Toward a Human-Centered Metric for Evaluating Trust in Artificial Intelligence Systems

Andrea Esposito[1],*, Giuseppe Desolda[1] and Rosa Lanzilotti[1]

[1] *Department of Computer Science, University of Bari Aldo Moro, Via E. Orabona 4, 70125 Bari, Italy*

Abstract

Trust is a critical human factor in the adoption and effectiveness of artificial intelligence (AI) systems, especially in high-risk domains such as medicine and cybersecurity. This paper introduces a novel probabilistic metric to objectively quantify trust in AI systems, using the degree of reliance on AI by human users as a proxy. The proposed metric leverages changes in human decision-making influenced by AI decisions and can be computed using a flexible data-driven approach. To validate this metric, we present user study protocols applicable to diverse domains, providing examples in the medical field (for Alzheimer's disease detection) and cybersecurity (for phishing detection). These protocols employ a frequentist definition of probability to estimate and compute trust, allowing for a comparison with trust questionnaires to establish correlations. This work aims to contribute a standardized, interpretable, and adaptable method for evaluating trust in human-centered AI systems.

Keywords

Human-Centered Artificial Intelligence, Trustworthiness, Reliability, Human Reliance on AI, Probability

1. Introduction

In recent years, Artificial Intelligence (AI) has become a driving force for both innovation and economy, with applications spanning from low-risk domains, like image generation [1], to high-risk ones, like medicine [2, 3]. Transparency in the AI decision-making process might be critical, especially in high-risk scenarios [4]. Therefore, a lot of research is now focused on eXplainable Artificial Intelligence (XAI), aiming at making AI model decision explainable and interpretable [5]. However, especially in high-risk scenarios, major concerns still linger, including fear of bias and discrimination, or fear of being replaced by machines.

In this context, a novel research area is gaining traction in the academic community: Human-Centered Artificial Intelligence (HCAI). HCAI proposes the adoption of human-centered methodologies throughout the design and development of AI models [6, 7]. HCAI emphasizes creating AI systems by actively involving users in their design, development, and evaluation, aiming to enhance human performances and satisfaction in specified tasks [8]. HCAI systems aim to be useful and usable for specified users, who might be described through a formal model, to help them achieve their specified goals within their context of use while being reliable, safe to use, and *trustworthy* [8, 9].

Focusing on the latter property, trustworthiness is thus a fundamental property of HCAI systems. Trustworthiness is an inherent attribute of the system, reflecting its ability to foster trust—which is the user's willingness to rely on the system in a given context [10, 8]. Therefore, *trust* is a critical human factor influencing the adoption of AI systems. A system that lacks trustworthiness can lead to user disengagement, misuse, or rejection, regardless of its technical capabilities. Exploring trust is therefore crucial for the relatively novel discipline of HCAI: *trustworthiness* is generally accepted as an interesting metric for HCAI systems, on par with usability for classical systems [8].

Despite extensive research on AI reliability and performance, there is a lack of standardized methodologies for evaluating trust from a user-centered perspective. The main existing solutions in the literature to measure the trust of an AI system involve using questionnaires that measure the trust

Joint Proceedings of the ACM IUI Workshops 2025, March 24-27, 2025, Cagliari, Italy

*Corresponding author.

andrea.esposito@uniba.it (A. Esposito); giuseppe.desolda@uniba.it (G. Desolda); rosa.lanzilotti@uniba.it (R. Lanzilotti)
0000-0002-9536-3087 (A. Esposito); 0000-0001-9894-2116 (G. Desolda); 0000-0002-2039-8162 (R. Lanzilotti)

perceived by users. Although they may be useful tools, questionnaires collect a user perspective (or perception) that may also be distorted or untrue concerning the user's real trust due to biases (e.g., due to the peak-end rule [11]). Moreover, administering questionnaires requires the execution of usability tests with real users, but usability tests are difficult to perform or replicate over time to see how trust evolves.

This paper proposes a preliminary idea of a metric to measure trust more objectively than questionnaires. This metric models trust in AI as the human reliance on the AI decision-making process. It leverages probabilistic models to quantify trust, providing researchers and practitioners with a scalable and interpretable approach.

This paper is structured as follows. Section 2 reviews related work; Section 3 presents the core contribution of this paper, defining trust in AI systems, with Section 3.1 detailing the proposed trust metric; Section 5 discusses the metric; Section 6 outlining a methodology to validate the metric, providing examples; finally, Section 7 concludes the paper.

2. Related Work

Previous studies have explored trust in human-computer interaction, focusing on reliability, explainability, and ethical alignment [10]. Trust is an abstract and complex concept, making it difficult to define [8]. However, recent research suggests that a context-dependent definition may benefit the understanding of "trust" [10]. An initial definition by Mayer et al. expresses trust as stemming from the dialog between two agents (e.g., the user and AI), defining it as the *willingness of a party (the trustor) to be vulnerable to the actions of another party (the trustee) based on the expectation that the latter will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party* [12].

Other definitions of trust emphasize its context-dependent nature. For instance, trust has been described by Lee and See as the *attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability* [13]. This definition focuses on the relational and situational dimensions of trust, highlighting the importance of the trustor's goals and the inherent risks in relying on another agent. Similarly, trust has also been defined by Yan et al. as *the belief on whether an application could fulfill a task as expected* [14], focusing on the trustor's expectations and the system's perceived capability to meet them.

Metrics for evaluating trust often rely on qualitative measures, such as user surveys or interviews. However, some quantitative approaches are available through questionnaires. Examples of quantitative techniques to assess trust, which are widely used in the context of AI systems evaluations, include the "Trust in Automated System Test (TOAST)" [15], the "Multidimensional Measure of Trust" [16], and the "Trust in Automation (TiA)" [17] questionnaires.

While these methods provide valuable insights into user *perceptions* of trust, they also present several challenges. Mainly, qualitative measures and questionnaire-based quantitative measures depend on self-reported data, which may not always reflect trust-related behaviors due to biases (e.g., peak-end rule [11]) or simply due to inconsistencies in user responses. Furthermore, these methods can be resource-intensive and lack scalability for large-scale or longitudinal studies. More formal approaches, such as the probabilistic metric proposed in this work, aim to address these limitations by quantifying trust in measurable terms. By relying on observable decision-making behaviors (namely, the final outcomes), such methods offer a more objective and scalable framework for evaluating trust in AI systems, *complementing* (not entirely replacing) traditional qualitative techniques.

3. Defining Trust in AI Systems

By analyzing definitions of trust provided in Section 2, some commonalities are clear. All definitions expect some *uncertainty*, or in general, a lower level of *control* over a situation. Similarly, all definitions expect the "trustor" to *rely* on the "trustee," hoping the latter will perform a task as expected.

Therefore, we define trust in AI as a user's comfort in relying on AI, thus employing *reliance as a proxy of trust*. In other words, we define trust in AI as the confidence in the system's ability to perform its intended functions reliably, transparently, and appropriately within a specific context.

3.1. Proposed Trust Metric

We propose a probabilistic metric for trust, denoted as $\mathcal{T} \in [-1, 1]$, defined as the level of reliance (which is used as a proxy) on the AI decision given human and AI decision-making. Specifically, considering a n-ary classification task with labels in $\mathcal{Y} = \{y_1, \dots, y_n\}$:

$$\mathcal{T} = p\left(d_{\mathrm{H}} \mid d_{\mathrm{AI}}\right) - p\left(d_{\mathrm{H}}\right) \tag{1}$$

where:

- $d_{\mathrm{H}} \in \mathcal{Y}$ represents the human decision.
- $d_{\mathrm{AI}} \in \mathcal{Y}$ represents the AI decision.
- $p\left(d_{\mathrm{H}}\right) \in [0, 1]$ is the probability of the human decision independent of the AI.
- $p\left(d_{\mathrm{H}} \mid d_{\mathrm{AI}}\right) \in [0, 1]$ is the probability of the human decision given the AI decision.

In other words, since we define trust as the level of human reliance on AI, this definition captures the change in human decision-making when influenced by the AI. A higher value of $| \, T \, |$ suggests greater reliance on the AI's decision, whether it reinforces the user's initial judgment or causes a significant shift in their decision. This makes the metric sensitive to alignment and divergence between human and AI decisions, potentially offering insights into how and when users integrate AI outputs into their decision-making.

This definition does not aim to be operational. In other words, it intentionally avoids prescribing specific methods for calculating the probabilities involved ($p\left(d_{\mathrm{H}}\right)$ and $p\left(d_{\mathrm{H}} \mid d_{\mathrm{AI}}\right)$). This open-endedness ensures that the metric definition may adapt to different levels of knowledge about the probability distributions. For instance, probabilities may be derived through frequentist approaches, Bayesian inference, or even machine learning models, depending on the context and the available data.

3.1.1. Computing the Metric Using a Frequentist Approach

The simplest way of computing the metric $\mathcal{T}$ assumes a frequentist definition of probability. In this approach, through a controlled user test, human decisions and AI decisions are recorded. Using these data, $\mathcal{T}$ can be computed using Equation (1) by estimating $p\left(d_{\mathrm{H}}\right)$ and $p\left(d_{\mathrm{H}} \mid d_{\mathrm{AI}}\right)$ through observed frequencies and leveraging the definition of conditional probability [18]:

$$\mathcal{T} = p\left(d_{\mathrm{H}} \mid d_{\mathrm{AI}}\right) - p\left(d_{\mathrm{H}}\right) = \frac{p\left(d_{\mathrm{H}} \wedge d_{\mathrm{AI}}\right)}{p\left(d_{\mathrm{AI}}\right)} - p\left(d_{\mathrm{H}}\right) \tag{2}$$

To illustrate how the proposed metric can be computed, consider a controlled user study where participants are asked to decide on a specific topic. Researchers first collect the participants' initial decisions (d_{H}) before exposing them to the AI's outcome. Subsequently, participants are presented with the AI's decision (d_{AI}), potentially accompanied by explanations, and they may be asked to provide a new decision.

Leveraging a frequentist definition of probability, researchers can approximate $p\left(d_{\mathrm{H}}\right)$ by calculating the frequency of each possible human decision $d_{\mathrm{H}} \in \mathcal{Y}$ for each participant. Then, if the users are asked to provide a new decision after their exposure to the AI outcome, researchers can similarly compute $p\left(d_{\mathrm{H}} \mid d_{\mathrm{AI}}\right)$. Otherwise, the frequency of each AI decision $d_{\mathrm{AI}} \in \mathcal{Y}$ can be used to approximate $p\left(d_{\mathrm{AI}}\right)$ (although such probability is often inherent in machine learning models and can thus be easily inferred), while to estimate $p\left(d_{\mathrm{H}} \wedge d_{\mathrm{AI}}\right)$ researchers count the number of instances where both the human decision d_{H} and the AI decision d_{AI} occurred simultaneously. Once these probabilities are obtained, the trust metric $\mathcal{T}$ can be computed using the formula defined in Equation (2).

4. Example: Evaluating Trust in AI-Assisted Diagnosis

To provide a simple example to clearly show how $\mathcal{T}$ can be used to evaluate trust in human-AI interaction, we present a simple and fictitious example in the medical domain, assuming the existence of a generic AI-based diagnosis support tool for pneumonia detection. In this scenario, the AI system receives a chest X-ray and provides a binary classification, flagging the input as either "with pneumonia" (which we denote $\top$) or "no pneumonia" ($\bot$). Therefore, in this case, $\mathcal{Y} = \{\top, \bot\}$. The proposed trust metric $\mathcal{T}$ quantifies changes in physicians' decisions influenced by AI recommendations. A total of 5 participants, is asked to diagnose 40 chest X-ray scans, divided across four experimental conditions:

1. High Accuracy + With Explanations **(HA-WE)**
2. High Accuracy + No Explanations **(HA-NE)**
3. Low Accuracy + With Explanations **(LA-WE)**
4. Low Accuracy + No Explanations **(LA-NE)**

After making their initial diagnosis, they are presented with the AI classification. Finally, participants are asked to reassess their decision and to provide a potentially revised diagnosis.

4.1. Data Collection

Mock data for a subset of 40 images (10 per condition) was generated to illustrate the process. For each condition, the frequencies of d_{H}, and $d_{\text{H}} \mid d_{\text{AI}}$ (which represents the decision after exposure to the AI classification) were recorded, as shown in Table 1.

Table 1
Frequencies of decisions across experimental conditions.

Condition	$d_{\text{H}} = \top$	$d_{\text{H}} = \top \mid d_{\text{AI}}$	Total Images
HA-WE	4	9	10
HA-NE	4	6	10
LA-WE	6	5	10
LA-NE	4	4	10

4.2. Computation of Trust Metric

The trust metric $\mathcal{T}$ was calculated using Equation (1) for each condition:

- **HA-WE:** $p(d_{\text{H}}) = 0.4 \land p(d_{\text{H}} \mid d_{\text{AI}}) = 0.9 \implies \mathcal{T} = 0.5.$
- **HA-NE:** $p(d_{\text{H}}) = 0.4 \land p(d_{\text{H}} \mid d_{\text{AI}}) = 0.6 \implies \mathcal{T} = 0.2.$
- **LA-WE:** $p(d_{\text{H}}) = 0.6 \land p(d_{\text{H}} \mid d_{\text{AI}}) = 0.5 \implies \mathcal{T} = -0.1.$
- **LA-NE:** $p(d_{\text{H}}) = 0.4 \land p(d_{\text{H}} \mid d_{\text{AI}}) = 0.4 \implies \mathcal{T} = 0.0.$

4.3. Results and Discussion

The results, alongside a possible interpretation, are summarized in Table 2.

Leveraging the way the experimental conditions were defined, it is possible to speculate on some possible discussions. Examples of possible outcomes that can be derived from an analysis like the above are:

- **Impact of Explanations:** Explanations increased reliance on AI in both accuracy conditions, highlighting their importance in fostering trust. However, over-reliance in low-accuracy scenarios suggests the potential for misplaced trust when explanations create undue confidence.
- **Impact of Accuracy:** High accuracy reduced skepticism, while low accuracy diminished reliance, especially in the absence of explanations.
- **Balancing Trust:** The metric T captured variations in trust under different conditions. Negative values of T indicated increased reliance on AI, with more negative values reflecting higher trust.

Table 2
Trust metric ($\mathcal{T}$) values and their interpretation.

Condition	$\mathcal{T}$	Interpretation
HA-WE	0.5	High reliance on AI, facilitated by accuracy and explanatory support.
HA-NE	0.2	Low reliance, slightly reduced due to lack of explanations.
LA-WE	-0.1	Low reliance despite low accuracy, influenced by explanations.
LA-NE	0.0	No reliance, reflecting mistrust in low-accuracy AI without explanations.

5. Discussion

By utilizing the metric defined in Equation (1), it becomes possible to quantitatively measure the trust of a specific user in particular decision-making scenarios. Drawing an analogy with explainable AI, $\mathcal{T}$ can be interpreted as a contextual and “local” trust measure. This definition enables its application in evaluating either the overall level of trust for a specific user or the trust associated with individual decisions (for example, by averaging multiple measures). Additionally, it facilitates the analysis of trust dynamics throughout an interaction, such as those modeled in the interaction framework proposed by [19].

The metric defined in Equation (1) takes values between -1 and 1, inclusive. Specifically, $\mathcal{T} = 0$ when $p\left(d_{\mathrm{H}}\right) = p\left(d_{\mathrm{H}} \mid d_{\mathrm{AI}}\right)$, meaning the probability of the user making a particular decision remains unchanged after considering the AI’s outcome. In this case, the user does not rely on AI. $\mathcal{T} = -1$ when $p\left(d_{\mathrm{H}}\right) = 1$ and $p\left(d_{\mathrm{H}} \mid d_{\mathrm{AI}}\right) = 0$, indicating that the user’s probability of making a particular decision is completely nullified. This represents a scenario where the user fully relies on the AI’s decision-making process, disregarding their initial judgment. Conversely, $\mathcal{T} = 1$ occurs when the user initially disregards a possibility but completely changes their decision after being exposed to the AI outcome, i.e., when $p\left(d_{\mathrm{H}}\right) = 0$ and $p\left(d_{\mathrm{H}} \mid d_{\mathrm{AI}}\right) = 1$. In other words, $\mathcal{T} = 1$ and $\mathcal{T} = -1$ represent analogous events: in the case of binary classification, if $\mathcal{T} = 1$ for one decision, then $\mathcal{T} = -1$ for the opposite decision.

It is important to emphasize that the proposed metric reflects the degree of human reliance on AI, capturing the extent to which the system’s outputs influence a user’s decision-making. This focus on observed changes rather than the specific decisions themselves ensures that the metric defined in Equation (1) remains context-agnostic, making it adaptable to various domains and scenarios. The metric provides a dynamic and interaction-dependent view of trust by isolating the shifts in human behavior caused by the introduction of AI-generated recommendations or classifications.

This approach recognizes that trust is not inherently tied to the correctness or quality of any decision but rather to the user’s perception of the system’s utility, reliability, and alignment with their goals, in line with HCAI view [6, 9, 8]. Therefore, the metric is well-suited to evaluate trust across different types of AI systems, from classification models to more complex decision-support tools, and across different domains.

5.1. Implications for XAI and AI Development

A formal definition of a trust metric, such as the one proposed in Equation (1), provides researchers and practitioners with a rigorous and systematic approach to assessing how different design choices influence users’ perception and reliance on AI systems. By offering a quantifiable measure, $\mathcal{T}$ enables the objective evaluation and comparison of various design solutions in XAI, ultimately guiding the development of more effective and user-aligned systems.

For instance, as demonstrated in Section 4, the metric can be used to assess the overall impact of explanations on user trust. Additionally, $\mathcal{T}$ facilitates a comparative analysis of different explanation modalities, helping identify which approach enhances the system’s trustworthiness most effectively. However, it is crucial to recognize that there is no universal solution, even in defining what constitutes an “improvement.” The optimal level of trust depends on the specific application domain. For example, in high-stakes areas such as tumor detection, an excessive increase in user trust—particularly if it leads to over-reliance on AI—could be detrimental, potentially compromising critical decision-making and posing risks to society. Thus, trust calibration remains an essential factor in designing responsible and reliable AI systems.

6. Metric Validation

To validate the metric proposed in Equation (1), different steps are required. First, to check *construct validity*, $\mathcal{T}$ should be better inserted within the relevant context in the academic literature, for example, contrasting it with theoretical frameworks of trust from psychology and HCI literature. Second, a *convergent validity* check should be performed by comparing this metric against different trust metrics. Third, *internal consistency* should be checked, ensuring the metric fully represents the intended human factor.

To verify the convergent validity and internal consistency, a series of user tests will be conducted across multiple domains, following a standardized protocol to ensure consistency in the evaluation process. Each test will involve users interacting with an AI system engaged in decision-making tasks. These tests aim to assess users' trust in AI decisions and whether the proposed metric aligns with user responses.

In each study, users will be presented with various data samples relevant to the domain under investigation. They will be asked to make decisions independently based on the presented data. This independent decision-making process serves as a baseline for understanding how users assess situations without the influence of AI-generated suggestions. Once users have made their initial decisions, they will be presented with an AI decision, either with or without an accompanying explanation: the inclusion of explanations allows us to explore whether transparency influences users' trust in the AI system's decisions. After seeing the AI decision, users will be asked whether they agree with it. This question serves to detect potential changes in the users' decisions regarding their initial classification.

Finally, users will complete a series of questionnaires to evaluate trust: we leverage existing quantitative questionnaires for user trust, described in Section 2, to provide a baseline to compare our metric against.

The trust metric $\mathcal{T}$ will then be computed using a frequentist approach to probability, taking into account the users' decisions and their agreement with the AI, as described in Section 3.1.1. The computed metric will then be compared against the questionnaires' results to identify any correlations. These correlations will help determine whether the metric accurately captures the trust that users place in the AI system, potentially validating the metric definition.

To further ensure the robustness and generalizability of the findings, the user studies will be designed to cover a variety of domains, e.g., healthcare, cybersecurity, finance, and autonomous vehicles. Each domain presents unique challenges and opportunities, and the overall protocol may thus be slightly tweaked to ensure its applicability to the chosen domain. In some domains, the studies are currently in the planning and recruitment phase, and once completed, the data will provide valuable insights into the robustness of the metric definition.

We will now outline examples of user studies designed to validate this metric, based on a frequentist definition of p.

6.1. Example 1—Medical Domain: Alzheimer's Detection

This study leverages an existing AI system for Alzheimer's disease detection [3]. Participants include professionals such as physicians, neurosurgeons, and neuroscientists. During the study, users will be provided with Magnetic-Resonance Imaging (MRI) scans and supplementary clinical information (e.g., Mini-Mental State Examination results, patient age, etc.) extracted from a pre-defined dataset of patients [20]. Initially, participants will independently determine whether they believe the patient is affected by Alzheimer's disease. Following this, depending on the study's experimental condition, they will be presented with the AI's classification, with or without accompanying explanations. Participants will then indicate whether they agree with the AI's assessment. Finally, they will complete questionnaires to evaluate their trust in the AI system. The trust metric $\mathcal{T}$ will be calculated as outlined in Section 3.1.1 and compared against the questionnaire responses to assess correlations and validate the metric.

6.2. Example 2—Cybersecurity Domain: Phishing Warning

This study utilizes an existing AI system to detect phishing attacks and provide warnings with explanations [21]. Participants are non-experts in cybersecurity, reflecting typical end-users of such systems. Users will be presented with a series of legitimate and phishing emails sourced from a ground-truth dataset. Initially, participants will independently decide whether they believe each email is phishing. Next, they will review the AI's classification, which may include explanation-based warnings depending on the experimental condition. Participants will then state whether they agree with the AI's judgment. To conclude, they will complete trust evaluation questionnaires. The trust metric $\mathcal{T}$ will be computed as detailed in Section 3.1.1 and compared to the questionnaire data to validate the metric's reliability.

7. Conclusion

This paper introduces a novel probabilistic metric for evaluating trust in AI systems, defined as the difference between the probabilities of a human decision with and without the AI's input. This addresses the need for standardized and scalable approaches to evaluate trust. By leveraging reliability as a proxy of trust, the metric provides a human-centered yet objective approach to evaluating trust in AI systems. The metric's definition offers flexibility and generalizability by avoiding pre-defined methods for computing probabilities. However, a frequentist implementation provides a simple way for initial computations, while the general definition remains open to different approaches.

While the proposed metric offers a novel approach to quantifying trust in AI systems, it has certain limitations that require further research. First, the metric has not yet undergone validation. Future work should focus on conducting user studies and experiments, for example following the outlined validation protocol, to establish its reliability and practical applicability. Validation will also allow refining the metric based on empirical findings, ensuring it accurately captures trust.

Second, the metric has not been tested. Throughout the validation process, rigorous testing should be performed to evaluate the quality of its formalization and applicability across diverse scenarios. This includes analyzing its sensitivity to different experimental conditions and assessing its robustness in various contexts of human-AI interaction.

Finally, the current definition of the metric is primarily tailored for classification tasks. Extending its applicability to regression problems and generative AI systems is fundamental. A more generalized formulation will enhance the metric's versatility, enabling its use in a broader range of AI applications.

Acknowledgments

The research of Andrea Esposito is supported by the Italian Ministry of University and Research (MUR) and by the European Union - NextGenerationEU, under grant PRIN 2022 PNRR "PROTECT: imPROving ciTizEn inClusiveness Through Conversational AI" (Grant P2022JJPBY) – CUP: H53D23008150001.

The research of Giuseppe Desolda is partially supported by the Italian Ministry of University and Research (MUR) under grant PRIN 2022 "DevProDev: Profiling Software Developers for Developer-Centered Recommender Systems" – CUP: H53D23003620006.

The research of Rosa Lanzilotti is partially supported by the co-funding of the European Union - Next Generation EU: NRRP Initiative, Mission 4, Component 2, Investment 1.3 – Partnerships extended to universities, research centers, companies and research D.D. MUR n. 341 del 15.03.2022 – Next Generation EU (PE0000013 – "Future Artificial Intelligence Research – FAIR" - CUP: H97G22000210007);

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022. URL: http://arxiv.org/abs/2204.06125. doi:10.48550/arXiv.2204.06125. arXiv:2204.06125.

[2] S. Cerri, O. Puonti, D. S. Meier, J. Wuerfel, M. Mühlau, H. R. Siebner, K. V. Leemput, A contrast-adaptive method for simultaneous whole-brain and lesion segmentation in multiple sclerosis, Neuroimage 225 (2020) 117471. URL: https://www.ncbi.nlm.nih.gov/pubmed/33099007. doi:10.1016/j.neuroimage.2020.117471. arXiv:33099007.

[3] G. Castellano, A. Esposito, E. Lella, G. Montanaro, G. Vessio, Automated detection of Alzheimer's disease: A multi-modal approach with 3D MRI and amyloid PET, Scientific Reports 14 (2024) 5210. URL: https://www.nature.com/articles/s41598-024-56001-9. doi:10.1038/s41598-024-56001-9.

[4] C. Combi, B. Amico, R. Bellazzi, A. Holzinger, J. H. Moore, M. Zitnik, J. H. Holmes, A Manifesto on Explainability for Artificial Intelligence in Medicine, Artificial Intelligence in Medicine (2022) 102423. URL: https://www.sciencedirect.com/science/article/pii/S0933365722001750. doi:10.1016/j.artmed.2022.102423.

[5] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A Survey of Methods for Explaining Black Box Models, ACM Computing Surveys 51 (2019) 1–42. URL: https://dl.acm.org/doi/10.1145/3236009. doi:10.1145/3236009.

[6] W. Xu, Toward Human-Centered Ai: A Perspective from Human-Computer Interaction, Interactions 26 (2019) 42–46. URL: https://dl.acm.org/doi/10.1145/3328485. doi:10.1145/3328485.

[7] B. Shneiderman, Human-Centered AI, 1 ed., Oxford University Press, Oxford, 2022.

[8] G. Desolda, A. Esposito, R. Lanzilotti, A. Piccinno, M. F. Costabile, From human-centered to symbiotic artificial intelligence: A focus on medical applications, Multimedia Tools and Applications (2024). URL: https://rdcu.be/d1RF4. doi:10.1007/s11042-024-20414-5.

[9] B. Shneiderman, Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy, International Journal of Human–Computer Interaction 36 (2020) 495–504. URL: https://www.tandfonline.com/doi/full/10.1080/10447318.2020.1741118. doi:10.1080/10447318.2020.1741118.

[10] T. A. Bach, A. Khan, H. Hallock, G. Beltrão, S. Sousa, A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective, International Journal of Human–Computer Interaction 40 (2024) 1251–1266. URL: https://www.tandfonline.com/doi/full/10.1080/10447318.2022.2138826. doi:10.1080/10447318.2022.2138826.

[11] A. M. Do, A. V. Rupert, G. Wolford, Evaluations of pleasurable experiences: The peak-end rule, Psychonomic Bulletin & Review 15 (2008) 96–98. URL: https://doi.org/10.3758/PBR.15.1.96. doi:10.3758/PBR.15.1.96.

[12] R. C. Mayer, J. H. Davis, F. D. Schoorman, An Integrative Model of Organizational Trust, The Academy of Management Review 20 (1995) 709. URL: http://www.jstor.org/stable/258792?origin=crossref. doi:10.2307/258792. arXiv:258792.

[13] J. D. Lee, K. A. See, Trust in Automation: Designing for Appropriate Reliance, Human Factors 46 (2004) 50–80. URL: http://hfs.sagepub.com/cgi/doi/10.1518/hfes.46.1.50_30392. doi:10.1518/hfes.46.1.50_30392.

[14] Z. Yan, Y. Dong, V. Niemi, G. Yu, Exploring trust of mobile applications based on user behaviors: An empirical study, Journal of Applied Social Psychology 43 (2013) 638–659. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1559-1816.2013.01044.x. doi:10.1111/j.1559-1816.2013.01044.x.

[15] H. M. Wojton, D. Porter, S. T. Lane, C. Bieber, P. Madhavan, Initial Validation of the Trust of Automated Systems Test (TOAST), The Journal of Social Psychology 160 (2020) 735–750. URL: https://doi.org/10.1080/00224545.2020.1749020. doi:10.1080/00224545.2020.1749020.

[16] D. Ullman, B. F. Malle, What Does it Mean to Trust a Robot?: Steps Toward a Multidimensional Measure of Trust, in: Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, ACM, Chicago IL USA, 2018, pp. 263–264. URL: https://dl.acm.org/doi/10.1145/3173386.3176991. doi:10.1145/3173386.3176991.

[17] J.-Y. Jian, A. M. Bisantz, C. G. Drury, Foundations for an Empirically Determined Scale of Trust in Automated Systems, International Journal of Cognitive Ergonomics 4 (2000) 53–71. URL: http://www.tandfonline.com/doi/abs/10.1207/S15327566IJCE0401_04. doi:10.1207/S15327566IJCE0401_04.

[18] A. Gut, Probability: A Graduate Course: A Graduate Course, volume 75 of *Springer Texts in Statistics*, Springer New York, New York, NY, 2013. URL: https://link.springer.com/10.1007/978-1-4614-4708-5. doi:10.1007/978-1-4614-4708-5.

[19] G. Desolda, G. Dimauro, A. Esposito, R. Lanzilotti, M. Matera, M. Zancanaro, A Human–AI interaction paradigm and its application to rhinocytology, Artificial Intelligence in Medicine 155 (2024) 102933. URL: https://linkinghub.elsevier.com/retrieve/pii/S0933365724001751. doi:10.1016/j.artmed.2024.102933.

[20] P. J. LaMontagne, T. L. S. Benzinger, J. C. Morris, S. Keefe, R. Hornbeck, C. Xiong, E. Grant, J. Hassenstab, K. Moulder, A. G. Vlassenko, M. E. Raichle, C. Cruchaga, D. Marcus, OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease, medRxiv : the preprint server for health sciences (2019). URL: https://doi.org/10.1101/2019.12.13.19014902. doi:10.1101/2019.12.13.19014902.

[21] G. Desolda, F. Greco, L. Viganò, APOLLO: A GPT-based tool to detect phishing emails and generate explanations that warn users, 2024. URL: https://arxiv.org/abs/2410.07997. doi:10.48550/ARXIV.2410.07997.