

I Feel Safe with the Prediction: The Effect of Prediction Accuracy on Trust

Lisa Graichen¹^a and Matthias Graichen²^b

¹Department of Psychology and Ergonomics, TU Berlin, Marchstraße 23, Berlin, Germany

²Universität der Bundeswehr München, Munich, Germany

Keywords: XAI, HCXAI, Mental Model, Trust.

Abstract: Mental model building and trust are important topics in the interaction with Artificial Intelligence (AI)-based systems, particularly in domains involving high risk to user safety. The presented paper describes an upcoming study on applying methods from Human-Centered Explainable AI (HCXAI) to investigate the influence of AI accuracy on user trust. For the interaction with an AI-based system, we use an algorithm designed to support drivers at intersections by predicting turning maneuvers. We will investigate how drivers subjectively feel towards different presented and implemented accuracy levels. Participants will be asked to rate their respective levels of trust and acceptance. Moreover, we will investigate, whether trust and acceptance are depending on personality traits.


1 INTRODUCTION


Artificial Intelligence (AI) is being applied in more and more areas of research and daily life. However, each technology has predefined conditions under which it can reach its potential or constraints that may lead to system failure that users should be aware of in order to react properly. Depending on the context and purpose of the application, system failure can have critical consequences, such as in critical traffic situations while driving with AI-based Advanced Driver Assistance Systems (ADAS) or in autonomous driving.

The development of ADAS using sensors and algorithms and In-Vehicle Information System to support the driver with the primary and secondary driving task has already been a topic for decades (Bengler et al., 2014), the next huge step in this development is supposed to be the availability of fully autonomous driving functions (e.g., see @CITY, 2018 or KARLI, 2022). However, the bare development cannot be the only goal, but also facilitating the user to build trust and adequate mental models about functions and constraints to realize possible benefits like increased driving safety, comfort and efficiency (Bengler et al., 2018).

When it comes to user's trust in AI applications we need to mention that trust cannot only be too low, but also too high. If trust is too uncritical, which means it cannot be justified by the real possibilities of the application, it may lead to undesired or even dangerous situations. So called overtrust can have several dimensions according to Itoh (2012): 1) User are not aware of possible malfunctions, 2) users misuse applications for situations they were not developed for and 3) users fundamentally misunderstand the functioning of the system. On the other hand, if users would not trust a system enough, they will may not use the respective system on a regular basis or even not at all and cannot make use of its benefits. The mentioned relation between a user's mental model, their trust in automated systems, and its actual use has been investigated in several studies (e.g. see Ghazizadeh et al., 2012 or Beggiato & Krems, 2013).

From a developer and manufacturer perspective, it is important to know which level of accuracy users would find acceptable or worth trusting. Methods of (HC)XAI have the potential to support the process of understanding the AI's functionality and constraints and to build appropriate trust levels, therefore we used an explanation that was developed and evaluated

^a <https://orcid.org/0000-0001-8634-6843>

^b <https://orcid.org/0009-0009-2774-154X>

before (L. Graichen et al., 2022). In the presented study we are addressing the following research questions:

RQ1: How do users feel towards different prediction accuracy levels?

RQ2: How do different personality traits influence trust towards prediction accuracy levels?

2 METHODS

2.1 Design and Independent Variables

A one-way repeated measure design was chosen, with three prediction accuracy levels representing the factor levels. The experiment will have two parts. In the first part participants will rate their trust levels for different prediction accuracy levels in an online questionnaire, as this gives us the opportunity to ask for more different accuracy levels. Prediction accuracy levels used in this part will be starting with 55% up to 95% in steps of five percent. In the second part participants will sit in a driving simulator and perceiving three of the prediction accuracy levels and respective warnings of cyclists during real driving. Prediction accuracy levels used in this part of the experiment will be 70%, 80% and 90%. Accuracy levels have been chosen based on realistic levels the algorithm reaches at different points in time between 100m and 0m distance to real life intersections.

2.2 Participants

An opportunity sample of about 35 persons will be selected using the website of TU Berlin. It will mostly contain human factors students. This research will comply with the tenets of the Declaration of Helsinki, and informed consent will be obtained from each participant.

2.3 Facilities and Apparatus

A fixed-based driving simulator will be used for the study. Participants will sit in a driving simulation mock-up with automatic transmission. We will use Carla as a driving simulation environment (Dosovitskiy et al., 2017). To record the driver interactions with the IVIS, a camera will be positioned towards the driver on a table close to the mock-up. To investigate the visual behavior of the participants towards the displays, a Pupil Labs eye tracker will be used (see <https://pupil-labs.com/>). As a driving scenario, a city scenario was chosen from the tracks that are provided in Carla (see Figure 1a).

A 15-inch screen will be mounted on the center console to display the IVIS with the warnings (see Figure 1b).

2.4 Training Material

Before participants start the two parts of the experiment, there will be a short training of the AI used. We will use an algorithm that was originally designed to support drivers at intersections by predicting turning maneuvers. When a driver approaches an intersection, the algorithm predicts whether they are likely to turn right or go straight (see M. Graichen, 2019 and Liebner et al., 2013). One application for such an algorithm is to present drivers a dynamic warning in situations where they intend to turn and another traffic participant (e.g., a cyclist) could potentially cross the anticipated trajectory. To predict a turning maneuver, the algorithm mainly uses vehicle data about the driver's speed and acceleration and compares this information to the behavioral pattern drivers typically show when approaching intersections. However, in certain situations, the algorithm may falsely predict right turning maneuvers when the preceding traffic is not considered. For example, a vehicle decelerating ahead has an indirect impact on the driver, which can lead to a velocity profile similar to the pattern shown before turning maneuvers.

For the training material we created two short videos, which are based on vehicle data and videos recorded on a trip through Munich, Germany. The videos show the road scenery from a driver's perspective to achieve a realistic impression of approaching an intersection. For the video showing a system limit (e.g., the driver intends to go straight but the system predicts a turning maneuver, as there is a preceding vehicle), respective driving scenarios were chosen.

The video material used consists of a compilation of the video showing the road scenery and pre-processed results of the prediction algorithms, which are presented as diagrams at the bottom left corner (see Figure 2). One diagram shows the dynamic maneuver probability for "turning right" or "going straight." The second diagram provides information on the driver's velocity as well as the anticipated velocity models for turning right or going straight for the respective intersection. Additionally, the user sees information about the potential velocity corridor for the two possible maneuvers. This makes it possible for the user to compare if the current velocity approximates closer to a trajectory typically shown

when turning right or going straight at any given time of the video.

2.5 Procedure

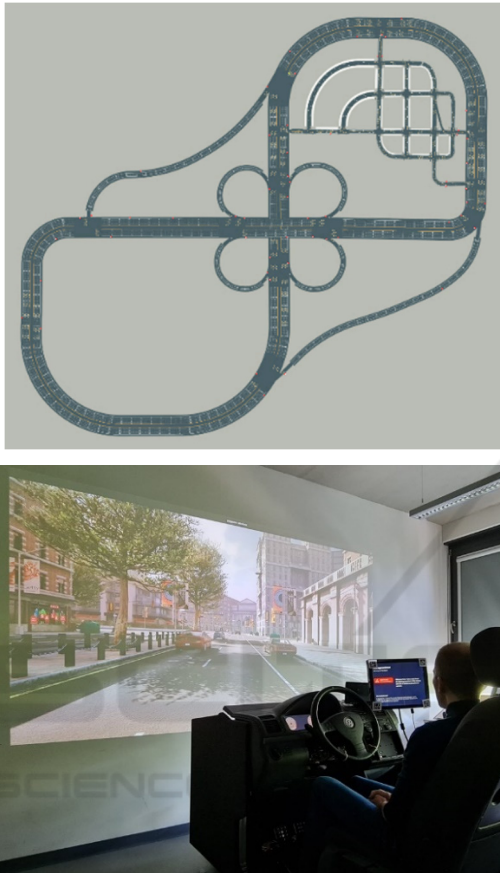


Figure 1: 1a-1b: Track used for the study/Simulator setting.

Upon arrival, participants will be introduced to the simulator, navigation device, and cyclist warning. Afterwards the two training videos, one with a system malfunction, will be presented. Before the experiment starts, participants will be introduced to the eye-tracking device, which will then be calibrated individually. Participants will then rate their individual trust levels towards different prediction accuracy levels in the first part of the experiment. Afterwards they will drive one training trip and three experimental trips in the driving simulator. Before each trip they will be told how accurate the AI is detecting their turning intention (70% vs. 80% vs. 90%). The number of perceived malfunctions during the trip will be adjusted to this accuracy level. After each trip they will answer questionnaires pertaining to trust, personality traits, and workload afterwards. Participants will be told to drive according to the

German Road Traffic Act and keep to the standard velocity allowed in urban areas.

2.6 Dependent Variables

For dependent variables, participants will complete a questionnaire pertaining to trust in the system's decision making (Pöhler et al., 2016), which is based on (Jian et al., 2000). Moreover, participants were asked to rate their level of workload using NASA-TLX (Hart & Staveland, 1988), acceptance (Van der Laan et al., 1997), affinity towards technology (Neyer et al., 2012) and driving behavior (Reason et al., 1990).

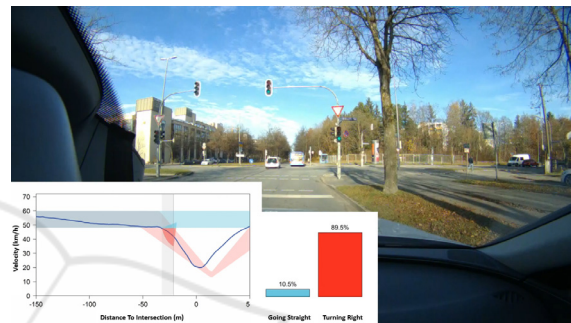


Figure 2: Training material used in the comprehensive explanation of the algorithm.

3 IMPLICATIONS

Understanding the constraints and predefined conditions under which AI-based systems are designed to operate and produce results is critical, especially when these systems are used in domains with high risk to user safety like driving. There are indications that users tend to both a) request a high level of accuracy to be willing to use fully automated cars (Shariff et al., 2021), as they tend to believe they perform better than the average of all drivers (better-than-average effect, see Alicke & Govorun, 2005, but also b) may overtrust AI and technical systems in general. Therefore, it is important to understand how users build appropriate mental models and gain trust and acceptance in using AI-based systems. With the presented study, we aim to address the question of how users feel towards different prediction accuracy levels.

REFERENCES

- @CITY. (2018). *Automated driving in the city*. <https://www.atcity-online.de/>
- Alicke, M. D., & Govorun, O. (2005). The Better-Than-Average Effect. In M. D. Alicke, D. A. Dunning, & J. Krueger (Eds.), *The Self in Social Judgment* (pp. 85–106). Psychology Press.
- Beggiato, M., & Krems, J. F. (2013). The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation Research Part F: Traffic Psychology and Behaviour*, *18*, 47–57. <https://doi.org/10.1016/j.trf.2012.12.006>
- Bengler, Klaus; Dietmayer, Klaus; Färber, Berthold; Maurer, Markus; Stiller, Christoph; Winner, Hermann (2014): Three decades of driver assistance systems. Review and Future Perspectives. In: IEEE Intelligent Transportation Systems Magazine 6 (4), S. 6–22. DOI: 10.1109/MITS.2014.2336271.
- Bengler, K., Drüke, J., Hoffmann, S., Manstetten, D., & Neukum, A. (Eds.). (2018). *UR:BAN Human Factors in Traffic. Approaches for Safe, Efficient and Stressfree Urban Traffic*. Springer. <https://doi.org/10.1007/978-3-658-15418-9>
- Dosovitskiy, A., Ros, G., Codevilla, F., & López, Antonio, Koltun, Vladlen. (2017). CARLA: An Open Urban Driving Simulator. In *Proceedings of the 1st Annual Conference on Robot Learning* (pp. 1–16).
- Ghazizadeh, M., Lee, J. D., & Boyle, L. N. (2012). Extending the Technology Acceptance Model to assess automation. *Cognition, Technology & Work*, *14*(1), 39–49. <https://doi.org/10.1007/s10111-011-0194-3>
- Graichen, L., Graichen, M., & Petrosjan, M. (2022). How to Facilitate Mental Model Building and Mitigate Overtrust Using HCXAI. In *Workshop Human-Centered Perspectives in Explainable AI*. Symposium conducted at the meeting of ACM, fully virtual.
- Graichen, M. (2019). *Analyse des Fahrverhaltens bei der Annäherung an Knotenpunkte und personenspezifische Vorhersage von Abbiegemanövern* [Doctoral thesis]. Universität der Bundeswehr München, Neubiberg. <http://athene-forschung.rz.unibw-muenchen.de/129783>
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology. Human Mental Workload* (Vol. 52, pp. 139–183). Elsevier. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Itoh, M. (2012). Toward overtrust-free advanced driver assistance systems. *Cognition, Technology & Work*, *14*(1), 51–60. <https://doi.org/10.1007/s10111-011-0195-2>
- Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04
- KARLI. (2022). *Künstliche Intelligenz (KI) für Adaptive, Responsive und Levelkonforme Interaktion im Fahrzeug der Zukunft*. <https://karli-projekt.de/>
- Liebner, M., Klanner, F., Baumann, M., Ruhhammer, C., & Stiller, C. (2013). Velocity-Based Driver Intent Inference at Urban Intersections in the Presence of Preceding Vehicles. *IEEE Intelligent Transportation System Magazine*, *5*(2), 10–21. <https://doi.org/10.1109/MITS.2013.2246291>
- Neyer, F. J., Felber, J., & Gebhardt, C. (2012). Entwicklung und Validierung einer Kurzsкала zur Erfassung von Technikbereitschaft. *Diagnostica*, *58*(2), 87–99. <https://doi.org/10.1026/0012-1924/a000067>
- Pöhler, G., Heine, T., & Deml, B. (2016). Itemanalyse und Faktorstruktur eines Fragebogens zur Messung von Vertrauen im Umgang mit automatischen Systemen. *Zeitschrift Für Arbeitswissenschaft*, *70*(3), 151–160. <https://doi.org/10.1007/s41449-016-0024-9>
- Reason, J., Manstead, A., Stradling, S., Baxter, J., & Campbell, K. (1990). Errors and violations on the roads: A real distinction? *Ergonomics*, *33*(10-11), 1315–1332. <https://doi.org/10.1080/00140139008925335>
- Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2021). How safe is safe enough? Psychological mechanisms underlying extreme safety demands for self-driving cars. *Transportation Research Part C: Emerging Technologies*, *126*, 103069. <https://doi.org/10.1016/j.trc.2021.103069>
- Van der Laan, J. D., Heino, A., & Waard, D. de (1997). A simple procedure for the assessment of acceptance of advanced transport telematics. *Transportation Research Part C: Emerging Technologies*, *5*(1), 1–10. [https://doi.org/10.1016/S0968-090X\(96\)00025-3](https://doi.org/10.1016/S0968-090X(96)00025-3)