision about something important (i.e. when the stakes are high), they often discount the advice coming from a machine, especially when compared to advice coming from a human. Hopefully there is a reason behind this, however, when there is an irrational discounting of machine advice, then it is called algorithm aversion.

A study shows that the advice utilization coming from an algorithm was greatly reduced in comparison to the advice utilization coming from a human when in both cases the initial advice was proved to be equally wrong (Prahl & Van Swol, 2017). This could be explained by the “perfection schema” (Madhavan & Wiegmann, 2007) which describes the fact that we generally expect the advice coming from an algorithm to be flawless and perfect, therefore “when they make a mistake this feels especially negative to the advice recipient and they lose trust rapidly in the advisor” (Prahl & Van Swol, 2017). Furthermore, if the advice coming from the algorithm is shown to be correct after the first wrong prediction, humans seem to choose human advice. So it seems that humans are more empathetic with other humans since they are aware that we are not perfect and that we are quite likely to make mistakes and, hopefully, learn from them. On the other hand, algorithm aversion seems to come from the “perfection schema” mentality and from the idea that computers do not learn from their mistakes so that they are either consistently right or consistently wrong.

Other factors can influence algorithm aversion, such as perceived task objectivity. Human willingness to use an algorithm is negatively related to the objectivity of the task (Castelo et al., 2019). This means that the more subjective a human perceives the task to be, the less likely it is to be considered trustworthy, then in such cases other humans are preferred to perform the task. This can be interpreted the other way, so that the more objective a task needs to be we should reduce the amount of subjectivity present which could be achieved by relying on a computer than on a human, just like the study previously discussed where two different resumes were analysed.

6. Ethics

I have been focusing on the aspects that constitute human-computer trust and offered suggestions on how we could increase this trust by making computers more human, but we should be aware of the concerns and ethical questions about some of these approaches.

Jaana Porra from the University of Houston, raises the question of “how human should computer-based human-likeness appear?” It emphasizes that if computers keep displaying emotions which they do not feel (since they are programmed to), humans will progressively lose humanness: “We are used to hearing machines say: ‘I am sorry’ but we need to consider what meaningful is left for us to say when we really feel sorry” (Porra, 2019). Consideration must be taken when implementing emotions in computers since it could impact genuine humanness.

7. Conclusion

In this paper we have discussed various situations where the basis of human-computer trust proved to be more complex and trivial than expected. We have seen that humans’ own bias and beliefs about the computer itself might induce an erroneous reaction against the algorithm’s output which leads to bad decision-making (Prahl & Van Swol, 2017). Thus, human-computer trust goes beyond just the fact of the algorithm proving to be right or wrong in contrast to the human, but rather there are factors such as human’s observations of the computer’s perceived intentions (i.e. empathy, kindness, selfishness, etc.) that affect trust (Judd et al., 2005). These familiar behaviours are characteristic of humans, therefore we see that computers that show a degree of humanness are more likely to be trusted than the ones that do not display this feature (Parasuraman & Miller, 2004). In addition to that, we have seen that non-related computer external factors, i.e factors that are out of the machine’s control and therefore reside in humans’ brain, such as perceived risk, directly condition the willingness of humans to rely on a computer without having previous knowledge of the algorithm itself (Hansen, 2017). Furthermore, studies relating autonomous vehicles (Walker, 2020; Ruijten et.al., 2018) have indicated the importance of the degree of awareness of the user about the machine’s thinking, meaning that algorithms that operate as “black boxes” transmit uncertainty, which consequently leads to a decrease of the participant’s trust.

Therefore, I can now conclude that human-computer trust is not only determined by the degree of competence proved by the algorithm, but it is highly influenced by human’s own bias and beliefs such as perceived risk, perceived computer’s intentions and computer manifestation of humanness.

Endnotes

(1) Gricean maxims: the maxim of quantity ("give as much information as is needed, and no more"), the maxim of quality ("do not give information that is false or that is not supported by evidence"), the maxim of relation ("be relevant and say things that are pertinent to the discussion"), the maxim of manner ("be clear, brief, and orderly, and avoid obscurity and ambiguity").

References

Alqahtani, Saeed, et al. "Human-Generated and Machine-Generated Ratings of Password Strength: What Do Users Trust More?" *Eai Endorsed Transactions on Security and Safety*, vol. 6, 2019. https://eudl.eu/pdf/10.4108/eai.13-7-2018.162797

Baier, A. C. *Moral Prejudices: Essays on Ethics*. Cambridge, MA: Harvard University Press, 1994.

Brell, Teresa, et al. "Scary! Risk Perceptions in Autonomous Driving: The Influence of Experience on Perceived Benefits and Barriers." Risk Analysis, vol. 39, no. 2, 2019, pp. 342–357., doi:10.1111/risa.13190.

Castelo, Noah, et al. "Task-Dependent Algorithm Aversion." *Journal of Marketing Research*, vol. 56, no. 5, 2019, pp. 809–825., doi:10.1177/0022243719851788.

Dietvorst BJ, et al. "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err." *Journal of Experimental Psychology. General*, vol. 144, no. 1, 2015, pp. 114–26., doi:10.1037/xge0000033.

Hansen, Jared & Saridakis, George & Benson, Vladlena. "Risk, Trust, and the Interaction of Perceived Ease of Use and Behavioral Control in Predicting Consumers' Use of Social Media for Transactions." Computers in Human Behavior. 2017. https://www.researchgate.net/publication/320994032_Risk_Trust_and_the_Interaction_of_Perceived_Ease_of_Use_and_Behavioral_Control_in_Predicting_Consumers'_Use_of_Social_Media_for_Transactions

Hegner, Sabrina M, et al. "In Automatic We Trust: Investigating the Impact of Trust, Control, Personality Characteristics, and Extrinsic and Intrinsic Motivations on the Acceptance of Autonomous Vehicles." *International Journal of Human-Computer Interaction*, vol. 35, no. 19, 2019, pp. 1769–1780.

Judd, C. M., James-Hawkins, L., Yzerbyt, V., and Kashima, Y. (2005). Fundamental dimensions of social judgment: understanding the relations between judgments of competence and warmth. *J. Personal. Soc. Psychol.* 89, 899–913. doi: 10.1037/0022-3514.89.6.899

Kaspar Raats, et al. "Trusting Autonomous Vehicles: An Interdisciplinary Approach." *Transportation Research Interdisciplinary Perspectives*, vol. 7, 2020, pp. 100201–100201. doi:10.1016/j.trip.2020.100201.

Kulms, Philipp, and Stefan Kopp. “A Social Cognition Perspective on Human–Computer Trust: The Effect of Perceived Warmth and Competence on Trust in Decision-Making with Computers.” *Frontiers in Digital Humanities*, vol. 5, 2018, doi:10.3389/fdigh.2018.00014.

Liu, Peng & Yang, Run & Xu, Zhigang. “How Safe Is Safe Enough for Self-Driving Vehicles?” *Risk Analysis*, vol. 39, no. 2, 2019, pp. 315–325., doi:10.1111/risa.13116.

Logg. Jennifer M., Julia A. Minson & Don A. Moore, “Algorithm Appreciation: People Prefer Algorithmic To Human Judgment.” Harvard Business School (2018) https://www.hbs.edu/faculty/Publication%20Files/17-086_610956b6-7d91-4337-90cc-5bb5245316a8.pdf

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

Parasuraman, Raja, and Christopher A Miller. “Human-Computer Etiquette: Managing Expectations with Intentional Agents - Trust and Etiquette in High-Criticality Automated Systems.” *Communications of the Acm*, vol. 47, no. 4, 2004, p. 51. doi:10.1145/975817.975844.

Porra, Jaana, et al. “‘Can Computer Based Human-Likeness Endanger Humanness?" - a Philosophical and Ethical Perspective on Digital Assistants Expressing Feelings They Can't Have.’” Vol. Day:9, 2019. https://doi.org/10.1007/s10796-019-09969-z

Prahl, Andrew, and Lyn Van Swol. “Understanding Algorithm Aversion: When Is Advice from Automation Discounted?” *Journal of Forecasting*, vol. 36, no. 6, 2017, pp. 691–702., doi:10.1002/for.2464.

Reeves, B., and Nass, C. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge: Cambridge University Press. 1996. https://www.researchgate.net/publication/37705092_The_Media_Equation_How_People_Treat_Computers_Television_and_New_Media_Like_Real_People_and_Pla

Ruijten, P.A.M.; Terken, J.M.B.; Chandramouli, S.N. Enhancing Trust in Autonomous Vehicles through Intelligent User Interfaces That Mimic Human Behavior. *Multimodal Technol. Interact.* 2018, *2*, 62. https://doi.org/10.3390/mti2040062

Simpson, Thomas W. “What Is Trust?” *Pacific Philosophical Quarterly*, vol. 93, no. 4, 2012, pp. 550–569., doi:10.1111/j.1468-0114.2012.01438.x.

Swayne, Matt, “People more likely to trust machines than humans with their private information.” *Pennsylvania State University*. Phys.org. May 10, 2019. https://phys.org/news/2019-05-people-machines-humans-private.html

“The top 10 causes of death.” *World Health Organization*. 9 December 2020. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death#:~:text=The%20top%20global%20causes%20of,birth%20asphyxia%20and%20birth%20trauma%2C

Verberne, F. M. F., Ham, J., and Midden, C. J. H. (2015). Trusting a virtual driver that looks, acts, and thinks like you. *Hum. Fact.* 57, 895–909. doi: 10.1177/0018720815580749

Walker, F, et al. “Do Engineer Perceptions About Automated Vehicles Match User Trust? Consequences for Design.” *Transportation Research Interdisciplinary Perspectives*, vol. 8, 2020. doi:10.1016/j.trip.2020.100251.

Willis, Janine & Todorov, Alexander. “First Impressions : Making Up Your Mind After a 100-Ms Exposure to a Face.” *Psychological Science*, vol. 17, no. 7, 2006, pp. 592–598., doi:10.1111/j.1467-9280.2006.01750.x.