Trust in AI: Transparency, and Uncertainty Reduction. Development of a new theoretical framework

Letizia Aquilino[1,2], Piercosma Bisconti[2] and Antonella Marchetti[1]

[1] *Università Cattolica del Sacro Cuore, Largo Gemelli 1, Milan, 20123, Italy*
[2] *DEXAI – Etica Artificiale, Rome, Italy*

Abstract

Trust plays a pivotal role in the acceptance of AI (Artificial Intelligence), particularly when it involves people's health and safety. AI systems have proven to hold great potential when applied to the medical field. However, users still find it challenging to trust AI over a human doctor for decisions regarding their health. This paper establishes a new theoretical framework, drawing upon the integration of the Uncertainty Reduction Theory (URT) and the theorization on agency locus. This framework aims to examine the influence of transparency, agency locus, and human oversight, mediated by uncertainty reduction, on trust development. Transparency has already been revealed as a key element in fostering trust, as AI systems showing some kind of transparency, providing insights into their inner workings, are generally perceived as more trustworthy. One explanation for this can pertain to the system becoming more understandable and predictable to the user, which reduces the uncertainty of the interaction. The framework also focuses on the differences entailed by the application in different fields, namely healthcare and first response intervention. Moreover, the paper foresees multiple experiments that will validate this model, shedding light on the complex dynamics of trust in AI.

Keywords

Artificial Agents, Perceived Trustworthiness, Transparency, Trust, Uncertainty

1. Introduction

Trust is a key component in the acceptance of artificial intelligence (AI). Considering the numerous benefits that can be unlocked by relying on AI applications [1, 2], and their growing spread in people's lives, understanding how trust in AI is built is a crucial matter. In medical settings, patients have shown to be motivated to participate in treatment and to be satisfied with the service when they trust the medical agent [3–7]. This becomes particularly significant when the decision-making process involves AI. Studies show that people are reluctant to trust AI technology for medical treatment [8–10] and still prefer to receive services from human doctors, even when AI demonstrates equal or superior performance in prevention, diagnosis and treatment [8, 9]. This unwillingness to rely on AI can be due to its novelty. Drawing from literature on the development of relationships between humans, the Uncertainty Reduction Theory (URT) [11] can help better understand the evolution of trust in AI. The URT explains how people are inherently motivated to reduce the uncertainty that initially characterizes relationships, in order to be able to better predict or explain others' behaviour. Furthermore, the theory states that uncertainty reduction builds on effectance motivation, a basic human motivation to be a competent agent in one's environment [12]. The process of trust formation is highly linked to uncertainty reduction, as it has been found that predictive and explanatory knowledge about automated systems enhances user's trust [13]. Moreover, making AI systems transparent to users can be a fundamental tool to reduce uncertainty and enhance trust [14, 15]: the more information the person has about the other agent, the more they will feel able to predict its future actions, being given information about its inner workings. Besides transparency, the degree of

MultiTTrust: 2nd Workshop on Multidisciplinary Perspectives on Human-AI Team, Dec 04, 2023, Gothenburg, Sweden

Letizia.aquilino1@unicatt.it (L. Aquilino); piercosma.bisconti@dexai.eu (P. Bisconti); antonella.marchetti@unicatt.it (A. Marchetti)

0009-0001-4862-5910 (L. Aquilino); 0000-0001-8052-0142 (P. Bisconti); 0000-0001-9985-0539 (A. Marchetti)

uncertainty (and consequently of trust) can depend also on the assumptions the person makes about the system's functioning, based on the information they are given. In fact, people can perceive the way the system "thinks" as more or less similar to the way humans do. The perception of machines as human-like has been found to be helpful in reducing uncertainty as it facilitates simulation of human intelligence, a type more familiar to users [14, 16, 17]. The machine's rules or the cause of its apparent agency have been described as agency locus, which can be external as created by humans, or internal, generated by the machine. This distinction has become relevant with the development of AI technologies and the massive use of machine learning techniques, which are substantially different from systems programmed by humans, because they have the capacity to define or modify decision-making rules autonomously [18]. For the purposes of this paper, agency locus will be referred to as something made explicit to the user, so that actual and perceived agency locus necessarily coincide.

After conducting an in-depth literature review, a new theoretical framework has been developed [See Appendix 1]. Said framework is intended to fully capture the dynamics of trust development in AI systems, by including characteristics of both the human agent and the AI system. Keeping an eye on both parts of the interaction is an element of novelty within the human-machine interaction research. The model aims to provide information about how to tailor the system's transparency to accommodate the person's needs, based on their way of thinking and working. Furthermore, its objective is to understand how to reach the optimal level of uncertainty by managing human oversight, both as presence of a human mediator during the interaction and in terms of agency locus. The framework is set to be tested within multiple studies.

2. Research Questions and Hypotheses

As previously mentioned, the development of trust in AI systems is a complex process, highly influenced by uncertainty reduction, which in turn is affected by AI's transparency and agency locus. In this paper we set out the fundamental theoretical framework for future experiments. The objective of the experiments will be to understand whether trust in a specific AI-decision-making system is determined by its transparency, agency locus (external/programmed by humans or internal/machine learning), and by the presence of human oversight, with perceived uncertainty as mediating variable. Additionally, the studies will highlight the effects of the salience the AI-made decisions have on the human agent's health, by investigating the development of trust in two different fields of application: medical and first response. In the first case, AI decisions have a direct impact on the user's welfare, as they can regard diagnosis or treatment plans; in the second, the information provided has more influence on the health and safety of the person being rescued rather than the first responder's. Therefore, the following research questions can be posed: (RQ1) Is trust in an AI-decision-making system determined by the level of transparency shown and by the type of information given about it? (RQ2) Is trust in an AI-decision-making system determined by its agency locus? (RQ3) Is trust in an AI-decision-making system determined by the presence of human oversight? (RQ4) Do perceived uncertainty mediate the effect of transparency, agency locus, and human oversight on trust?

These questions are crucial, as the level of trust in an AI-decision-making system could be easily optimized by modifying information given about its way of working and by the level of human intervention in the decision-making process.

We hypothesize that: (H1) Users' trust in the system will increase if given an intermediate level of information [19], making the system transparent enough to reduce uncertainty, without excessively increasing the user's cognitive load; (H2) users will place more trust in the system if it is programmed by a human [20]; (H3) users will trust the system more if it operates under human oversight; (H4) users will trust the system more easily when their personal health and safety is not directly affectable by the system's decisions.

The theoretical framework also includes other variables pertaining to the user's characteristics and attitudes, which need to be explored in additional experiments. These are hypothesized to act as moderating variables (personality traits, propensity to trust, and locus of

control) or to be other variables that could have a direct influence on trust in the AI system (trust in the human intermediary and job demands).

3. Study Design

The research is structured in two separate studies: the first sees the employment of medical AI, whilst the second regards an AI system used as aid to first responders in rescue missions. In the first study participants will take part in a simulation that sees them as patients receiving a diagnosis and a treatment plan from an AI-system; in the second study, participants will be involved in a simulation as first responders using an AI system that advises them about the conditions of the environment they need to explore and of the people that need rescuing. Recruitment for study 2 will target people that are currently or have been employed as first responders, as already familiar with rescue situations and gear. On the other hand, in study 1 participants will be selected to create a sample of 50% AI experts and 50% laypeople with no specific experience in the field. This distinction will be useful to identify possible differences in the optimal level of transparency based on the level of expertise. Both studies follow a 3x2x2 between-subjects design, with three independent variables: level of transparency (low, intermediate, high), agency locus (external, internal), and human oversight (present, absent).

3.1. Independent variables

Transparency. During the interaction, the system can give information about its working to the user. More specifically, the information will be about: the process through which the output has been created, explaining the source and the technologies used to collect data; and the level of accuracy of the output, expressed in percentage (e.g., "This system has a success rate of 74% in diagnosing correctly). The quantity of information given to the users will vary, defining the level of transparency, with each condition defined by a different percentage of information made available: 100% (high level), 50% (intermediate level), 25% (low level).

Agency locus. Agency locus can be defined as the system's rules, or the cause of its apparent agency, which can be external as created by humans, or internal, generated by the machine. The users will be informed that the system's agency locus is either external (programmed by humans) or internal (working with machine learning).

Human oversight. The system's output is either communicated directly by the system itself, without the mediation of a person who could add to it based on their experience and judgment, or communicated by a human agent, who acts as filter to the system's output. In study 1 the information provided by the AI regarding diagnosis and treatment can either be communicated by the medical doctor or directly by the AI system. In study 2, AI's output can be provided to the first responder directly or filtered by the operation manager.

3.2. Mediating variables

Perceived uncertainty. Perceived uncertainty is to be measured, as in [14], using 6 items measured on a 7-point-Likert scale from [21] [See Appendix 2].

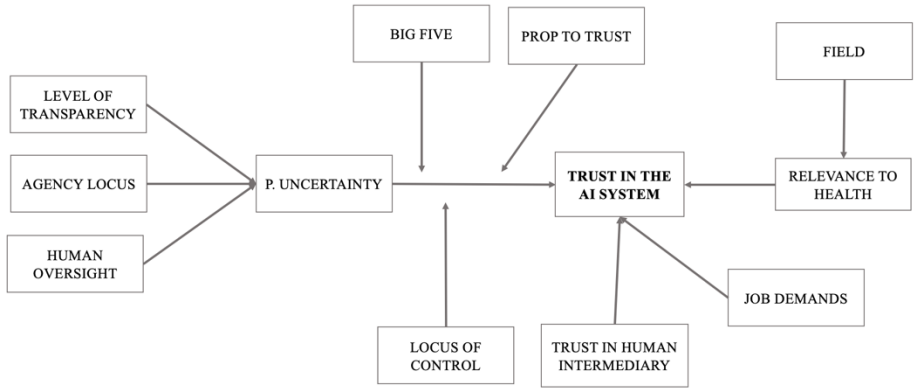
3.3. Dependent variables

Trust in the AI system. Trust in the AI system is measured with 8 items adapted from [22] by [23] as: willingness to depend (3 items) and Subjective probability of depending (5 items) [See Appendix 2].

4. Conclusion and Future Research

The current paper intends to be a prelude for future research with the aim of testing a new theoretical model, while verifying the role of transparency, agency locus, and human oversight in the trust-development process. It also represents a contribution to shedding light on the specifics of trust development based on the field of application. Previous research has investigated which variables influence trust in AI [24, 25]. Nevertheless, there is still need to consider the full complexity of the process, as research usually focuses either on the characteristics of the human agent [26, 27] or on the way AI systems work and present their functioning logics [28, 29]. The two categories are highly intertwined, and both contribute to the relational evolution of trust. Therefore, the user and the system need to be considered as equally important agents, and their characteristics as equally impactful on interaction quality. Moreover, this framework will also contribute to reach a better understanding of the application of the Uncertainty Reduction Theory framework to the adoption of AI, with the aim of providing information on how to optimize the encounter between users and AI. The framework presented is arguably complex and could not be tested in a single trial, due to resource limitations. Therefore, a series of experiments need to be planned to fully explore all the variables. A first distinction, as already stated, can be made according to the field of application of AI: medical AI and first response, which allows to appreciate the effect of salience of health and safety on trusting beliefs and intentions, as they are expected to vary depending on the consequences of AI decisions. Additionally, variables pertaining to the user's characteristics need to be fully explored, to achieve further knowledge about the adjustments of the level of transparency and human intervention (agency locus and human oversight) needed to make the AI system more trustworthy to the eyes of the user.

Appendix 1: Theoretical framework



Appendix 2: Questionnaires*

Perceived uncertainty (from [21])
How confident were you in predicting the system's judgments?
How confident were you in predicting the results of the system given an input?
How confident were you in explaining why the system gives certain output given its input
How well did you feel you know this system?
How well did you feel you know how the system acts?
How certain were you about what this system is really like?

Trust in the AI system ([23]'s adaptation from [22])
Willingness to depend
1. When an important issue or problem arises, I would feel comfortable depending on the information provided by the AI system.
2. I can always rely on the AI system in a tough situation.
3. I feel that I could count on the AI system to help with a crucial problem.
Subjective probability of depending
1. If I had a challenging legal problem, I would want to use the AI system again.
2. I would feel comfortable acting on the information given to me by the AI system.
3. I would not hesitate to use the information the AI system supplied
4. I would confidently act on the advice I was given by the AI system
5. I would feel secure in using the information from the AI system

References

[1] McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020; 577: 89–94.

[2] De Carvalho TM, Noels E, Wakkee M, et al. Development of Smartphone Apps for Skin Cancer Risk Assessment: Progress and Promise. *JMIR Dermatol* 2019; 2: e13376.

[3] Baker R, Mainous Iii AG, Gray DP, et al. Exploration of the relationship between continuity, trust in regular doctors and patient satisfaction with consultations with family doctors. *Scandinavian Journal of Primary Health Care* 2003; 21: 27–32.

[4] Birkhäuer J, Gaab J, Kossowsky J, et al. Trust in the health care professional and health outcome: A meta-analysis. *PLoS ONE* 2017; 12: e0170988.

[5] Gill L, Cassia F, Cameron ID, et al. Exploring Client Adherence Factors Related to Clinical Outcomes. *Australasian Marketing Journal* 2014; 22: 197–204.

[6] Hillen MA, De Haes HCJM, Smets EMA. Cancer patients' trust in their physician-a review. *Psycho-Oncology* 2011; 20: 227–241.

[7] Wu Y-H, Cristancho-Lacroix V, Fassert C, et al. The Attitudes and Perceptions of Older Adults With Mild Cognitive Impairment Toward an Assistive Robot. *J Appl Gerontol* 2016; 35: 3–17.

[8] Bigman YE, Gray K. People are averse to machines making moral decisions. *Cognition* 2018; 181: 21–34.

[9] Longoni C, Bonezzi A, Morewedge CK. Resistance to Medical Artificial Intelligence. *Journal of Consumer Research* 2019; 46: 629–650.

[10] Promberger M, Baron J. Do patients trust computers? *J Behav Decis Making* 2006; 19: 455–468.

[11] Berger CR, Calabrese RJ. SOME EXPLORATIONS IN INITIAL INTERACTION AND BEYOND: TOWARD A DEVELOPMENTAL THEORY OF INTERPERSONAL COMMUNICATION. *Human Comm Res* 1975; 1: 99–112.

[12] White RW. Motivation reconsidered: The concept of competence. *Psychological Review* 1959; 66: 297–333.

[13] Jian J-Y, Bisantz AM, Drury CG. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 2000; 4: 53–71.

[14] Liu B. In AI We Trust? Effects of Agency Locus and Transparency on Uncertainty Reduction in Human-AI Interaction. *Journal of Computer-Mediated Communication* 2021; 26: 384–402.

[15] Shin D, Park YJ. Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior* 2019; 98: 277–284.

[16] Fox J, Ahn SJ (Grace), Janssen JH, et al. Avatars Versus Agents: A Meta-Analysis Quantifying the Effect of Agency on Social Influence. *Human–Computer Interaction* 2015; 30: 401–432.

[17] Oh CS, Bailenson JN, Welch GF. A Systematic Review of Social Presence: Definition, Antecedents, and Implications. *Front Robot AI* 2018; 5: 114.

[18] Mittelstadt BD, Allo P, Taddeo M, et al. The ethics of algorithms: Mapping the debate. *Big Data & Society* 2016; 3: 205395171667967.

[19] Kizilcec RF. How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. San Jose California USA: ACM, pp. 2390–2395.

[20] Liu B, Wei L. Machine gaze in online behavioral targeting: The effects of algorithmic human likeness on social presence and social influence. *Computers in Human Behavior* 2021; 124: 106926.

[21] Clatterbuck GW. ATTRIBUTIONAL CONFIDENCE AND UNCERTAINTY IN INITIAL INTERACTION. *Human Comm Res* 1979; 5: 147–157.

[22] McKnight DH, Choudhury V, Kacmar C. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research* 2002; 13: 334–359.

[23] Zarifis A, Kawalek P, Azadegan A. Evaluating If Trust and Personal Information Privacy Concerns Are Barriers to Using Health Insurance That Explicitly Utilizes AI. *Journal of Internet Commerce* 2021; 20: 66–83.
[24] Xiang H, Zhou J, Xie B. AI tools for debunking online spam reviews? Trust of younger and older adults in AI detection criteria. *Behaviour and Information Technology*. Epub ahead of print 2022. DOI: 10.1080/0144929X.2021.2024252.
[25] Kim T, Song H. Communicating the Limitations of AI: The Effect of Message Framing and Ownership on Trust in Artificial Intelligence. *International Journal of Human-Computer Interaction*. Epub ahead of print 2022. DOI: 10.1080/10447318.2022.2049134.
[26] Kandoth S, Shekhar SK. Social influence and intention to use AI: the role of personal innovativeness and perceived trust using the parallel mediation model. *Forum Scientiae Oeconomia* 2022; 10: 131–150.
[27] Huo W, Zheng G, Yan J, et al. Interacting with medical artificial intelligence: Integrating self-responsibility attribution, human–computer trust, and personality. *Computers in Human Behavior*; 132. Epub ahead of print 2022. DOI: 10.1016/j.chb.2022.107253.
[28] Shin D. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human Computer Studies*; 146. Epub ahead of print 2021. DOI: 10.1016/j.ijhcs.2020.102551.
[29] Yu L, Li Y. Artificial Intelligence Decision-Making Transparency and Employees' Trust: The Parallel Multiple Mediating Effect of Effectiveness and Discomfort. *Behavioral Sciences*; 12. Epub ahead of print 2022. DOI: 10.3390/bs12050127.