Assessing and Quantifying Perceived Trust in Interpretable Clinical Decision Support

Mohsen Abbaspour Onari[1,2(✉)], Isel Grau[1,2], Chao Zhang[3], Marco S. Nobile[4], and Yingqian Zhang[1,2]

[1] Information Systems, Eindhoven University of Technology, Eindhoven, The Netherlands
[2] Eindhoven Artificial Intelligence Systems Institute, Eindhoven University of Technology, Eindhoven, The Netherlands
[3] Human-Technology Interaction, Eindhoven University of Technology, Eindhoven, The Netherlands
{m.abbaspour.onari,i.d.c.grau.garcia,c.zhang.5,yqzhang}@tue.nl
[4] Computational Biology, Bioinformatics and Biomedicine, Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Venice, Italy
marco.nobile@unive.it

Abstract. Technical and ethical concerns impede the establishment of trust among healthcare professionals (HCPs) in developing artificial intelligence (AI)-based decision support. Yet, our understanding of trust models is constrained, and a standard accepted approach to evaluating trust in AI models is still lacking. We introduce a novel methodology to assess and quantify HCPs' perceived trust in an interpretable machine learning model that serves as clinical decision support for diagnosing COVID-19 cases. Our approach leverages fuzzy cognitive maps (FCMs) to elicit and quantify HCPs' trust mental models for understanding trust dynamics in clinical diagnosis. Our study reveals that HCPs rely predominantly on their own expertise when interacting with the developed interpretable clinical decision support. Although the model's interpretations offer limited assistance in diagnostic tasks, they facilitate the HCPs' utilization of it. However, the impact of these interpretations on the establishment of perceived trust varies among HCPs, which can lead to an increase in trust for some while decreasing it for others. To validate quantified perceived trust, we employ the degree of agreement metric, which quantitatively assesses whether HCPs lean more towards their own expertise or rely on the model's recommendations in diagnostic tasks. We found significant alignment between the conclusions of the two metrics, indicating successful modeling and quantification of perceived trust. Plus, a moderate to strong positive correlation between the two metrics confirmed this conclusion. This means that FCMs can quantify HCPs' perceived trust, aligning with their actual diagnostic advice shift after interacting with the model.

R. Guidotti et al. (Eds.): xAI 2025, CCIS 2578, pp. 202–222, 2026.
https://doi.org/10.1007/978-3-032-08327-2_10

Keywords: Perceived Trust · User Study · Explainable AI · Interpretable AI · Clinical Decision Support

1 Introduction

The need for trustworthy artificial intelligence (TAI) systems is clear in driving the integration of AI into healthcare, primarily due to the limited measurable benefits observed in real-world patient care, despite the promising results demonstrated by an increasing number of AI-driven clinical decision support systems in preclinical and in silico studies [37]. The European Ethics Guidelines for TAI [11] outlines specific criteria to establish trustworthiness that AI systems must comply with. The existing literature reveals a scarcity of prospective studies to validate proposed AI solutions in real-world settings [12]. This scarcity has resulted in diminished trust from healthcare professionals (HCPs) towards the developed solutions. Therefore, in the current study, we propose a methodology to assess and quantify *perceived trust* of HCPs in interpretable clinical decision support. We aim to adhere to the guidelines set forth by the General Data Protection Regulation (GDPR) [10], emphasizing the critical role of transparency and explainability in establishing TAI. Miller [23] recommends incorporating interpretable machine learning (IML) models, particularly in high-stakes tasks, to enhance the comprehensibility and reliability of AI systems. Doshi-Velez and Kim [9] presented a taxonomy of IML model evaluation methodologies, including application-based assessments that involve domain experts using the IML model. Aligned with the goals of this study, which aim to assess and quantify HPCs' perceived trust in the developed clinical decision support, and considering the problem's high sensitivity, we involve HCPs to develop our methodology. Our proposed methodology considers perceived trust as a dynamic entity affected by different elements. Hence, we aim to elicit and quantify HCPs' perceived trust mental model.

Initially, IML serves as clinical decision support, recommending and interpreting diagnostic advice. This helps categorize suspected COVID-19 patients into positive or negative cases. Then, we have structured a diagnostic task that engages the HCPs in diagnosing the COVID-19 status of selected patients under two distinct scenarios: (i) relying on their expertise and (ii) interacting with the IML model. Then, they will express their satisfaction with the effectiveness of interpretations in the diagnostic task, using the Explanation Satisfaction Scale (ESS) proposed by Hoffman *et al.* [13] using fuzzy linguistic variables. This phase seeks four main objectives: (i) its efficacy in assisting HCPs in diagnosing the disease, (ii) the effectiveness of interpretations in diagnosing the disease, (iii) the impact of interpretations on establishing HCPs' trust, and (iv) incorporating HCPs' subjectivity and uncertainty through the use of fuzzy variables. In the next phase of the research, HCPs will contribute to eliciting their mental models of perceived trust in the IML model based on the influence of ESS on their perceived trust using fuzzy cognitive maps (FCM) [19] using fuzzy linguistic variables. FCMs model and simulate dynamic systems with complex interac-

tions, allowing decision-makers to forecast future system states through scenario-making, learning algorithms, and current state analysis [1]. Upon constructing FCMs by HCPs, a distinctive perceived trust mental model will be established for each HCP and we can derive a quantified value indicative of perceived trust. To validate the results, we measure the *degree of agreement (DoA)* between the diagnostic advice of HCPs when relying on their expertise and after interaction with the IML model. This measure can elucidate whether HCPs exhibit reliance on their expertise or lean toward the IML model's recommendations. Considerable alignment and correlation between the two metrics can indicate whether FCMs could successfully measure HCPs' perceived trust. In undertaking this research, we contribute to the literature in several ways by addressing the following gaps.

- Prospective studies validating AI solutions remain limited, as noted by Nauta *et al.* [26], revealing a gap in the literature on eXplainable AI (XAI) with respect to application-based performance assessments of IML models. The contribution of medical experts is crucial to this study, as their involvement is essential for establishing the realism and reliability of a trust analysis. Although the number of participating experts is limited to 15, their input plays a vital role in understanding trust behaviors toward AI models.
- Trust in AI models is often evaluated using Hoffman's trust scale [13], which relies on Likert-scale questions to provide a simplified representation of users' trust perception. To better model this perception, we utilize FCM to extract mental models of HCPs, capturing their trust perception following their interaction with the XAI model.
- Existing methodologies overlook the role of transparency, interpretability, and explainability in shaping trust [22]. By embedding the model's interpretability into the diagnostic decision-making process and leveraging FCM's capability to model the impact of interpretability on trust, we emphasize the critical role of interpretability in modeling and measuring perceived trust.

The subsequent sections of this paper are structured as follows. In Sect. 2, we investigate the XAI literature, reviewing studies that address trust in XAI models. Section 3 encompasses the primary definition of trust and used methods to develop the proposed methodology. The experimental task designed to measure and quantify perceived trust is described in Sect. 4. In Sect. 5, we will validate the proposed methodology. Lastly, Sect. 6 encompasses the discussion of the results and outlines potential avenues for future research.

2 Background

Nauta *et al.* [26] found that a minority of XAI papers engage users in evaluating model explanations, a trend consistent even when domain experts are involved in assessments. Also, Vereschak *et al.* [38] conducted a comprehensive study revealing a lack of organized research on modeling decision-makers' trust, inspiring us to assess the impact of model interpretations on the trust levels of

HCPs. Lakkaraju and Bastani [20] conducted groundbreaking research aimed at empirically establishing how user trust in black box models can be manipulated through misleading explanations. However, the study did not compare the results with users' own perceptions regarding the efficacy of explanations in decision-making and their trust levels to elicit their mental models. Zhang *et al.* [42] underscored that local explanations for AI-assisted decision-making struggle to accurately calibrate human trust in AI. Nonetheless, they did not also directly assess users' perceptions regarding the effectiveness of explanations in facilitating the decision-making process.

In their empirical evaluation of XAI methods, Wang and Yin [39] conducted a comparison of established XAI techniques, analyzing their impact on AI-assisted decision-making and user trust. However, their study did not delve into users' perceptions regarding the effectiveness of these explanations in shaping their decision-making processes. Bansal *et al.* [4] conducted mixed-method user studies on three datasets. In these studies, participants were assisted by an AI system, with accuracy comparable to humans, in completing tasks. The AI system explained itself in some conditions, and the researchers studied whether users trusted the XAI model or not. However, the results may not be generalizable to high-stakes domains with expert users, such as medical diagnosis. Yang *et al.* [41] investigated the effects of example-based explanations for an ML classifier on end users' appropriate trust. However, we contend that they primarily measured agreement rather than trust. Additionally, their focus was solely on the efficacy of explanations in terms of helpfulness, neglecting other essential aspects of ESS. In Huber *et al.*'s study [14], which explored the impacts of global and local explanation methods on reinforcement learning agents, the methodology primarily focuses on assessing users' agreement rather than their trust. The interpretable decision support interface for sepsis treatment proposed by Sivaraman *et al.* [35] predominantly examines the influence of AI model explanations on HCPs' confidence in their diagnoses, yet it only marginally addresses their trust in the IML model.

Wysocki *et al.* [40] introduced a pragmatic evaluation framework for XAI within clinical decision support in a separate study. However, their approach merely assesses HCPs' trust with a simplistic survey, lacking a systematic method to assess trust in the AI model. In an extensive study, Mehrotra *et al.* [21] showed the impact of various integrity-based explanations made by an AI agent on the appropriateness of human trust in that agent. However, their evaluation focused solely on the usefulness of the provided explanations in decision-making tasks and corresponding trust, neglecting other essential factors of ESS. Joshi *et al.* [18] presented a Wizard of Oz study comparing low- and high-explainability versions of a vacation planning chatbot in a between-subjects design, examining the effect of explainability on users' understanding, trust, and acceptance. Chanda *et al.* [6] developed an XAI model to generate domain-specific, interpretable explanations to support melanoma diagnosis. In this study, medical experts assessed their trust in the model using a 10-point Likert scale. Perlmutter *et al.* [32] also investigated the impact of an example-based XAI interface on

trust, understanding, and performance in highly technical populations using a 10-point Likert scale.

Therefore, the main shortcomings of existing studies in the literature can be summarized as follows: misdefining the concept of trust, conducting studies at a general user level without involving domain experts, and neglecting to assess the efficacy of explanations in facilitating decision-making tasks and their impact on users' trust, which causes overlooking the elicitation of users' trust mental models.

3 Methodology

In this section, we outline the foundational concepts of this study. Sub-sect 3.1 explores the definition of trust, while Subsect. 3.2 introduces FCM.

3.1 XAI and Perceived Trust

Trust is generally defined as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, regardless of the ability to monitor or control that other party [36]." In the same way, when a user trusts the AI model, the anticipation depends on whether the model can fulfill its expectations. Here, we refer to the definition of Jacovi et al. [16] for Human-AI trust:

"If H (human) perceives that M (AI model) is trustworthy to contract C and accepts vulnerability to M's actions, then H trusts M contractually to C. The objective of H in trusting M is to anticipate that M will maintain C in the presence of uncertainty; consequently, trust does not exist if H does not perceive risk."

Ribeiro *et al.* [33] asserted the importance of trust for effective human interaction with ML systems, emphasizing the importance of explaining individual predictions as a key factor in assessing trust. By hypothesis, effective and satisfying explanations enable users to construct a good mental model. So, this sound mental model can facilitate the development of trust in AI and enhance user performance when using it [13]. Miller [22] states that trust as a mental attitude must be measured in field studies, lab experiments, and surveys/interviews with human participants. The main reason is that trust can rapidly deteriorate when subjected to factors such as time constraints, noticeable system defects, high error rates, or frequent false alarms [13]. Like the diverse forms of trust, various manifestations of negative trust exist, including mistrust and distrust [13]. The proposed trust continuum by Cho et al. [7] can demonstrate this behavior (see Fig. 1).

In the XAI domain, the trust assessment is based mainly on the trust scale proposed by Hoffman et al. [13]; however, Miller [22] declares that the trust scale presented does not explicitly measure the effect of trust. In fact, this scale measures users' trust through a set of Likert scale questions, primarily focusing on

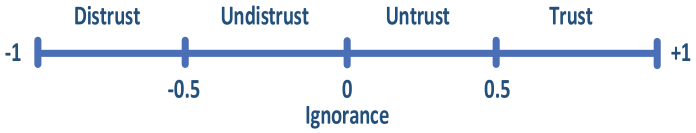


Fig. 1. Trust continuum [7].

the users' perception of trust rather than their demonstrated trust when interacting with an XAI model. Besides, the existing techniques do not measure the impact of transparency, interpretability, and explainability methods on human participants' trust [22]. In essence, trust measurement efforts have often focused on precisely defining the elements of trust to measure perceived trust, often overlooking the underlying mental models that shape users' perceptions. Therefore, it is imperative to introduce a methodology that delves into the influence of explanations on trust establishment and examines how they can impact users' perceived trust.

3.2 Eliciting Perceived Trust Mental Models by FCM

The ESS proposed by Hoffman *et al.* [13] serves as the foundational elements of HCPs' perceived trust mental models in this study so that we can analyze the contribution of interpretation in building trust. We slightly modified the ESS for our specific context by adding a "Functionality" scale, as outlined in Table 1. In this study, trust is considered a dynamic entity with intricate interactions among ESS, and we model it using FCM.

Table 1. Explanation Satisfaction Scale and description

ESS	Description
Understandability (US)	The interpretation was understandable in diagnosing the disease.
Sufficiency of details (SD)	The interpretation had sufficient details to help me diagnose the disease.
Completeness (CL)	The interpretation was complete enough to diagnose the disease.
Feeling of satisfaction (FS)	I am satisfied with the quality of the interpretation for diagnosing the disease.
Accuracy (AC)	The interpretation was accurate enough to diagnose the disease.
Usability (US)	Interpretation is easy to use to diagnose the disease.
Functionality (FC)	In general, the interpretation helped me diagnose the disease.

Kosko [19], for the first time, introduced FCMs to mitigate the limited ability of cognitive maps [3] to represent causal beliefs in social scientific knowledge [25]. Multiple domain experts who have knowledge in a particular area contribute as knowledge engineers to manually develop an FCM or a mental model [30]. They start by identifying key domain components or concepts (C) and then determine the influence (edges) of concepts, including their strength on each other or weight (w) [30]. A semantic representation of an FCM (including concepts, edges, and

weights) is shown in Fig. 2. In our study, ESS serves as FCM concepts, and their initial value, edges, and weights are determined by HCPs.

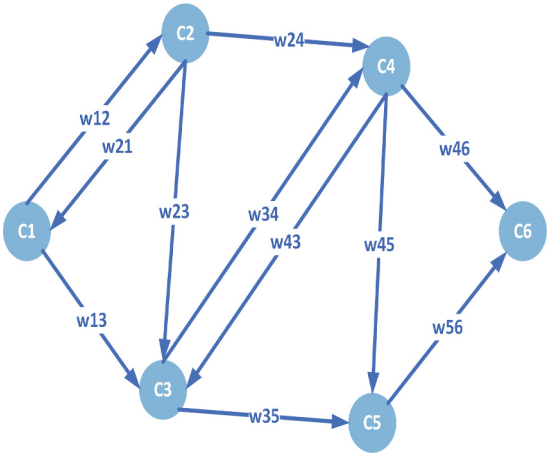


Fig. 2. A semantic representation of an FCM.

There are three types of relationships between concepts in the FCM [28]:

- $w_{ij} > 0$, direct influence between concepts C_i and C_j,
- $w_{ij} < 0$, inverse influence between concepts C_i and C_j,
- $w_{ij} = 0$, no relationship between concepts C_i and C_j.

The established reasoning process of an FCM [19,25,28,30], uses the following simple mathematical formula:

$$C_i^{(k)} = f\left(\mathrm{C}_i^{k-1} + \sum_{j=1, j\neq i}^{N} C_j^{(k-1)} \cdot w_{ji}\right), \tag{1}$$

where, $C_i^{(k)}$ represents the value of concept i at iteration k of the reasoning process. w_{ji} indicates the weight of the edge from C_j to C_i, and N is the number of entered edges to C_i. Our study utilizes a state vector of size 1×8, encompassing ESS and a target concept denoted as perceived trust (PT).

The initial values of these concepts reflect HCP's subjective satisfaction with ESS effectiveness in diagnostic tasks, employing fuzzy linguistic variables detailed in Table 2. With this approach, we achieve two primary objectives: firstly, we gain insight into the satisfaction level of HCP with interpretations; secondly, we embed HCP's satisfaction impact in establishing the perceived trust, which, in fact, models their trust mental model based on the model's interpretability. To convert these linguistic variables into actionable data to develop FCM, defuzzification is applied to convert them into crisp numbers (see Table 2),

employing the center of gravity (CoG) method [31]. Due to its high accuracy, the CoG defuzzification method is the most widely used in practice [2]. It effectively satisfies important criteria such as continuity, disambiguity, and plausibility, contributing to its reliability and interoperability [2]. These initial values fall within the interval [0, 1], with proximity to 1 indicating higher importance.

w is an 8×8 weighted matrix defining relationships between ESS and PT, determined by HCPs using the linguistic variables outlined in Table 3. In the same way, defuzzification is applied to weights as well (see Table 3), transforming them within the range [-1, 1], with values closer to 1 indicating stronger influence and the sign denoting direct or inverse influence between concepts. Following this approach, we integrate the influence of each individual ESS on one another, ultimately culminating in their collective impact on the perceived trust of HCP. The activation function $f(x)$, typically sigmoid or hyperbolic tangent, is employed to constrain the state vector's values within [0, 1] and [-1, 1], respectively. Our study adopts the hyperbolic tangent function to align perceived trust values with the trust continuum outlined in Fig. 1. According to FCM literature, interaction among concepts persists until one of the following states occurs [5]:

- stable state: The model reaches an equilibrium fixed point, with output values settling at constant numerical levels.
- limit cycle: The concept values fall in a loop of numerical values.
- chaotic behavior: The model exhibits non-deterministic, random fluctuations in concept values.

We set a maximum iteration limit for the algorithm, ensuring that it terminates after this number of iterations, regardless of convergence status. Finally, the ultimate value of PT in the state vector quantifies the corresponding HCP's perceived trust level. This process will be repeated for all HCPs to elicit a unique mental model for each participant involved in this study.

Table 2. Linguistic variables for the initial values of C.

	Linguistic variables	Membership function	Defuzzified value
1	I disagree strongly	(0, 0, 0.25)	0
2	I disagree somewhat	(0, 0.25, 0.5)	0.25
3	I'm neutral about it	(0.25, 0.5, 0.75)	0.5
4	I agree somewhat	(0.5, 0.75, 1)	0.75
5	I agree strongly	(0.75, 1, 1)	1

4 Experimental Design

This section outlines the step-by-step process used to quantify perceived trust in this study. Subsect. 4.1 introduces the implemented dataset, the training of the

Table 3. Linguistic variables to determine w.

	Linguistic variables	Membership function	Defuzzified value
1	Inversely high	(-1, -1, -0.5)	-1
2	Inversely low	(-1, -0.5, 0)	-0.5
3	No influence	(-0.5, 0, 0.5)	0
4	Directly low	(0, 0.5, 1)	0.5
5	Directly high	(0.5, 1, 1)	1

IML model, and its interpretation. Subsect. 4.2 explains how HCPs were selected for the study. Subsect. 4.3 analyzes the shift in diagnostic advice during the decision-making task before and after interaction with the IML model. Subsect. 4.4 evaluates HCPs' satisfaction with the model's interpretability. Subsect. 4.5 presents the elicited mental models of HCPs' using FCM, followed by presenting the quantified perceived trust for each HCP based on FCM implementation in Subsect. 4.6.

4.1 Clinical Setting and Exploited IML Model

Data Set: The data set comprises the results of blood sample tests obtained from suspected patients with COVID-19 upon their arrival in the emergency department, encompassing a minimum of 30 distinct clinical measurements. The data set comprises 12873 patients with 32 clinical features derived from blood samples. We followed the ethical aspects of the AI application by signing written agreements regarding the limited use of data. Second, we adhered to security measures to protect data privacy per the agreements. Third, patients' identities were removed. The data set includes missing values in both the features and labels. Certain observations collected before the COVID-19 outbreak were classified as negative cases. Observations with no labels and missing values exceeding 40% were discarded as they offer no meaningful information for the IML model. Patients under the age of 18 years were also excluded. Ultimately, the data set comprises 8781 observations, of which 8461 are negative and 320 are positive.

IML Model: Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [8] is an IML algorithm that operates on rules directly learned from the data. Abbaspour Onari *et al.* [27] showed its high predictive performance compared to other ML models in COVID-19 prediction. RIPPER produces IF-THEN classification rules using the separate-and-conquer technique and the reduced-error pruning approach. Afterward, a set of rules is returned, which can be applied to classify new objects [27]. Before implementing RIPPER, KNN data imputation is applied to correct 2563 missing values in the data set. Then, the correlation between the features is calculated, and features with a higher correlation value of 0.7 with each other are dropped from the data set, leaving 27 features to build the IML model. The data set is split into training and test

data sets with 80%–20% partition, respectively. Although RIPPER shows high capability in unbalanced data set classification, we implement the SMOTE oversampling technique to have the same number of positive and negative cases in the training phase. Furthermore, the model's hyperparameters are optimized using grid search. The model undergoes 5-fold cross-validation on the training data set to validate its performance. The results on the test dataset demonstrate performance metrics of 0.9841 for accuracy, 0.8667 for precision, 0.6393 for recall, and 0.7358 for the F1 score.

Interpretations: To interpret the prediction's logic, RIPPER generates three rules on the test data set represented in Table 4. The instances that satisfy either of these rules are classified as positive cases, and all others are considered as negative cases. Building upon the insights of Huysmans et al. [15], which demonstrated that representing decision rules in decision tables enhances respondents' understanding of the rules, we will present RIPPER rules in the same format. We represented RIPPER's logic in correctly diagnosing a truly affected patient in Table 5 as a visual representation in Fig. 3. The legend in the figure explains the colors used: orange indicates that the conditions based on the patient's clinical features are not verified in the RIPPER's conditions, blue shows that the patient's features are verified in the RIPPER's conditions, and purple highlights when all conditions are satisfied, and the rule is applied to the patient. In cases where only a single rule is satisfied in RIPPER, that specific rule becomes the sole basis for the classification decision.

Table 4. Rules generated by RIPPER to classify patients into positive cases.

Feature	Rule 1	Rule 2	Rule 3
Albumin	≤ 37.9	-	-
Alkaline Phosphatase	≤ 82	≤ 83.6	-
Calcium	≤ 2.28	-	-
Erythrocytes	≥ 3.94	-	≥ 4.29
Glucose	≥ 5.66	-	-
Lactate Dehydrogenase	≥ 302	-	≥ 320
Basophils	-	≤ 0.01	-
C-Reactive Protein	-	≥ 19.62	-
Leukocytes	-	≤ 7.69	≤ 7.68
Lipase	-	≥ 30.5	-
Mean Cellular Haemoglobin	-	-	≥ 1.85

4.2 Selection of Participants

In the current study, we use an IML model to recommend and interpret diagnostic advice to HCPs due to the high-stakes nature of decision-making. Inter-

Table 5. Blood sample test results of a patient diagnosed as a positive case by RIPPER.

	Features	Test result
1	Albumin	37
2	Alkaline phosphatase	70
3	Basophils	0.03
4	Calcium	2.09
5	C-reactive protein	1.43
6	Erythrocytes	4.75
8	Glucose	10.36
9	Lactate dehydrogenase	392
10	Leukocytes	7.13
11	Lipase	47.8
12	Mean Cellular Haemoglobin	1.895

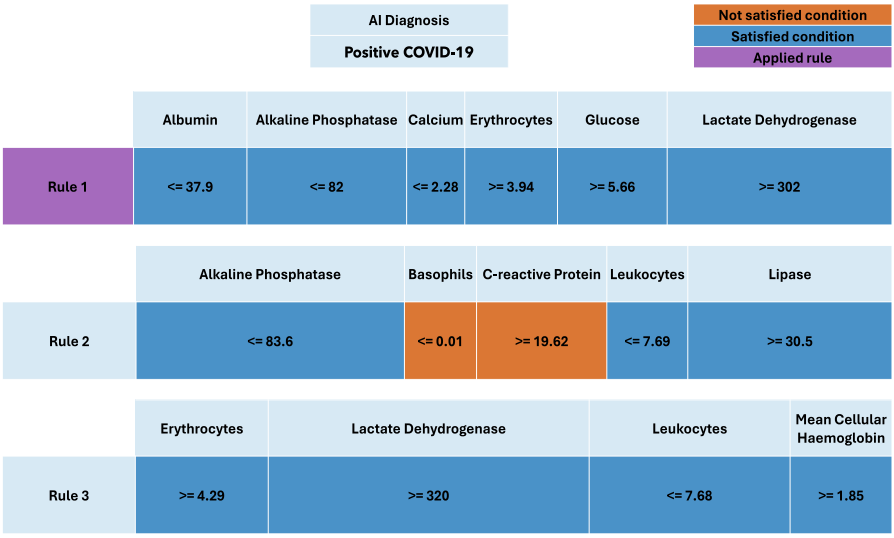


Fig. 3. Representation of RIPPER's rules as decision tables.

pretable models rely on a limited set of features characterized by a low complexity. The underlying assumption is that the model encompasses the necessary explanatory information due to its interpretability [23]. This study will focus on understanding how interpretability can build perceived trust among HCPs in the IML model. The university's ethical board granted ethical approval for this research project[1]. The participants in our study, including HCPs, were identi-

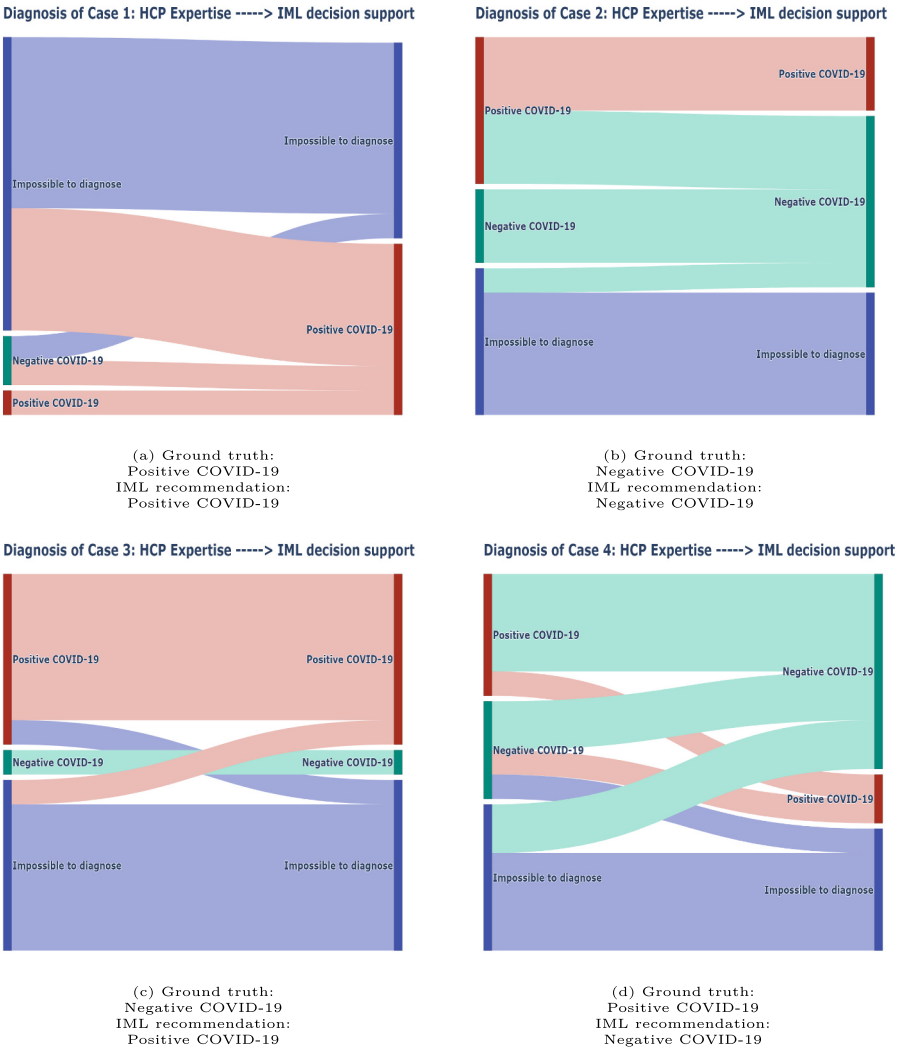
[1] This study has been approved by Ethical Board of the university with reference number: ERB2023IEIS10.

fied by snowball sampling. Initially, we contacted our network of clinically active HCPs with practical experience in healthcare centers during the COVID-19 pandemic. Subsequently, we requested them to send our participation request to individuals who meet our criteria and might be interested. Our research involved a total of N=15 HCPs. While this sample size classifies the study as a pilot, the valuable contributions of the HCPs make it highly insightful and meaningful. We used Qualtrics software as our main tool to design the user study. First, the HCPs responded to three questions about their professional background, professional tenure, and whether they wanted to participate in this research voluntarily. This question is apart from the ethical consent forms sent to them. If they had opted not to participate, their survey would have been terminated immediately. HCPs are general practitioners, senior medical students, cardiovascular imaging specialists, medical specialists in infectious diseases, and internal medicine specialists. Our participants have at least two years of professional work experience in healthcare centers and, at most, 13 years. HCPs from diverse geographic locations participated: Iran (10), Italy (2), Canada (1), Australia (1) and the UK (1). The gender distribution comprised 7 men and 8 women.

4.3 Diagnostic Task: Diagnostic Advice Shift

Four instances were selected from the test data set to present to all HCPs. In two cases, the ground truth status aligns with the recommendation of the IML model, while in the remaining two, there are contradictions. In the first sub-task, the clinical blood sample test results (As shown in Table 5) are presented to HCPs, and they are asked to offer their diagnostic advice relying on their expertise. The same question is asked in the next sub-task, including generated rules by IML and recommendations (see Fig. 3) functioning as clinical decision support to diagnose the disease. For both sub-tasks, HCPs can choose an option between "Positive COVID-19," "Negative COVID-19," and "Not possible to diagnose." The results of the diagnostic task have been outlined in Fig. 4. Using Sankey diagram, we show how HCPs change their diagnostic advice after interaction with the clinical decision support.

In Case 1, following their interaction with the IML model, seven HCPs adjusted their diagnostic advice. Remarkably, six of them aligned their advice with the model's recommendation. Notably, one HCP revised their initial diagnosis from "Negative COVID-19" to "Not possible to diagnose." This adjustment can be deemed a positive impact of the model, revealing the HCP's initial lack of confidence in their initial diagnostic advice. Moving on to Case 2, four HCPs modified their diagnostic advice to "Negative COVID-19" after engaging with the model. Regarding Case 3, there was no discernible shift in the diagnostic advice patterns of HCPs. It appears that the clinical features recommended by the model lacked sufficient information for the HCPs. It is plausible that these features resembled those of a patient with "Positive COVID-19," prompting HCPs to err on caution. In Case 4, six HCPs followed the model's recommended advice after interacting with it. In conclusion, we assert that the IML model can influ-



(a) Ground truth: Positive COVID-19 IML recommendation: Positive COVID-19

(b) Ground truth: Negative COVID-19 IML recommendation: Negative COVID-19

(c) Ground truth: Negative COVID-19 IML recommendation: Positive COVID-19

(d) Ground truth: Positive COVID-19 IML recommendation: Negative COVID-19

Fig. 4. Diagnostic advice shift of HCPs before and after interaction with IML model's recommendations and interpretation.

ence HCPs' diagnostic advice in at least three tasks to some extent, though not drastically.

4.4 HCPs' Satisfaction with Interpretations

After completing the diagnostic tasks, the HCPs expressed their satisfaction with the effectiveness of the model's interpretation as outlined in Sect. 3.2. The results have been demonstrated in Fig. 5. The results indicate that HCPs perceived the

model's interpretations as insightful across three scales: understandability, satisfaction, and usability. However, regarding completeness and accuracy, HCPs did not deem the model's interpretations sufficiently informative. This observation might help explain why HCPs were inclined to refrain from providing precise diagnostic advice on the diagnostic task and prefer to rely on their expertise. Finally, when it comes to evaluating the sufficiency of details and functionality, a notable lack of meaningful consensus among HCPs is apparent. Consequently, no informative conclusion can be drawn from these aspects. In conclusion, the utility of the model's interpretability appears more evident when HCPs intend to utilize it for their understanding rather than as a significant source of information for offering diagnostic advice.

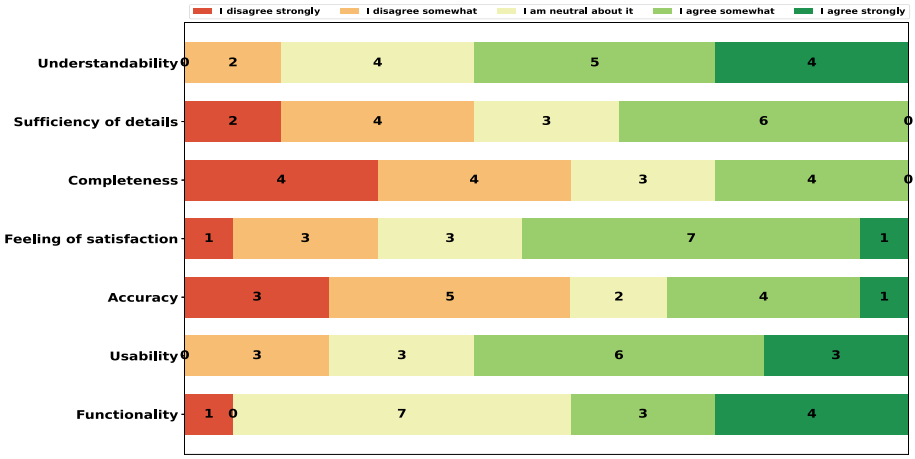


Fig. 5. HCPs' satisfaction with model's interpretations.

4.5 Eliciting HCPs' Mental Models Using FCM

In the conclusive phase of the experiment, HCPs contributed to eliciting their perceived trust mental models, as detailed in Sect. 3.2. Using FCMExpert tool [24] to semantically visualize mental models as FCMs, we identified four discernible patterns: trust, distrust, neutrality, and unknown, as illustrated in Fig. 6. The positive-weighted edges that originate from the ESS and enter PT in Fig. 6a represent a direct influence of ESS on HCP trust. An elevation in ESS values corresponds to an increase in perceived trust in them. In contrast, the negative-weighted edges from ESS to PT in Fig. 6b indicate that an increase in ESS values leads to a decreased perceived trust of HCPs. This observation may stem from the realization that the model falls short of meeting their expectations when its interpretability is increased. The neutrality behavior emerges when the cumulated weight in PT converges to zeros, signifying that edges' weights neutralize

each other, and HCPs feel neutral about the IML model (see Fig. 6c). Finally, the solitary unknown pattern indicates that the HCP perceives no influence from ESS on PT and vice versa (see Fig. 6d). Extracting meaningful information from this pattern regarding HCP's perceived trust is challenging. To keep the paper concise, we present mental models for four HCPs to illustrate key patterns.

Table 6. Quantified perceived trust and DoA results. The "Alignment" column shows whether both metrics converge in a common conclusion.

HCP	PT	DoA	Alignment
HCP1	0.8098	1.1660	✓
HCP2	−0.7619	0.6111	✓
HCP3	0.9992	0.7849	×
HCP4	unk	0.9668	-
HCP5	0.9836	0.8876	×
HCP6	0.0	0.8326	✓
HCP7	−0.7805	0.6489	✓
HCP8	0.5884	1.0694	✓
HCP9	−0.0746	0.7849	✓
HCP10	0.9997	1.1660	✓
HCP11	0.2061	0.6111	✓
HCP12	0.9997	1.1660	✓
HCP13	0.9999	1.6988	✓
HCP14	0.9942	1.0694	✓
HCP15	−0.4631	1.0694	×

4.6 Quantified Value of Perceived Trust for Each HCP

The FCM for each HCP was implemented following the reasoning process outlined in Sect. 3.2. The reasoning process is terminated once the algorithm reaches its maximum iteration limit (k = 35). To determine this value, we began by testing smaller iteration counts (e.g., 10) and observed whether FCM demonstrates triple stop criteria. If it did not, we incrementally increased k and reassessed it. Eventually, at k = 35, all concepts either reached steady state convergence or showed chaotic behavior, thereby satisfying the stop criteria outlined in Subsect. 3.2.

Upon completing the FCM implementation, we derived the quantified PT values for each HCP, as presented in Table 6. Comparing the results obtained with the trust continuum illustrated in column "PT" of Fig. 1 reveals distinctive patterns. Two HCPs (2 and 7) demonstrate distrust towards the model, while two others (9 and 15) exhibit an undistrusting stance. HCP6 expresses a neutral

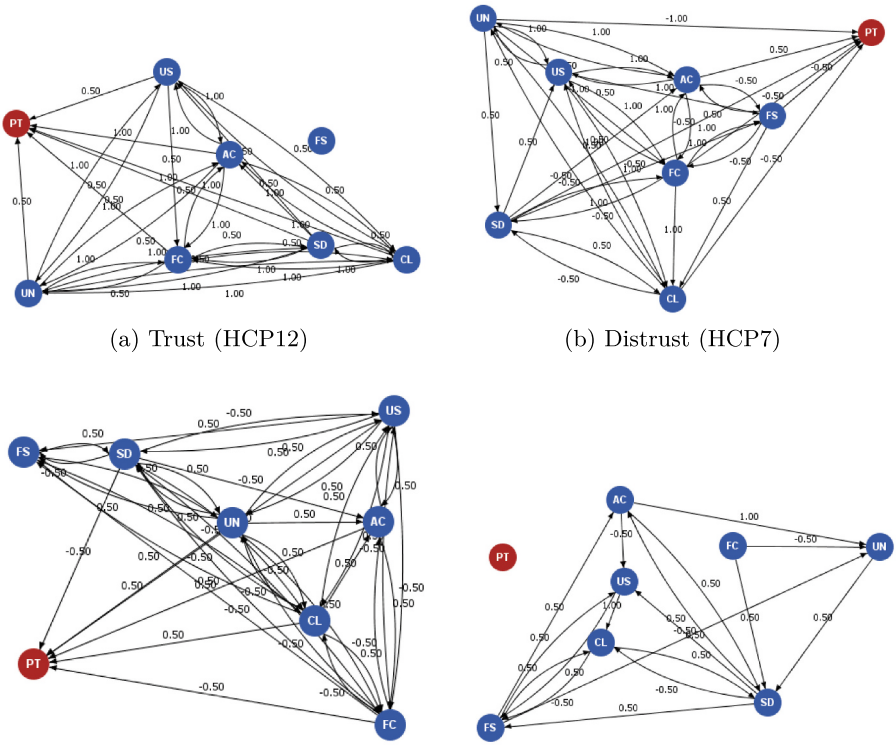


Fig. 6. Semantic representation of mental models (FCMs) of four HCPs representing trust, distrust, neutrality, and unknown behavior.

stance towards the model, and HCP11 displays an untrusting disposition. Also, the quantified value of trust for HCP4 is unknown, as depicted in Fig. 6d, and we exclude it in our future analysis. While the remaining HCPs express trust in the model, the extent of trust varies among them.

5 Validation of the Quantified Perceived Trust

Formal validation of FCMs is challenging due to their subjective nature. The difficulty lies in the fact that FCMs represent different interpretations of the system, and assessing their accuracy requires comparing them against yet another interpretation of reality [29]. To achieve this, we adopted the approach outlined by Schmidt and Biessmann [34], who introduced a metric to quantify trust by incorporating the concept of mutual information. However, Miller [22] believes that the metric primarily measures the agreement of users with the ML model's recommendations rather than trust. So, we have adjusted the terminology to refer to this metric as DoA. This metric measures the shift of diagnostic advice

among HCPs after interacting with the IML model, indicating reliance on it. To do so, using the information collected in Sect. 4.3, the mutual information between the IML model recommendation and the diagnostic advice of each HCP after interacting with it is measured using the following formula:

$$I(\widehat{y}_{IML}, \widehat{y}_{HI}) = \sum_{\widehat{y}_{IML}, \widehat{y}_{HI}} p(\widehat{y}_{IML}, \widehat{y}_{HI}) \log_2 \frac{p(\widehat{y}_{IML}, \widehat{y}_{HI})}{p(\widehat{y}_{IML}) p(\widehat{y}_{HI})} \tag{2}$$

In Eq. 2, the result is measured in bits. Similarly, $I(\widehat{y}_{GT}, \widehat{y}_{HE})$ indicates the mutual information between the ground truth status and HCP's diagnostic advice based on their expertise. Hence, the following equation can be used to measure DoA of each HCP with IML recommendations:

$$DoA = \frac{I(\widehat{y}_{IML}, \widehat{y}_{HI})}{I(\widehat{y}_{GT}, \widehat{y}_{HE})}, \tag{3}$$

where $DoA < 1$ represents that HCP does not have a high agreement with the model's recommendation and prefers to rely on its own expertise. $DoA > 1$ shows the HCP relies on the model's recommendation. The perfect agreement between the HCP and the IML model is established when the $DoA = 1$. The measured DoA for all HCPs is presented in DoA column of Table 6, showing eight HCPs rely on their expertise ($DoA < 1$). This tendency may be influenced by factors such as confirmation bias, general skepticism toward AI models, and the way IML rules are presented. HCPs found the rules lacking in completeness and accuracy, with no clear consensus on their sufficiency in terms of detail and functionality. While these factors are important and warrant further investigation to understand the underlying reasons for this behavior, they fall outside the scope of this research.

The results indicate that seven HCPs exhibit reliance on the model's diagnostic advice ($DoA > 1$), which is a sign of over-reliance on it. Possible factors contributing to this preference include the level of expertise, and general optimism toward AI. However, we acknowledge that further studies are needed to better understand the underlying causes of this over-reliance. The "Alignment" column of Table 6 evaluates whether both metrics lead to consistent conclusions regarding PT and DoA. Apart from three HCPs (3, 5, 15), all others adhere to the trust continuum pattern depicted in Fig. 1 and DoA. For instance, HCP3's PT is 0.9992, suggesting near-perfect perceived trust. However, during the diagnostic task, they relied on their own expertise ($DoA < 1$). These discrepancies may stem from inaccuracies in how their mental models were elicited. Weights and edges in an FCM reflect the subjective perspectives of HCPs, which is an advantage because it incorporates domain-specific expertise, but it is a limitation due to its inherent subjectivity. This highlights the need for further analysis in future studies.

Finally, we further analyze the obtained results by calculating the Pearson correlation between PT and DoA in Table 6. The correlation coefficient is 0.6851,

indicating a moderate to strong positive correlation between the two metrics. Thus, FCM demonstrates a high level of confidence in quantifying the perceived trust of HCPs, aligning closely with their propensity to adjust diagnostic advice after interaction with the IML model.

6 Discussion and Conclusions

This study introduces a novel methodology to measure the perceived trust of HCPs in interpretable clinical decision support by eliciting their mental models. Our findings suggest that while clinical decision support can somewhat influence HCPs' diagnostic advice, its impact is limited. Additionally, HCPs did not find interpretations very useful for diagnosing diseases; instead, they were more helpful in implementing them in the diagnostic task. This finding resonates with Jin *et al.*'s [17] conclusion that existing XAI algorithms often fall short of meeting clinical needs. The study validates the quantified perceived trust obtained via FCMs by comparing it with the DoA measure, which shows in most cases, both metrics converge to the same conclusion about the behavior of HCPs. Finally, the moderate to strong correlation between perceived trust and DoA suggests that FCM can effectively measure HCPs' perceived trust.

Our developed methodology is applicable across all realms in which domain experts are accessible. The pivotal aspect of this research lies in identifying key components that contribute to trust establishment within the domain of interest. This can be achieved by involving experts to pinpoint the principal elements of their trust. The strength of the proposed methodology lies in its ability to model the trust mechanisms of participants and reflect their subjectivity. Leveraging the high interpretability of FCMs, we can detect crucial aspects contributing to participants' trust refinement. Subsequently, this understanding enables us to refine and improve the IML model to increase trust. Furthermore, FCMs offer the flexibility to be updated or modified based on new information or changes in the system, allowing for continuous refinement and improvement.

This study is limited by its small sample size, which categorizes it as a pilot study and potentially renders the results statistically unreliable. While a larger sample might reveal a greater discrepancy between PT and DoA, it is important to note that the core contribution of modeling perceived trust through FCM remains unaffected. Because FCM is a subjective model grounded in each HCP's mental model, the quantified perceived trust precisely reflects what HCPs report and how they conceptualize their mental models due to the mathematical basis of FCM is robust. Despite these limitations, the insights gained—particularly from user studies involving HCPs—remain valuable. Trust, though central to this study, is a nuanced concept encompassing multiple facets beyond the scope of ESS alone. Future research will involve participants in identifying and articulating these broader trust elements.

Acknowledgments. I. Grau and C. Zhang are supported by the European Union's HORIZON Research and Innovation Program under grant agreement No. 101120657, project ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI).

Disclosure of Interests. All authors declare that they have no conflicts of interest.

References

1. Abbaspour Onari, M., Jahangoshai Rezaee, M.: Implementing bargaining game-based fuzzy cognitive map and mixed-motive games for group decisions in the healthcare supplier selection. Artif. Intell. Rev. 1–34 (2023)
2. Arun, N., Mohan, B.: Modeling, stability analysis, and computational aspects of some simplest nonlinear fuzzy two-term controllers derived via center of area/gravity defuzzification. ISA Trans. **70**, 16–29 (2017)
3. Axelrod, R.: The cognitive maps of political elites (1976)
4. Bansal, G., et al.: Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–16 (2021)
5. Beena, P., Ganguli, R.: Structural damage detection using fuzzy cognitive maps and Hebbian learning. Appl. Soft Comput. **11**(1), 1014–1020 (2011)
6. Chanda, T., et al.: Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. Nat. Commun. **15**(1), 524 (2024)
7. Cho, J.H., Chan, K., Adali, S.: A survey on trust modeling. ACM Comput. Surv. (CSUR) **48**(2), 1–40 (2015)
8. Cohen, W.W.: Fast effective rule induction. In: Machine Learning Proceedings 1995, pp. 115–123. Elsevier (1995)
9. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
10. Regulation (eu) 2016/679 (general data protection regulation) (2016). https://eur-lex.europa.eu/eli/reg/2016/679/oj
11. Ethics guidelines for trustworthy AI (2019). https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
12. González-Gonzalo, C., et al.: Trustworthy AI: closing the gap between development and integration of AI systems in ophthalmic practice. Prog. Retin. Eye Res. **90**, 101034 (2022)
13. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Measures for explainable AI: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. Front. Comput. Sci. **5**, 1096257 (2023)
14. Huber, T., Weitz, K., André, E., Amir, O.: Local and global explanations of agent behavior: integrating strategy summaries with saliency maps. Artif. Intell. **301**, 103571 (2021)
15. Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., Baesens, B.: An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decis. Support Syst. **51**(1), 141–154 (2011)
16. Jacovi, A., Marasović, A., Miller, T., Goldberg, Y.: 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, pp. 624–635 (2021)
17. Jin, W., Li, X., Hamarneh, G.: Evaluating explainable AI on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements? In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 11945–11953 (2022)

18. Joshi, R., Graefe, J., Kraus, M., Bengler, K.: Exploring the impact of explainability on trust and acceptance of conversational agents–a wizard of OZ study. In: International Conference on Human-Computer Interaction, pp. 199–218. Springer (2024)
19. Kosko, B.: Fuzzy cognitive maps. Int. J. Man Mach. Stud. **24**(1), 65–75 (1986)
20. Lakkaraju, H., Bastani, O.: "How do I fool you?". manipulating user trust via misleading black box explanations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 79–85 (2020)
21. Mehrotra, S., Jorge, C.C., Jonker, C.M., Tielman, M.L.: Integrity-based explanations for fostering appropriate trust in AI agents. ACM Trans. Interact. Intell. Syst. **14**(1), 1–36 (2024)
22. Miller, T.: Are we measuring trust correctly in explainability, interpretability, and transparency research? CHI TRAIT Workshop (2022)
23. Miller, T.: Explainable AI is dead, long live explainable AI! hypothesis-driven decision support using evaluative AI. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 333–342 (2023)
24. Nápoles, G., Espinosa, M.L., Grau, I., Vanhoof, K.: FCM Expert: software tool for scenario analysis and pattern classification based on fuzzy cognitive maps. Int. J. Artif. Intell. Tools **27**(07), 1860010 (2018)
25. Nápoles, G., Grau, I., Bello, R., Grau, R.: Two-steps learning of fuzzy cognitive maps for prediction and knowledge discovery on the HIV-1 drug resistance. Expert Syst. Appl. **41**(3), 821–830 (2014)
26. Nauta, M., et al.: From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. ACM Comput. Surv. **55**(13s), 1–42 (2023)
27. Onari, M.A., et al.: Comparing interpretable AI approaches for the clinical environment: an application to Covid-19. In: 2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 1–8. IEEE (2022)
28. Onari, M.A., Yousefi, S., Rezaee, M.J.: Risk assessment in discrete production processes considering uncertainty and reliability: Z-number multi-stage fuzzy cognitive map with fuzzy learning algorithm. Artif. Intell. Rev. **54**(2), 1349–1383 (2021)
29. Özesmi, U., Özesmi, S.L.: Ecological models based on peoples knowledge: a multi-step fuzzy cognitive mapping approach. Ecol. Model. **176**(1–2), 43–64 (2004)
30. Papageorgiou, E.I.: A new methodology for decisions in medical informatics using fuzzy cognitive maps based on fuzzy rule-extraction techniques. Appl. Soft Comput. **11**(1), 500–513 (2011)
31. Pedrycz, W.: Fuzzy control and fuzzy systems. Research Studies Press Ltd. (1993)
32. Perlmutter, M., Gifford, R., Krening, S.: Impact of example-based XAI for neural networks on trust, understanding, and performance. Int. J. Hum Comput Stud. **188**, 103277 (2024)
33. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
34. Schmidt, P., Biessmann, F.: Quantifying interpretability and trust in machine learning systems. In: AAAI-19 Workshop on Network Interpretability for Deep Learning (2019)
35. Sivaraman, V., Bukowski, L.A., Levin, J., Kahn, J.M., Perer, A.: Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–18 (2023)

36. Ueno, T., Sawa, Y., Kim, Y., Urakami, J., Oura, H., Seaborn, K.: Trust in human-AI interaction: scoping out models, measures, and methods. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts, pp. 1–7 (2022)
37. Vasey, B., et al.: Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: decide-AI. Nat. Med. **28**(5), 924–933 (2022)
38. Vereschak, O., Bailly, G., Caramiaux, B.: How to evaluate trust in AI-assisted decision making? a survey of empirical methodologies. Proc. ACM Hum. Comput. Interact. **5**(CSCW2), 1–39 (2021)
39. Wang, X., Yin, M.: Are explanations helpful? a comparative study of the effects of explanations in AI-assisted decision-making. In: 26th International Conference on Intelligent user Interfaces, pp. 318–328 (2021)
40. Wysocki, O., et al.: Assessing the communication gap between AI models and healthcare professionals: explainability, utility and trust in AI-driven clinical decision-making. Artif. Intell. **316**, 103839 (2023)
41. Yang, F., Huang, Z., Scholtz, J., Arendt, D.L.: How do visual explanations foster end users' appropriate trust in machine learning? In: Proceedings of the 25th International Conference on Intelligent user Interfaces, pp. 189–201 (2020)
42. Zhang, Y., Liao, Q.V., Bellamy, R.K.: Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 295–305 (2020)