Why we need to measure Trust in Artificial Intelligence

SIMON TEVES-HUMME, AILEEN BÜCKER, ANILCAN POLAT, JELLE HÜNTELMANN, and PROF. DR. EMANUELL MÜLLER

Artificial Intelligence (AI) keeps evolving and grows in importance in everyday life but the usefulness and the success of this integration depend on the user's trust in AI technology. Therefore a constant cycle of (re)designing an AI model and measuring user trust in that AI model is highly recommended to compare to and match the desired interaction. This review discusses current methods of measuring trust, highlights difficulties in universalizing trust, and presents influencing aspects for trust development. Regarding trust measures it focuses on the Trust between People and Automation Scale (TPA) and the Trust scale for explainable AI (TXAI) and the need for standardized trust scales in the AI context. Trust development impacts are divided into cognitive and emotional trust aspects and separated for robotic, virtual, and embedded AI appearances. We suggest establishing TPA and TXAI as standardized questionnaires and appealing to the importance of integrating regular measuring in the development of (high-risk) AI models.

ACM Reference Format:
Simon Teves-Humme, Aileen Bücker, Anilcan Polat, Jelle Hüntelmann, and Prof. Dr. Emanuell Müller. 2023. Why we need to measure Trust in Artificial Intelligence. 1, 1 (November 2023), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

With Artificial Intelligence Systems cultivating more and more components of our society, the psychological factors of the interactions with such become more apparent. In this paper specifically, the factor of trust will be the subject of discussion. Ella Glikson defines trust in her paper "Human trust in artificial intelligence: review of empirical research" as follows: "[The] tendency to take a meaningful risk while believing in a high chance of positive outcome." [2]. She thereby highlights the weighing up of risk and reward when it comes to trust, which is also reflected in later aspects covered in this report like calibrated trust and trust trajectories for robotic, virtual, and embedded AI. The need for a belief is necessary due to the opaque nature of black box models, regarding what the output is versus what we expect from a model. As in every other context that covers the human psyche, the emotion of trust is difficult and oftentimes ambitious to quantify, a very large amount of factors flow into it and various research tried to measure it in different ways.

Therefore we start in Section 3.1 by giving a first overview of what is trust. We divide trust into cognitive and emotional trust as both types develop quite individual.

Going on with Section 3.2 we present two recently recommended questionnaires used in the AI context to measure trust. TPA is currently commonly used but was, as the name suggests, originally created for measuring trust in automation instead of AI and had to be reevaluated. TXAI however was designed for the context of AI but is rather new and lacks empirical evaluation.

Continuing with section 3.3 we give a first general overview of why trust in AI needs to be measured

Authors' address: Simon Teves-Humme, simon.teveshumme@tu-dortmund.de; Aileen Bücker, aileen.buecker@tu-dortmund.de; Anilcan Polat, anilcan.polat@tu-dortmund.de; Jelle Hüntelmann, jelle.huentelmann@cs.tu-dortmund.de; Prof. Dr. Emanuell Müller.

XXXX-XXXX/2023/11-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

including the need to identify AI model misuse, abuse, and disuse and minimize it. With Artificial Intelligence merging into everyday life and thereby being used by ordinary people, it is a duty to check for suitable usage.

After that we highlight difficulties in measuring trust in Section 3.4 as trust is a subjective and therefore not observable construct. Trust is developed individually by each person as it depends on the person's background, ethnicity, and general attitude toward AI technology. The Negative Attitude Towards Robots scale (NARS) can be a good indicator of what attitude to expect, to be able to include in the model design. Furthermore, people even have different expectations originating from varying system appearances, which illustrates just one more dimension compared to the common expected dimensions of reliability and transparency.

Section 3.5 explores the key cognitive aspects that influence the development of trust in human-AI interactions, emphasizing the earlier implied dimensions such as Tangibility, Reliability, Task Characteristics, and Immediacy Behaviors. Findings for each aspect are divided by robotic, virtual, and embedded AI as they show varying effects for each appearance. These findings can be considered as a guide for AI design to achieve optimal user experience and alignment with intended use.

The final Section 3.6 is similar to the above section but explores the key emotional aspects that influence the development of trust. As mentioned before cognitive and emotional trust operate quite differently. This section therefore emphasizes the dimensions of Tangibility, Anthropomorphism, and Immediacy Behaviors. While Anthropomorphism introduces a whole new dimension concerning the human likeness of the model, Tangibility and Immediacy Behaviors are reevaluated as well as hold different effects compared to cognitive trust.

2 RELATED WORK

Describing a different approach in their publication "Monitoring Model Deterioration with Explainable Uncertainty Estimation via Non-parametric Bootstrap" [5], Carlos Mougan and Dan Saattrup Nielsen introduce methods for monitoring running machine learning models without labeled training data, trying to establish a numeric value for the uncertainty of the correctness given an output. Intending to give the end user explainable metrics, the goal is to give a better sense of trust in a given machine learning system. In their paper, they put emphasis on the connection of **difference in distribution shift and deterioration of a model**. Here, the difference in distribution shift is explained as the change of distribution of differences between what the model gives us versus the actual underlying values. To give an example of this phenomenon, one could picture a machine learning trained on economic data, for example, to predict the movement of stock prices over time. When the economy now undergoes a more or less substantial change, the accuracy and reliability of its predictions differ now that the model is not accurately trained on the actual circumstances that it is presented with. This might especially cause harm in a scenario where the end user still trusts the model due to prior successful predictions, while the trustworthiness has declined unnoticeably. For the evaluation of a model deterioration (meaning its decline of performance over a frame of time) they describe a method called non-parametric bootstrapped uncertainty estimates. Non-parametric means that no assumptions about the basic data distributions are made and bootstrapping means the frequent sampling with replacement of the underlying data to calculate properties of this data. Furthermore, they use "SHAP" values, short for "Shapely Additive Explanations", a method in the field of explainable artificial intelligence giving a numeric value to a certain feature which describes its overall participation in the final prediction that is given out. Deviating further from existing methods, Mougan and Nielsen perform these estimations in timed intervals instead of estimating at a point in time, the intervals being produced using bootstrapping with theoretical guarantees. This basically means sampling over and over as stated before, creating multiple data sets in the process out of which the intervals are calculated. Theoretical guarantees are certain statistical characteristics that are always upheld, leading to a general out-performance of existing models. The main emphasis of this report is on why trust needs to be measured, in the following psychological factors will be examined more closely, therefore the main emphasis of this publication is not highly relevant. The aspect of why the difference in distribution shift causes deterioration of a model should still be kept in mind, as it is a risk when dealing with machine learning models.

3 METHODS

3.1 What is trust?

A main differentiation is between **cognitive** and **emotional** trust. Cognitive trust describes the conscious, logical evaluation of competence, reliability, and consistency of a given person in our machine learning model. This evaluation happens in a relatively objective manner. Emotional trust on the other hand is formed on a more subconscious basis, involving factors like character traits, ethics, personal beliefs, or past experience, together forming almost like a mental predisposition towards the opposite. Most of the time the decision-making process and interpersonal interactions are influenced by both dimensions of trust simultaneously, with a lot of factors not being covered in the extent of this report.

With the prior points in mind, one should ask oneself the question of the right quantity of trust given to an arbitrary receiver. If the trust in a system is low from the beginning and the user questions the credibility in any instance, the use of its outputs becomes obsolete from the get-go. If the quality of its output is trusted in any instance on the other hand, the vulnerability in case of a wrong output grows due to it not being critically questioned. It makes sense to search for a medium between the two extremes, which we call calibrated trust in the following.

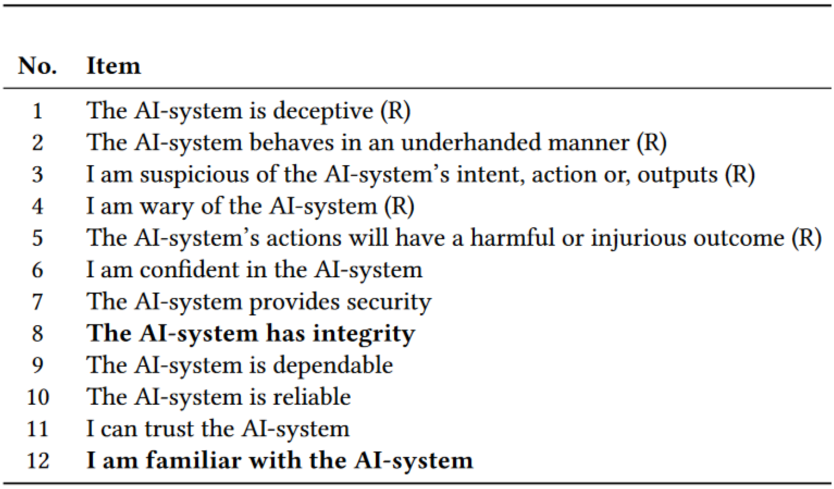
In ideal case there is a linear connection between trustworthiness, meaning the actual competence of an AI system, and the trust given to the system by the user. The trustworthiness can be evaluated in a couple of different ways, a simple example would be the metric of outputs perceived as correct, in contrast to those who are not. More imaginable dimensions would be Explainability, Traceability of decision-making processes, Security, and Robustness, as well as the user experience itself (for example the User Interface or the externalities of the model). This broad scope of dimensions often leaves behind ambiguity when trying to define a metric for trustworthiness.

3.2 How do we measure trust?

When trying to quantify trust in a scientific way, one approach is using questionnaires. Focusing on the **Trust between People and Automation Scale (TPA)** and the **Trust scale for explainable AI (TXAI)**, as they have been recently reevaluated and recommended [6].

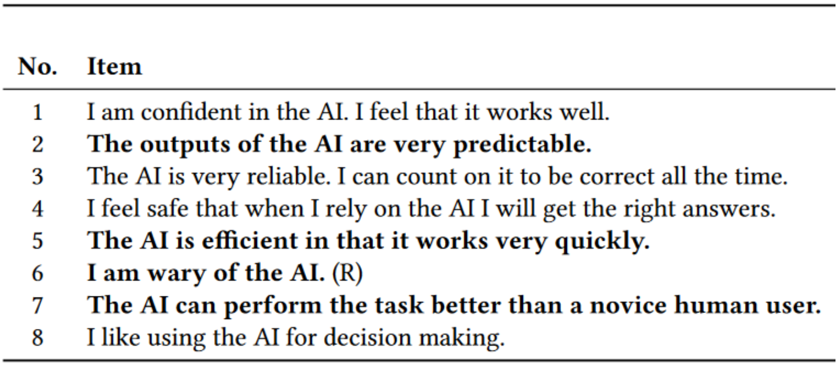
When presented with unobservable psychological constructs like trust, questionnaires are generally used. They provide a subjective estimation of an individual's personal perception and trust level regarding a trustee. Research oftentimes observes differences in a person's behavior towards the trustee and their personal estimated trust level. This type of methodology gives us a numeric value that represents an individual's subjective level of trust, oftentimes by using statistical inference. Most researchers at this current point in time use self-developed questionnaires to measure trust, causing differences in methodology in contrast to other studies, as well as a general lack of comparability. Ideally one would like for a standardized questionnaire to be put in place, which is, in the context of AI, still non-existent. This circumstance causes a reoccurring adaption in each case study, bringing up the question if the methodology still measures what initially was intended [6].

With the TPA being the most commonly used currently, initially it wasn't created for the purpose of evaluating the trustworthiness of an AI-System, rather to examine psychometric qualities in a psychological sense like mentioned before. The authors furthermore introduce a concept known as



No.	Item
1	The AI-system is deceptive (R)
2	The AI-system behaves in an underhanded manner (R)
3	I am suspicious of the AI-system's intent, action or, outputs (R)
4	I am wary of the AI-system (R)
5	The AI-system's actions will have a harmful or injurious outcome (R)
6	I am confident in the AI-system
7	The AI-system provides security
8	**The AI-system has integrity**
9	The AI-system is dependable
10	The AI-system is reliable
11	I can trust the AI-system
12	**I am familiar with the AI-system**

Fig. 1. Trust between People and Automation Scale (TPA). Likert-type scale from 1 to 7 [6]



No.	Item
1	I am confident in the AI. I feel that it works well.
2	**The outputs of the AI are very predictable.**
3	The AI is very reliable. I can count on it to be correct all the time.
4	I feel safe that when I rely on the AI I will get the right answers.
5	**The AI is efficient in that it works very quickly.**
6	**I am wary of the AI.** (R)
7	**The AI can perform the task better than a novice human user.**
8	I like using the AI for decision making.

Fig. 2. Trust scale for explainable AI (TXAI). Likert-type scale from 1 to 5 [6]

the Likert-type scale, a rating scale oftentimes used in surveys or questionnaires where the respondent picks a certain level of (dis-)agreement with a presented statement, for TPA the selectable range is one to seven, for TXAI the range is one to five.

The TPA questionnaire consists of the following 12 metrics: Deceptiveness, underhanded behavior, suspicion of systems intent/action/output, wariness, harmfulness of outcome, confidence in the system, security, integrity, whether the system is dependable, reliability, trust in the system and familiarity with the system. The researchers found that removing the metrics integrity (8) and familiarity (12) provided a more consistent and improved outcome of the survey. Splitting the evaluation into the two factors of trust and distrust proved to produce more validity and reliability, this is the reason why this scale is recommended when there is an interest in exactly this differentiation. It is not entirely clear if trust and distrust are even distinct concepts [6].

The more recently recommended of the two trust evaluation principles (TXAI) has not been yet empirically evaluated due to there being no practical tests to this day. Being based on TPA at its core it presents the following metrics: confidence in the AI system, predictability of outputs, reliability, the feeling of safety when trying to receive the correct answers, efficiency, wariness, the superiority of the AI over the human and the liking of the system for decision making. Similar to the evaluation model looked at before, removing an aspect (point 6: wariness) leads to a more consistent result overall[6].

In general there is a high positive correlation between trust measured by the adapted results of TPA and TXAI, those results are comparable and the authors recommended to make these scales standardized questionnaires[6].

3.3 Why do we measure trust in AI?

The increasing dependence of our society on artificial intelligence (AI) underscores the significance of assessing trust in AI systems. This assessment extends beyond mere technical efficacy, addressing the broader implications of AI in various sectors that are pivotal to human interaction.

The role of trust in AI is especially critical in areas where autonomous decision-making is key. In high-stakes fields such as healthcare and autonomous vehicle navigation, the effectiveness of AI is not solely measured by its technical capabilities, but also by the level of trust it instills in users and stakeholders. This trust is not static; it evolves through ongoing evaluations of the system's ethical alignment, transparency, and performance. As such, the quantification of trust in AI is essential, serving as an indicator of the system's preparedness for integration into society and its acceptance by the public.

Public acceptance of AI technology heavily relies on trust. The extent to which AI technologies are perceived as reliable and ethical plays a significant role in their integration into societal frameworks. Trust assessment transcends its function as a feedback mechanism for AI developers. It becomes integral in ensuring that AI advancements align with societal values and expectations, fostering smoother integration into everyday life.

Trust assessment is also a vital component in the iterative development of AI systems. It captures user perceptions, expectations, and potential misconceptions about AI. Meaning that it is necessary to identify misuse, disuse, and abuse of the AI model. Disuse refers to the case where trust is too low and whereby the actual functionality is not fully used and therefore wasted. Misuse represents the opposite where trust is too high so that the AI model is used for more than it is capable of. This results in a high risk of unwanted events [3]. Lastly, there is abuse which is also an outcome of too low trust where users intentionally try to fool the model. To minimize these behaviors feedback is indispensable for refining AI algorithms, enhancing user interfaces, and ensuring the accessibility and intuitiveness of AI systems. This process enhances the reciprocal relationship between AI systems and their human counterparts, promoting a development approach that is centered around the user.

In the sphere of policy and governance, understanding levels of trust in AI is critical. As AI becomes more ingrained in various aspects of life, trust measurement aids in shaping policies, standards, and regulations to ensure responsible and ethical AI usage. It acts as a guide for regulators and policymakers, helping them craft governance structures that are responsive to the evolving interplay between AI and society. Addressing biases in AI algorithms is another key aspect where trust measurement is indispensable. AI systems, mirroring the biases in their training data, can perpetuate societal inequalities. Measuring trust helps in recognizing these biases, paving the way for the development of AI systems that are ethically sound and equitable.

Moreover, assessing trust in AI provides insights into diverse cultural attitudes towards technology. Trust levels in AI vary across different cultural contexts, shaped by a myriad of historical, social, and ethical factors. Understanding these variations is critical for developing AI technologies that are respectful and adaptable to the diverse cultural fabric of our global society.

In summary, the evaluation of trust in AI encompasses a broad spectrum of considerations including cultural sensitivity, societal acceptance, safety, and ethical alignment. It forms the foundation for the responsible and beneficial evolution of AI technologies, ensuring they align with the evolving standards and values of our global human community.

3.4 Difficulties

Next to the already discussed challenges in measuring a psychological interaction between humans and Artificial Intelligence systems, we face a range of other complications. When dealing with a problem from the standpoint of computer science, we like to have a clear outline or deterministic approach to find a solution. Given that we do not even understand the inner workings of AI models fully in the first place, we do not know if deterministic explanations will ever be possible. Furthermore, when trying to evaluate trust, it becomes apparent that the problem we face is not a purely technical one, adding a psychological dimension to the discussion. Overall we are at a point where the matter is not completely researched.

The development and perception of trust are dependant on the background of each person individually, especially the attitude towards a new, previously unknown technology. What is especially important to note is that the performance of a given AI model cannot be equated to the development of trust towards it, a way more complex review of factors is needed. Often we have a case where trust-developing factors are in direct conflict with each other. For example, a robotic AI that is competent to the extent that it increases cognitive trust, yet erroneous robots are liked more than flawless ones, due to a lowered sense of competition towards it.

To elaborate on this example further it is beneficial to look at the dynamics of **trust trajectories**(Figure 3). Given a robotic AI, the initial trust towards it is relatively low and develops stronger over time, similar to the trust dynamic between humans growing stronger the more (positive) interactions are experienced, the machine intelligence is of great importance in this case. This is potentially caused by the fact that we associate robots stronger with humans than we would virtual or embedded AIs.

Given a virtual AI, meaning an AI that has visual components without being a physical representation, the initial trust might start off high (due to a human-like look for example) and decreases with more interactions, in case the expectations towards it are not met. Oftentimes the visual components outshine the actual competence of the model behind it by design, causing the mentioned discrepancy. In this case, careful calibration is required to have the machine intelligence and its visual representation match.

When the AI is embedded, meaning the system is partially or fully invisible to the end user, the trust trajectory is primarily driven by reliability, while the user may not even be aware that he is interacting with artificial intelligence in the first place. With this form of AI, the trust restoration is very hard [2].

3.5 Cognitive trust aspects

Contrary to some expectations, transparency, and reliability are not the sole influencing dimensions for developing trust. The five main human-AI cognitive trust influencing dimensions are Tangibility, Transparency, Reliability, Task Characteristics, and Immediacy Behaviors [2]. The appearance of the AI model, whether it is robotic, virtual, or embedded interacts with each of the mentioned dimensions and can result in opposite effects for each appearance as presented below. Therefore these aspects should be carefully considered when designing and adapting an AI model.

Tangibility refers to the physical or virtually visible appearance of the Artificial Intelligence system. This dimension has an especially great impact on the first impression and so-called initial trust. Generally, it can be said that a physical presence increases trust more than a virtual presence

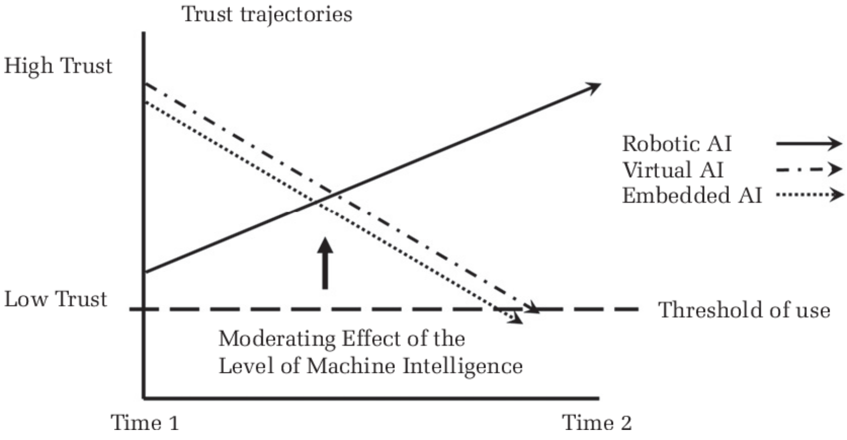


Fig. 3. Trajectories of Trust for Robotic, Virtual, and Embedded AI [2]

and the virtual presence increases trust more than an embedded invisible AI model but there are some conditions to consider as well. The robotic model's appearance has to match its assigned task as concerning itself overwritten digital texts only would have lower perceived trust by the user in its capabilities than a virtual AI model doing the same. When choosing a robotic appearance there is an additional challenge that the movements of the robot have to look smooth and competent to not influence the perceived trust negatively. It showed that human likeliness does not always have the model being perceived as highly intelligent. Concerning embedded AI, the emphasis is on the user's awareness of the model. While the precise impact of user awareness has not been properly researched, it is evident that hiding the model from the user has a significant negative, long-term effect on perceived trust [2].

Transparency involves the clarity of the system's operations, achieved through clear explanations that enable users to comprehend the decision-making process of AI model and gain the ability to know when to rely on the model and when not. This aspect increases the trustworthiness of the model. However, the impact of transparency on perceived trust varies across the AI types. Regarding robotic AI there is not much empirical research on how it influences perceived trust. For virtual AI knowledge of both reliability and algorithm contributes to increased trust. Virtual models often evoke unrealistically high expectations from the user which can be moderated by transparency. In the context of embedded AI, it increases trust and is crucial for highly intelligent systems. Given the inherent difficulty in perceiving embedded AI models, this is highly important as it otherwise leads to users guessing and attempting to manipulate the system. Although there have been cases explaining the limitation of the model's ability to lead users with high self-confidence to abandon the system and rather trust in their own abilities, making this an even more delicate matter for embedded AI systems [2].

Reliability concerns itself with the dependability of the AI model to return correct and consistent results. Reliability is crucial for the interaction-based development of trust. Low reliability decreased trust but timing and situation had a strong influence too. Experiments showed that early low reliability decreased trust more than late drops of reliability and in emergency situations, users tended to follow an AI systems command regardless of its low reliability [2]. Users having experience with low-reliability AI models showed higher trust than users without experience. Noteworthy is that medium reliability systems seemed to confuse the users as they lowered trust way more than low reliability. Regarding robotic AI a serious point is that low reliability did not decrease

trust when the model was perceived as highly intelligent once again calling the importance of good transparency into the foreground. For virtual AI low reliability especially decreased when the initial trust is high, reflecting the principle illustrated in figure 3. Concerning embedded AI low reliability strongly decreases trust and keeps being especially hard to restore for this AI model type.
Task Characteristics refer to the compatibility between the AI model's capabilities and appearance and the task it is supposed to perform. This has to align with the user's expectations for the AI system. This dimension affects various AI models similarly. For technical and data analysis tasks users tend to rely on AI models as much as or even more on their human counterparts. It is the other way around for social tasks. This behavioral trend aligns with the HABA-MABA research (Humans Are Better At/Machines Are Better At) [1]. In history, machines demonstrated superiority in technical tasks as well. However, as AI systems are rapidly evolving and getting more intelligent, there is a need for societal reevaluation of the possibility of AI surpassing humans in specific social tasks. Experimental integration of AI models as team members in group work revealed that participants valued human team members more than AI members. Nevertheless, AI was trusted to do management tasks and could even ease and boost communication inside the team as it could intervene during unprofessional communication. It was also found that when an AI model acted more emotionally it not only reduced tension but actually accelerated trust development [2].
Immediacy Behaviors involves the responsiveness of the AI model as well as its used gestures for active and passive communication. Regarding robotic AI high responsiveness, adaptability, and prosocial behavior were perceived to highly increase trust but also seemed to be expected for high machine intelligence resulting in a fast decrease of trust when the AI system could not meet those expectations [2]. Social gestures such as nodding while agreeing to something create sympathy but timing and fit to the specific situation are important similar to human-human interaction. For virtual AI personalising the model to the user and individual persuasion tactics were perceived as especially helpful for trust development. This was observed for embedded AI as well. Nevertheless, constant tracking and therefore a perfect short-term memory was perceived as uncomfortable and may decrease trust.
These findings are a summary of the research of two decades [2] and act as a general orientation for designing and adapting an AI model although it got clear that these dimension's effects are entangled in a complex way and therefore the exact effects are hard to predict. This again gives an understanding of how important it is to constantly measure the actual user experience and trust in the designed model. This proves the design decisions to be advantageous or not and finally reveals the consistency of the intended and actual use of the AI system.

3.6 Emotional trust aspects

Emotional trust is influenced differently from cognitive trust as this happens more subconsciously. Tangibility and Immediacy Behaviors were found to have a strong impact on emotional trust as well with Anthropomorphism emerging as a new influencing aspect [2]. Given that trust is composed of both cognitive and emotional trust careful consideration of these elements is important for designing or adapting an AI model especially when user behavior appears irrational, insights into these aspects can offer explanations. Emotional trust is especially hard to manage as the cultural background, the person's beliefs regarding functionality, and the general attitude towards robots and technology make each person react individually.
Beginning again with **Tangibility** the reactions for each AI model are quite different. Robotic AI is walking a delicate path between increasing liking or inducing fear through physical presence resulting in an increase or decrease of trust. Many people, especially people with negative attitudes towards robots are easily intimidated by a smart robot. For virtual AI the aspect of fear is not as dramatic, especially for older people the presence of a persona can increase liking and trust. It can

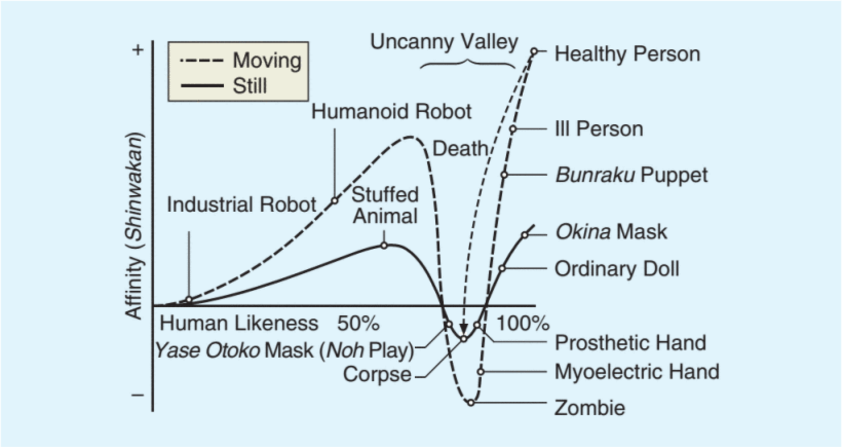


Fig. 4. Proposed relation between affinity and human likeness [4]

also trigger the release of oxytocin, often referred to as "the love hormone" creating a feeling of safeness and security and therefore increasing trust. Regarding embedded AI the presence of a persona is no option resulting in users tending to search for a representative trustee. User trust in this AI model is therefore correlated to the reputation of the developer or the technology. With the absence of a presence, there is a general lower perception of authenticity and ethically as well [2]. For **Immediacy Behaviors** there is a general liking of human-like behavior which highly increases emotional trust. Still, interestingly flawless models were liked less than erroneous models as erroneous models tend to lower the user's sense of competition. In an experiment, even an obviously cheating robot was liked more than an honest one. Users showed more positive emotions toward an agent that did not behave completely realistic [2]. The Uncanny Valley theory, as illustrated in Figure 4, provides the most comprehensive explanation for this phenomenon. First discovered by the robotics professor Masahiro Mori in 1970, the theory challenges the expectation of a consistently growing affinity by increasing human likeness. Contrary to this there is a drop in affinity when reaching near-perfect resemblance, only to rise again when reaching the perfect human resemblance with all of its imperfections. The so-called Uncanny Valley. Humans are feeling uncomfortable about uncannily realistic human appearances. A theory is that this behavior is a result of the evolutionary instinct of self-preservation [4].

Finally **Anthropomorphism** which refers to the human likeness of the AI system. For robotic AI this relates to the Uncanny Valley phenomenon as human likeness can increase positive emotions but can cause discomfort too. Some users experience thoughts of mortality when exposed to a human-like dead thing. Nevertheless, human-like robots were less blamed for errors. Regarding virtual AI human likeness strongly increases initial trust but again creates high expectations regarding the AI model's abilities. Notably attributes such as attractiveness, ethnicity, and facial similarity to the user accelerate trust development. This especially was visible in an experiment, where participants were supposed to ride in a driving simulator operated by a virtual driver that looked either similar or different from them. The Results showed that the participants with a similar-looking driver were more likely to trust it in varying situations [7]. In addition, the appearance of the AI and the task it is supposed to perform have to be calibrated to align with the user's expectations. Instances of poor calibration result in distrust and can lead to abandonment of the model. It is noteworthy that these behaviors were more evident in objective observations than in

subjective self-assessment. An intriguing revelation in human-AI interaction is that users tend to disclose more sensitive information when conversing with an AI model compared to another human, possibly as a result of a lowered sense of judgment [2]. This has to be considered regarding data security issues.
Emotional trust is even more challenging to predict compared to cognitive trust as it operates at a rather subconscious level. Though it is only slightly significant for embedded AI, it is important to consider it for robotic and virtual AI since it brings insights into the user affection and overall comfort with AI interactions. Therefore constantly measuring the actual trust and comparing the results to the predicted value by design is mandatory for running a human-AI trust calibrated AI model.

4 DISCUSSION

In the context of the previously mentioned methods, the question arises as to which extent it makes sense to use certain methods. Starting with the methodology of questionnaires in general, which first allows a view into the subjective perspectives of the questioned person, allowing them to quantify what trust means to them specifically. It also allows the user of said method to observe differences and nuances in their relationship with AI and gives a numerical value in connection with different factors influencing trust that can be compared from person to person. The downside to this form of question-based survey is the unavoidable biases and limitations that researchers face when designing a questionnaire that tries to measure an unobservable construct which is trust. Given that the concept has to be adapted to the different cases individually, standardization of such survey methods becomes nearly impossible as there is no general consensus between researchers on how this methodology should be applied. This does not strictly need to be a negative aspect, as researchers can find themselves with higher flexibility, allowing a better fit of the questions to the specific properties of a given AI system. What also has to be kept in mind in this matter is the general lack of agreed-upon validity and reliability of results that is caused by this style of research. **Ideally one would wish for a way to have a standardized questionnaire, which requires a lot of reciprocal review and cooperation between researchers to put this in place someday.**
To discuss the scales TPA and TXAI in particular, we start chronologically with TPA, which provides a rather established method as it is the most commonly used one to this day. Using the Likert-type scale, it provides a fairly nuanced space of values for each question between one and seven. This causes a generally more granular scale, allowing for more subtlety in evaluating with a questionnaire, as well as giving the person analyzing an easier time trying to look for patterns. As mentioned before, we have a case where a concept that was not initially designed for the purpose of analyzing AI was adopted for this purpose, bringing up the question of whether we have a case where it is truly applicable. The metrics integrity and familiarity as listed in this type of questionnaire are also not directly applicable to AI, rather in a figurative sense. It is also important to mention that when trying to formalize a concept that in itself is not clearly defined, there is no guarantee that a congruent interpretation will be reached on all fronts. **As the understanding of the world around us grows, the scientific findings should be applied over and over to the measurements of trust we perform, keeping the mental picture up to date as the consensus changes around it.**
TXAI on the other hand offers us the advantage that it picks up TPA and improves it in certain aspects, through the introduction of metrics like predictability of outputs and efficiency, which is generally more suited towards handling evaluation of AI, bringing a high relevancy to the discussion. This also causes a positive correlation with the results that TPA offers, for example allowing researchers that apply TXAI to fall back on material that uses TPA and vice versa, making links between the models and creating rather standardized results. This circumstance is not necessarily

a positive consequence though, limiting TXAI's scientific potential to create new unique insights, as possible biases or limits of the underlying TPA concept might be carried over when using the continued methodology that is TXAI.
Does it make sense that TXAI removes metrics for consistency (in our case wariness)? While achieving a general simplification of the measurement metric at hand, stabilizing results and strengthening the reliability of its output, a dimension in the evaluation is lost nonetheless, potentially destroying key insights gained from said metric, as well as not portraying the psychological concept of trust in the true complexity that it would need to be portrayed in. **Future research could extend questionnaires in certain aspects or rather shorten them by a few aspects while maintaining sensible insights into the concept of trust.**

5 CONCLUSIONS

Summing up the prior findings, examples, and considerations it is apparent that all technological factors involving AI, as well as all psychological factors involving the concept of trust, have to be considered with concern to their individual weight in regards to the overall conclusion. When trying to evaluate concepts underlying the human psyche or interpersonal interactions between humans we oftentimes find extremely complicated and convoluted connections that even modern science does not understand to this day. Furthermore, we have a situation where AI as a technology evolves constantly, maybe even at an exponential rate in the near future, where a balance between innovation and ethical considerations will most likely be essential to maintain a trustworthy AI ecosystem. The key takeaway from this discussion is that we as the end users have to constantly keep up the cycle of measuring user trust in the system, as well as redesigning the system in such a way that its properties match the desired interactions and results from it and up-keeping the essential indicators that classify it as trustworthy.

REFERENCES

[1] Jeffrey M Bradshaw, Paul Feltovich, and Matthew Johnson. 2017. Human-agent interaction. *Handbook of human-machine interaction* (2017), 283–302.

[2] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, 2 (2020), 627–660.

[3] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 624–635.

[4] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & automation magazine* 19, 2 (2012), 98–100.

[5] Carlos Mougan and Dan Saattrup Nielsen. 2023. Monitoring model deterioration with explainable uncertainty estimation via non-parametric bootstrap. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 15037–15045.

[6] Sebastian AC Perrig, Nicolas Scharowski, and Florian Brühlmann. 2023. Trust issues with trust scales: examining the psychometric quality of trust measures in the context of AI. In *Extended abstracts of the 2023 CHI Conference on human factors in computing systems*. 1–7.

[7] Frank MF Verberne, Jaap Ham, and Cees JH Midden. 2015. Trusting a virtual driver that looks, acts, and thinks like you. *Human factors* 57, 5 (2015), 895–909.