IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 July 2024, accepted 19 August 2024, date of publication 3 September 2024, date of current version 25 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3454061

RESEARCH ARTICLE

POTDAI: A Tool to Evaluate the Perceived Operational Trust Degree in Artificial Intelligence Systems

DAVID MARTÍN-MONCUNILL[1], (Member, IEEE), EDUARDO GARCÍA LAREDO[2], (Member, IEEE), AND JUAN CARLOS NIEVES[3]

[1]CAILab, Universidad Camilo José Cela, Madrid, 28692 Villafranca del Castillo, Spain
[2]Faculty of Health Sciences, International University of La Rioja (UNIR), 26006 Logroño, Spain
[3]Department of Computing Science, Umeå universitet, 901 87 Umeå, Sweden

Corresponding author: David Martín-Moncunill (david.martinm@ucjc.edu)

This work was supported in part by the HumanE AI Network (HumanE-AI-Net) Research Project, which has received funding from the European Union's Horizon 2020 Research and Innovation Program under Grant 952026 and the partners leading the microproject "To develop a trustworthy AI model for situation awareness by using mixed reality in Police interventions.": Umeå University - Computing Science Department, Umeå University - Police Education Unit, Comet Global Innovation S.L. and Institut de Seguretat Pública de Catalunya -ISPC.

ABSTRACT There is evidence that a user's subjective confidence in an Artificial Intelligence (AI)-based system is crucial in its use, even more decisive than the objective effectiveness and efficiency of the system. Therefore, different methods have been proposed for analyzing confidence in AI. In our research, we set out to evaluate how the degree of perceived trust in an AI system could affect a user's final decision to follow AI recommendations. To this end, we established trustworthy criteria that such an evaluation should meet by following a co-creation approach with a multidisciplinary group of 10 experts. After a systematic review of 3,204 articles, we found that none of the tools met the inclusion criteria. Thus, we introduce the so-called "Perceived Operational Trust Degree in AI" (POTDAI) tool that is based on the findings from the expert group and the literature analysis, with a methodology that adds rigor to that employed previously to create similar evaluation tools. We propose a short questionnaire for quick and easy application, inspired by the original version of the Technology Acceptance Model (TAM) with six Likert-type items. In this way, we also respond to the need pointed out by authors such as Vorm and Combs to extend the TAM to address questions related to user perception in systems with an AI component. Thus, POTDAI can be used alone or in combination with TAM to obtain additional information on its usefulness and ease of use.

INDEX TERMS Artificial intelligence, cooperative systems, human–computer interaction, human factors, trustworthy AI, technology acceptance model.

I. INTRODUCTION

Users' Trust in Artificial Intelligence (AI), which is one of the crucial factors that can influence the adoption and use of any system employing AI, has received considerable attention in recent years [1]. Consequently, many academic and industrial researchers have pointed out the need to develop qualitative and quantitative methods for evaluating the degree of trust in interaction studies using artificial intelligence-based tools [2], [3], [4].

In this regard, we must start highlighting the need to distinguish between the "real trustworthiness" – associated with the reliable statistical data that shows hit rate–and any other measurable aspects related to the effectiveness and efficiency linked to the reliability of the system – of the AI and the degree of trust perceived by the end user. This need cannot be ignored because evidence shows that both can differ significantly, influencing how the user interacts with an AI [5]. One can highlight that the statistical data that establish this "*real trustworthiness"* present two main problems:

First, factors such as hit rate are not always easy to quantify. There are use cases where verifying the hit rate is clear, such as an AI artifact that indicates whether a spot on the skin is

The associate editor coordinating the review of this manuscript and approving it for publication was Hiram Ponce.

a melanoma [6], [7], which can be verified in a laboratory, or an AI that can identify traffic signals [8]. However, use cases such as evaluating the quality of an image generated by artificial intelligence [9], a chatbot that generates the biography of a celebrity [10], or the accuracy rate in extracting keywords from a text [11], do not allow for precise verification of the hit rate, as in the previous examples. In such use cases, it is necessary to establish a ''gold standard,'' which is usually done either by comparing the performance of human experts with that of an AI when performing the same task, or by directly evaluating the results produced by each of them. These methods are not exclusive and can be used to set a gold standard.

Second, the same hit rate may be unacceptable in different use cases. AI with an 85% hit rate for identifying plates in a private parking space may be acceptable. The same percentage would be disastrous in an application that manages track changes of a train. Similarly, every individual may perceive trust differently when informed of the ''real trustworthiness'' of the system. Thus, circumstances, such as different quality expectation levels or fear of technology [8], can influence this perception, resulting in a discrepancy between perceived and actual trustworthiness. This phenomenon can occur with any type of system regardless of whether it employs artificial intelligence. However, the use of artificial intelligence introduces a distinct context that warrants further investigation.

If a user receives an incorrect answer from a system that does not use AI to make automated decisions, it is likely to be the result of the user's misuse, assuming that the system is working correctly and there is no bug or hardware failure. Hence, the system may fail; however, it is not wrong in a strict sense. Nor can it learn, and the reason for this result is always transparent. An AI that works correctly may produce incorrect results, but the reason for this result is not always clear [12].

Excess or lack of trust can make interactions with AI problematic. For example, a lack of trust can cause the user to constantly question and analyze whether an AI response is appropriate, resulting in a loss of efficiency [13]. In extreme cases, a lack of trust (*mistrust*) can lead the user to do the exact opposite of what an AI suggests [10]. However, an excess of trust (*overtrust*) may lead the user to relax and not pay the necessary attention required by the AI system [10]. Depending on the AI labor, the consequences of inappropriate use can range from an occasional mishap – a student doing his homework incorrectly – to a catastrophic situation – a car accident [14]. Because the user is aware of the consequences of these decisions, it is logical to conclude that their impulse to check that the AI response is appropriate may differ according to the different contexts of use.

One can argue that the perception of trust is not just another factor to be analyzed in an interaction or usability study, but also a key element that defines the user and must be considered before other types of testing are carried out. We believe that, in the same way that the level of experience that a user has with a system influences the results of its use – effectiveness, efficiency, etc. [4] – so does the level of user's trust in the system. It is human nature not to act in the same way as what we trust and towards what we do not trust [15]. Therefore, the outcomes of using a system should vary according to the user's perception of trust in the system.

From the above, it is clear that both factors characterize the user and must be assessed before moving on to other types of testing. Furthermore, both factors were modified in a symbiotic manner using this system. The more a user uses a system, the more experience he acquires. This experience gives the user a clearer idea of the system's capabilities, which will influence the level of trust. Similarly, if AI has learning capabilities, it can increasingly refine its actions [16] and positively modify users' trust perception. User trust can vary according to the results of previous interactions with the system or the emotional state of the subject. We must accept that trust perception is not necessarily stable and may even fluctuate drastically during an interaction with a given system [3].

The research reflected in this article is related to the ''HUMANE-AI'' project funded by the European Commission, and the sub-project ''AI Ethics and Responsible AI'' dedicated to developing trustworthy AI models for situation awareness by using mixed reality in police interventions. Our project is situated in a context where the consequences of IA use could lead to serious circumstances, so it is essential to appropriately manage users' trust. Our interest was to assess how a human operator's perceived level of trust in a particular AI system affects his or her decision to follow an AI's recommendations. To do this, we first developed a list of requirements with a multidisciplinary team of ten experts (Table 1) and systematically reviewed the literature to find an AI trust assessment tool or methodology suitable for those requirements.

As detailed in the epigraph 'Methods of trust assessment' in the Results section, we were unable to find an assessment system that met our requirements. This literature review also revealed that there is no clear definition of trust, with authors distinguishing between several types, such as cognitive trust, affective trust [17], [18], [19], [20], [21], trust propensity or dispositional trust [20], [21], history-based trust, situational trust, procedural trust, [20] and institutional trust [22], in which different authors identify a large number of factors associated with their view of trust [2], [4], [15].

The existence of these different types of trust and the fact that none of them completely fulfilled our objective of study led us to define the concept of ''Perceived Operational Trust Degree in AI'' (POTDAI). This, as mentioned above, focuses on whether the user tends to follow the indications of an AI, without going into the reasons for this, and is determined based on three factors: *overtrust*, *mistrust* and *monitoring*.

As our objective is focused on the perception of trust, we considered using the TAM model [23], [24], as it is a consolidated tool linked to the perception of usefulness and ease of use. We also found that, as confirmed by Vorm and Combs [25], there is no proposed version of this model that considers the aspects of technologies that use AI. However,

in its basic version, we found the TAM model to be useful in meeting several of the requirements set for our evaluation. Based on the above, we extended the TAM model with six additional questions to enable us to conduct our evaluation. In summary, our study provides an overview of the trust assessment methods for AI. We define the concept of ''Operational Trust in AI'' and provide a method for its evaluation that can be integrated into the TAM or used independently.

II. METHODOLOGY

The first step was to elicit the key requirements for the assessment tool to be designed. To this end, a group of ten experts involved in our ''HUMANE-AI'' project established their conclusions through two focus group sessions. Following which was stated by Goguen and Linde [26], about the use of focus groups in the requisites gathering context, the experts – listed in in Table 1 below – representing different knowledge areas and professional interests assembled in an informal discussion group format and a facilitator elicited views on issues and concerns about the key assessment tool features.

TABLE 1. Identification and characteristics of the experts.

ID	Main Career (Degree)	Knowledge area(s)
1	Computer Science	AI Education Research, Trustworthy AI
2	Cognitive Science	Software development, Human Computer Interaction (HCI)
3	Police Science	Police work, police education and research
4	Police Science	Police work and police education
5	Police Science	Police work and police education
6	Public health	Police education, Research
7	Police Science, Criminology Science	Research
8	Computer Engineering	Human computer interaction, usability, UX, Knowledge Organization Systems
9	Psychology	Neuroscience, Research Methodology, HCI
10	Law	Innovation, R&D, Lawtech

The experts started from the basic premise presented in the introduction: to be able to assess how the level of trust perceived by the user will lead them to act according to the indications provided by an AI system. It was emphasized that the aim was not to delve into the whys and wherefores of the decision, but rather the user's propensity to act or not in accordance with the AI – this approach is the reason why the name ''Perceived Operational Trust Degree in AI'' was chosen for the assessment tool.

Given that the group of experts was tasked with developing a system for assessing the perception of trust in AI, which would later be used in the police field, it was deemed that the most pertinent bias risk would be the potential to influence the results yielded by the application of POTDAI. This could result in the experts' developing questions that would favor their positions on the use of AI to support police tasks. We also detected as a potential that experts may be inclined to include questions related to their research or professional interests, rather than following the ''Operational Trust'' approach.

Although both events were considered unlikely to materialize, in order to avoid bias, efforts were made to (i) compose a multidisciplinary group, with experts from different professions and fields, as shown in Table 1; (ii) verify that the experts didn't have any type of direct interest (economic, political or of any other nature) related to the results of the study; (iii) ascertain that there were no links that would provide future benefits to third parties and (iv) start by establishing a clear set of requisites, as shown in Table 2.

TABLE 2. Requirements for an appropriate subjective AI trust assessment tool set by the experts.

Req. ID	Description
|R1|	Be suitable for situations where AI decisions could have a critical impact.
|R2|	The assessment tool should be useful for evaluating systems from their earliest stage, exposing the project through a time-limited demonstration with non-functional prototypes.
|R3|	Assess whether the user is predisposed to behaviors that lead to systematic under or over reliance on AI recommendations.
|R4|	Assess the role of user self-monitoring. This will enable an understanding of whether the user will feel uncomfortable because of the AI system continuously evaluating their performance, which may affect their decision-making.
|R5|	Assess the overall impression that potential users would have. This should be done by focusing on whether they perceive the usefulness of the system, whether they consider that use would lead to improvements in effectiveness and efficiency, and the difficulty of learning to use the system.

From Table 2, three main factors were identified to assess the user's propensity to act on AI indications. The first two factors, overtrust and mistrust, are related to the user's trust in the appropriateness of the indications provided by the AI and their ability to analyze them in a way that does not generate excessive dependence or mistrust [10], [13]. The third factor, which we call monitoring, shows if the user can be affected by the feeling that the AI is evaluating his work, both when the user must decide whether to follow the AI's indications and globally when the system provides no feedback at all. The first two factors are covered by the Requirement |R3|, whereas the third is covered by |R4|.

|R1| and |R2| are linked to the conditions of the context of use (critical situations, early evaluation) in which the tool must be able to operate, while |R5| seeks a general evaluation of the user's perception of the system. In addition, it was determined that the assessment tool should be user-friendly, easy to understand, require a short time to complete, and include a mechanism that will support the avoidance or detection of incomplete questionnaires has been filled incorrectly.

Considering the previous requirements, a review of the existing literature between 2019 and 2023 was conducted to find qualitative methods of trust assessment that could be applicable to our research project or, alternatively, recommendations that would allow us to reach its goals based on

the concepts and general ideas common to the definitions and assessments used in other studies.

The following databases were used to search for information on Web of Science (WOS) / SCOPUS / IEEE Xplore / ACM Digital Library / PubMed (although it is a medical database, the impact of AI trust is particularly relevant in the healthcare world, so it must be considered). The keywords used in the database searches were: (''user trust'' OR ''user confidence''[1] OR ''user perception'' OR ''user evaluation'' OR ''trust evaluation'' OR ''confidence evaluation'' OR ''perception evaluation'') AND ''Artificial Intelligence.'' The search for potential articles included the following steps.

1. Search for potential articles: Enter KEYWORDS (after adjusting the search criteria in the selected databases) and import the articles' metadata to the reference manager returned by each search (Table 3).
2. Screening by abstracts: The abstracts of the potentially selected articles were read to determine if they met the specified criteria.
3. Download and read the articles that met the selected criteria by abstract.
4. Final inclusion or non-inclusion in the study was determined according to specified criteria after reading the full text. The search process is detailed in Figure 1, including the results for each stage.

TABLE 3. Number of results found per database.

DATABASE	TOTAL
WOS	171 results
SCOPUS	265 documents found
IEEE Xplore	2078 results
ACM Digital Library	831 Results
Pubmed	6 Results
ALL	3.351 RESULTS

As depicted in Figure 1, 3,351 articles were identified, and the initial set excluding duplicates comprised 3,204 studies. Figure 1 also shows how, after reviewing the abstracts against the inclusion and exclusion criteria, 71 potential articles remained for full reading. The next step was a full reading of 71 papers. After this, 14 papers were withdrawn due to disagreement between two independent reviewers, leaving a total of 57 final papers, of which 33 provided scales for subjective assessment of trust in AI and a further 24 provided schemes and theories for their analysis (see Figure 2).

The papers that did not specifically address user trust evaluation (2548) covered a wide variety of topics (such as business community attitudes towards AI technologies, potential future research directions and topics, theoretical aspect clarification, etc.) but in general, they were primarily focused on studying aspects of the development and outcome of AI in various fields such as: the creation of new AI, improvements in AI algorithms, improvements in the use of data, signaling error rates committed by AI, solving problems in AI training, new AI for application to clinical and academic environments, implementing improvements in various aspects (such as navigation and/or security), applications to unmanned aerial vehicles (UAVs), improving the accuracy in the handling of vocabulary and languages, and so forth.

Although the abstracts of the articles mentioned trust, 20.2% of them (130) addressed the topic in a secondary manner and did not provide clear or useful information regarding its evaluation or conceptualization. In general, those articles dealt on various aspects of technology development, calibration and evaluation of its use, giving minimal attention to trust perception but focusing on methods for facilitating explanations during task performance, feedback, and strategies for enhancing security and privacy.

A total of 51.5% of the evaluated studies focused on user subjective and perceptual aspects, yet did not consider the role of trust in these evaluations (331 papers). The majority of studies addressed the various methods of evaluating and recognizing user emotions, with some employing a more generic approach and others focusing on specific applications, such as suicide prevention. Additionally, many studies explored subjective aspects influenced by cultural or media factors.

Nineteen percent of the studies evaluated trust but did so using automated methods or expert criteria (122 papers) from a reliability point of view. This was in regard to user attitudes towards the use of algorithms, focusing on the degree of success in executing tasks and on aspects of the frequency of use of AI; not from the perspective of perception. Some papers (3.7% of the total, 24 papers) discussed the necessity of evaluating user trust but did not present methods for doing so. Two of the potential papers (0.3%) were withdrawn by the journals as the publishers found that they uncovered evidence of systematic manipulation of the publication process. Finally, only 5.1% of the total provided information on the scales for the subjective evaluation of user trust (33 papers) (Table 4 and Figure 3).

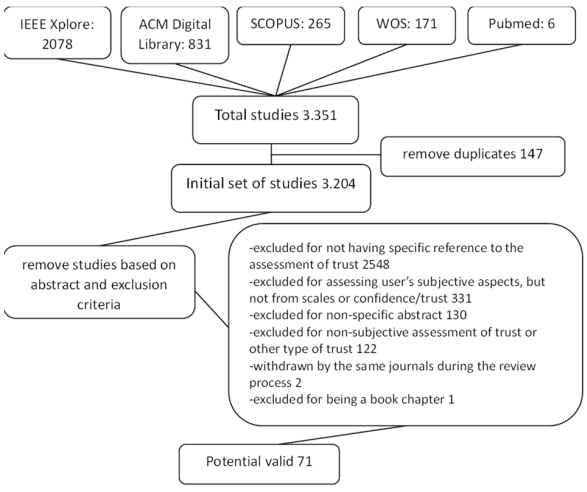


FIGURE 1. Diagram of the search for potential articles.

[1]Many authors employ the term ''confidence'' instead of ''trust.''

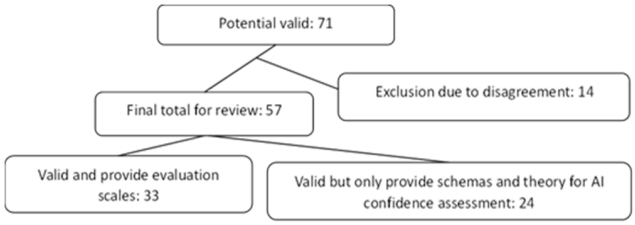


FIGURE 2. Selection of definitive papers.

TABLE 4. Classification of potential papers.

	Freq.	Percent.
Excluded for evaluating subjective aspects, but not from the concept of trust.	331	51,5
Excluded for having an abstract that did not provide clear information about its evaluation or conceptualization.	130	20,2
Excluded for assessing trust based on automated methods or expert criteria.	122	19,0
Withdrawn by the same journals during the review	2	0,3
Excluded for being a book chapter.	1	0,2
Valid: provide theory on subjective trust	24	3,7
Valid: provide methods for assessing subjective trust	33	5,1
Total	643	100,0

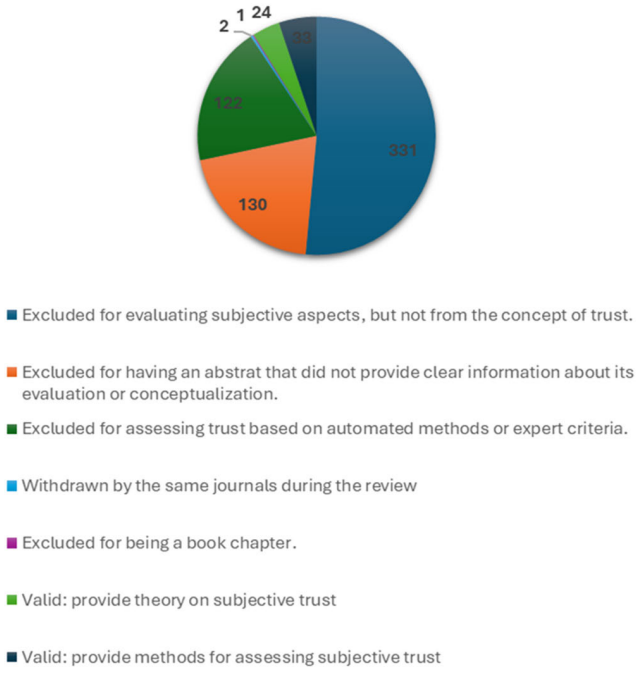


FIGURE 3. Classification graph of the possible documents.

As detailed in the results section, after analyzing the literature, we decided to start from the elementary Technology Acceptance Model (TAM) model [23], [24] and extend it with six additional questions to assess trust perception in AI, following the method proposed by Vorm and Combs [25]. That made a total of 18 questions: 6 for "perceived usefulness," 6 for "perceived ease of use" and 6 for the "perceived operational trust in the AI." This approach makes it possible to extract the information reflected in |R5| while accomplishing |R1| and |R2|. At this point, the experts focused on requirements related to overtrust/mistrust |R3| and monitoring |R4|.

A card-sorting approach [27] was chosen to define six questions on perceived operational trust. We found this technique suitable, as it allows different types of elements – in our case, questions – to be classified visually, and the board can be used to discuss and refine the questions. Categories were created based on the main factors previously identified as "overtrust," "mistrust, and "monitoring." The experts placed their questions in different categories; if a question covered several categories, it was placed in each category. Once the questions were positioned, they were refined and formulated, considering other decisions of a global nature that arose during the discussion. This process is described in detail in the Results section.

III. RESULTS

Our review of the existing literature indicates that studying the evaluation of users' subjective trust in AI presents a number of challenges. The studies reviewed offer varying definitions and classifications of "trust." There is no clear definition of the term "trust," and it is often presented in a vague or ambiguous manner, or it is only mentioned, leaving it open to interpretation by the reader. Furthermore, the concept of trust is subject to significant variation depending on the field of research and the characteristics of each study.

Trust is a complex construct, comprising numerous factors from both the user and the technology itself. These include the user's knowledge, experience, emotions, age, beliefs, as well as the technology's performance, usability, security, and explainability. Additionally, the purpose and context of use, including external user variables and socio-ethical considerations, must be considered. This multidimensional aspect is not sufficiently considered in many of the studies. Furthermore, there is no consensus on the specific components that should be included, the structure of these components, or the factors that could influence them.

A. USER'S TRUST ASSESSMENT TOOLS

Based on our literature review, we defined user trust as a construct – a subjective and complex process – that can be assessed with subjective measures (questionnaires or interviews) or behavioral measures (behavioral observation and task-based metrics) [5]. It is interesting to note that recent work has explored the use of psychophysiological measures, such as heart rate variability, galvanic skin response, or electroencephalography, to assess trust between humans and AI [28].

Of all these forms of assessment, questionnaires are the most common method for measuring trust because, despite their limitations, they can obtain important information and

have certain advantages such as (i) they allow capturing a person's attitude, what they feel, and what they think; (ii) participants may feel more comfortable reporting their psychological state because they are not in front of another person, but only in front of a screen or sheet of paper, and it is less invasive than taking sensor measurements on the individual; and (iii) they are relatively easy and quick to administer, allowing for the collection of a larger amount of data in a shorter period of time, compared to other means such as interviews and observations [28].

Among the 33 articles that provided scales for the assessment of subjective perception of user trust, there was wide variability in the forms of assessment used. In our review, we observed two cases where the assessment of trust was supplemented by an open-ended interview with the user [3], [29]. Three other studies discussed several potential scales without specifying their items and scores [3], [5], [28].

In other articles, the evaluation of trust in AI was limited to very precise domains, highlighting one case focused on online shopping [15], two studies on the use of ChatGPT [10], [30], three on the degree of trust in AI evaluations in the field of healthcare (one on the detection of serious pathologies such as cancer [29], one on trust in the use of AI in surgery [31], and another on clinical decision making [32]), two studies on the extent to which a user believes that his or her initial trust can be improved by explanations provided by the system itself (XAI) [33], [34], and another assessing aspects related to ethics rather than effectiveness of use [35].

We found three studies that approached their trust assessments by creating their own tests using items from other scales [1], [18], [32], one that extended existing assessments with their own items [36]), and two others that modified items from previous scales to adapt them to the purpose of their research [31], [37]. It is also common for item batteries to be created ad hoc according to the same authors' criteria (six cases: [3], [18], [19], [30], [38], [39]), with the assessment in two cases being a single direct question about the extent to which the AI was trusted, with a single response on a Likert scale [29], [40]. In a relevant amount of cases (13 studies), trust assessment was only part of a larger evaluation focused on measuring different aspects of user experience with AI [1], [5], [15], [18], [19], [22], [30], [31], [32], [33], [35], [38], [39].

While there were instances where the authors developed their own tests, the most prevalent finding in the reviewed studies was that they employed scales that had been previously designed by other authors. The most frequently cited scales are presented in Table 6.

B. HOW PREVIOUS USER'S TRUST ASSESSMENT TOOLS WERE CONSTRUCTED

Although almost all definitions of trust revolve around the balance of risks and benefits that a user assesses [10], one of the aspects that becomes clear when reviewing the various studies on the topic is that there is no single definition of trust [8], [19], with the result that different authors define and assess trust in different ways. It is common for studies to take one or more of the various definitions of trust offered by different fields as a starting point. We found definitions based on ISO standards, EU Guidelines on Trustworthy AI, definitions of trust from the social sciences, and even those found in dictionaries such as the Oxford Dictionary [4]. Similarly, it is very common to see different factors, such as user knowledge, technical competence, familiarity or even beliefs, faith, emotions, and personal attachments included in the conceptualization of trust, according to the objectives of each study.

Once a definition of user trust has been proposed, the authors indicate the factors that they consider relevant according to the literature or their own criteria. On this basis, they proceeded with the evaluation, either by using their own methods or by using some or all existing evaluations that fit the proposed definitions and factors. Just as trust can be defined in different ways, it can also be assessed in different ways. In general, because of its subjective nature, it is difficult to directly identify and measure the causes that determine the level of trust; therefore, the most used forms of evaluation are subjective measures. These measures consist of exploring a user's opinions before, during, and/or after working

TABLE 5. Forms of subjective trust evaluation in the review papers.

Complemented the questionnaires with an open interview with the user: 2	[3], [29].
Mentioned several scales without specifying their items and scores: 3	[3], [5], [28].
The trust evaluation was focused on very specific aspects:	
Online shopping usage: 1	[15].
ChatGPT usage: 2	[10], [30].
Healthcare usage: 3	[29], [31], [32].
XAI usage: 2	[33], [34].
Ethical issues: 1	[35].
They made their own tests from items of other scales: 3	[1], [18], [32].
Expanded existing assessments with their own items: 1	[36].
Modified the items of previous scales to adapt them to the purpose of their research: 2	[31], [37].
Ad hoc batteries developed from their own criteria: 6	[3], [30]. [18], [19], [38], [39].
The assessment was a single question to be responded to on a Likert scale: 2	[29], [40].
Trust evaluation was a part of a user experience evaluation: 13	[1], [5], [15], [30], [31], [32], [33], [35], [18], [19], [38], [39], [22].

TABLE 6. Scales most cited in the review and components of the scales.

Scale	Scale components
"Trust in Automation Questionnaire (TiA)" [41]	Presents 6 subscales with 19 items in total (Reliability/Competence with 6 items, Understanding/Predictability with 4 items, Familiarity with 2 items, Intention of Developers with 2 items, Propensity to Trust with 3 items and Trust in Automation with 2 items) with a Likert scale of response from Strongly disagree (1) to Strongly agree (5) .
"Checklist for Trust between People and Automation" [42]	12 items, presented on a 7-point Likert scale ranging from not at all (1) to all (7).
"Human-computer trust (HCT)" [43]	Composed of five constructs (Perceived Reliability, Perceived Technical Competence, Perceived Comprehensibility, Faith, Personal Attachment) which are each evaluated from 5 questions.
"Multidimensional Measure of Trust (MDMT)" [44] and [45]	version 1 with 16 items, assessing 4 dimensions divided into two trust factors: "Confidence of Ability" (reliable, capable) and "Moral Confidence" (ethical, sincere).
	version 2, with 20 items, which extends version 1 by evaluating five dimensions grouped into the factors: "Confidence in Performance" (subscales: reliable, competent) and "Moral Confidence" (subscales: ethical, transparent, benevolent).

with a system through interviews and questionnaires, usually Likert-type response questionnaires.

In this regard, Bach et al. [15] found that over two-thirds (69.56%) of the studies in their review employed ad hoc questionnaires to assess user trust. This demonstrates that questionnaires are the most prevalent method of measuring subjective user confidence, and that different authors tend to define ratings according to their own criteria.

Although, with the exception of the work of Park et al. [38], the vast majority of authors who use their own questionnaires do not explicitly indicate the stages of their development, we can observe that, in general, the following steps are followed: (i) a collection of documents, mostly research articles, are compiled from a set of keywords related to the concept to be evaluated; (ii) the researchers select the collected documents according to their relevance criteria; (iii) once the selection of documents is completed, the elements relevant to the concept to be evaluated are defined; and (iv) based on these definitions, the questions composing the survey are prepared based on the authors' criteria. In some cases experts were invited to analyze the suitability of the developed evaluation tool.

C. THE NEED TO BUILD A CUSTOM TOOL

In our review, we found that a wide range of assessments are available to measure trust. Subjective scales are considered the quickest and most convenient way to access subjects' thoughts and beliefs; not surprisingly, they are widely used for this purpose in psychology and psychiatry [46], and different authors use different criteria to conceptualize and represent trust, from which different trust assessment questions are derived. However, in our search for information, we did not find any tools that could be used directly to evaluate user confidence according to the requirements previously established by our working group and presented in the methodology section.

We found that, of the most used evaluations, none were oriented towards evaluating operational perception and were not optimal for cases where non-functional prototypes of the systems under development were available. Our aim is not to examine in depth the factors that may influence or define confidence but rather the ultimate consequences that the level of confidence may have during the operation of the system. As an illustrative example, Table 7 shows some of the problems we found with the most commonly used scales (based on our analysis of the literature) that prevented them from being adequate to meet our objectives.

TABLE 7. Main limitations in the existing scales to be used in our project.

Scale	Limitations
Trust in Automation Questionnaire (TiA)	-Presents trust in automation as a subscale. -It does not consider aspects related to ease of use and learning. -On the other hand, the questions of the Trust in Automation subscale items (item 9: I trust the system and item 14: I can rely on the system) can have a very open interpretation.
Checklist for Trust between People and Automation	Although it is a highly recommended scale and is generally one of those cited [49] , it should be noted that: -It was not designed to assess a subject's perception prior to interacting with an AI. -On the other hand, it does not provide items to assess the degree to which the system would facilitate a task in terms of learning. Furthermore, we believe that certain items, although correctly formulated (e.g., item 6: "I have trust in the system" and item 11: "I can trust the system"), could be more rigorously worded.
Human-computer trust (HCT).	Although this scale initially appeared to be consistent with our criteria, we have identified two significant limitations. - Firstly, the items in the scale, particularly those pertaining to the perceived reliability of the AI, often emphasize the idea that the AI is capable of behaving in the same way under the same conditions and at different times. However, this is not a consideration in a program under development. -Secondly, the scale does not assess fear or rejection of the AI, which are detrimental to the use of an AI and therefore require assessment. -In contrast, although the scale includes items to assess overtrust, it lacks items to assess fear or rejection of the AI. Our criteria indicate that both undertrust and overtrust are detrimental in the use of an AI, and therefore both need to be assessed.
MDMT: Multi-Dimensional Measure of Trust (version 1 y 2)	-Both versions of the test do not pose direct and detailed questions. Instead, they present a series of single words, which the user must indicate their level of agreement with regarding their thoughts on trust in AI. In this sense, the test asks about terms such as ethical, competent, sincere, etc., which it does not define or specify, thus limiting the degree of interpretation of the results.

In response to this situation, we decided to develop our own evaluation tool, which we have named "POTDAI" (**P**erceived **O**perational **T**rust **D**egree in **AI**).

D. CREATION OF POTDAI AND CONSTRUCTION OF THE INTEGRATED ASSESSMENT TOOL WITH TAM

Following the identification of the necessity of developing a bespoke assessment tool to gain a tangible understanding of this domain of AI trust, we proceeded to define it as "*Operational Trust." This is defined as "the degree of trust a user has in the system's ability to provide accurate and reliable guidance, leading to the user's final decision to follow the system's recommendations*."

Consequently, the POTDAI tool was designed to evaluate the extent to which a user's perception of an AI system influences their willingness to follow its recommendations. It also assesses whether this predisposition is influenced by attitudes of overtrust, mistrust or the AI's monitoring of the user's work. This definition does not include the aspects of the system's general perception that are covered in |R5|, such as usability. As will be detailed below, to evaluate general system perception, our tool has been integrated with the TAM model, which already has the capability to evaluate these aspects.

The model developed by [23] and [24] is one of the best known and most widely used models. Its primary format is specifically aimed at evaluating user perception, which is completely aligned with our case – the user's operational trust in AI perception. The model proposes that a series of interrelated factors influence the user's attitude towards a new technology as well as the decision of how and when to use it. The original approach is based on two main factors: Perceived Usefulness, PU, defined as "the degree to which a person believes that using a particular system will make them or their job performance stand out." In addition, the concept of Perceived Ease of Use (PEOU) is introduced. This is defined as the degree to which a person believes that using a particular system will relieve them of their effort.

The model proposes that users' perceptions of these two factors determine their intention to use a system. It should be noted that the selection of TAM items followed a process very similar to that discussed for other evaluations. From the conceptual definitions of perceived usefulness and perceived ease of use, two constructs that were identified as essential by the literature prior to the study when accepting or rejecting a technology – 14 potential items–were generated for evaluation. Subsequently, the wording of each item was studied to ascertain which items best aligned with the definitions of each construct, resulting in the selection of ten items for each construct. Later, [23] conducted a study with 112 users, whose results were analyzed through factorial, multitrait-multimethod, and reliability study methods, to refine the scales by reducing each construct to six items.

Despite criticism, TAM has been widely used as an evaluation tool, with positive results. The information obtained has been useful for the development of the technology in question [48] on several occasions. However, critics highlight its limited practical value and explanatory and predictive power [49]. This has prompted both proponents of the original model and other scholars to expand the model by incorporating additional variables and tailoring it to specific contexts of use.

The authors of the study agree that the TAM model – and thus the evaluation tool we propose in this paper – has limitations, but this is a characteristic fact of any evaluation tool in the field of human-computer interaction. To illustrate, if we wish to conduct an in-depth usability study of a technology, it is advisable to rely on a single evaluation method. There are numerous evaluation methods at our disposal, including cognitive walkthrough, scenarios and personas, brainstorming, interviews, the Think Aloud Protocol (TAP) [50]) and questionnaires such as the System Usability Scale (SUS) [51].

Each tool has a specific purpose and is useful for extracting particular types of information. For instance, the think-aloud protocol (TAP) is not an appropriate tool for measuring precise efficiency because it requires the user to verbalize their actions, which is time-consuming. In contrast, a benchmark test [52] would be more suitable for this purpose, as it allows for the collection of quantitative data; however, it will lack qualitative information (e.g., feelings, mental processes, etc.) available through the TAP.

Using questions on a Likert scale prevents delving into the "whys." If a user indicates that the system is not useful at all (i.e. selecting 1 on a 1 to 7 scale to reply the question "I would find the system useful in my job"), there is no other question that directly indicates the reason. The user simply provides his/her perception. This is an unquestionable limitation; however, following the previous example, other techniques, such as interviews, could be used to gain further insight. In any case, this type of scale is highly suitable for evaluating the subjective aspects of the user in a comfortable, simple, and quick manner [28].

Although the debate surrounding the validity of the TAM as a theory of information systems is beyond the scope of this study, what we require is to gain an understanding of the user's perception and how it will affect the decision to follow AIs recommendations. As previously explained, we don't seek to delve in the "whys."

Once the user's perception is known, it will be possible to guide the development in a specific direction or implement additional tests before taking the next step. The original version of the TAM allows us to obtain this initial perception in terms of usefulness and ease of use in a quick and simple manner, which is why we decided to use it as a basis. This is also in line with the study performed by Vorm and Combs [25], which concluded that the TAM model should be extended to gain insight into the degree of confidence in systems that use AI.

With regard to the number of response options on the Likert scale – seven according to the original TAM model and other questionnaires such as the Checklist for Trust between People and Automation or the MDMT (Multi-Dimensional Measure of Trust) – it is essential to highlight that a number of studies have been conducted with the objective of analyzing the impact of the number of response alternatives on the psychometric properties of Likert-type scales.

The findings of these studies indicate that the reliability of the scales is optimal when there are seven response alternatives [53]. It is also noteworthy that, in our scale, three items (14, 17, and 18) were scored inversely. This is a deliberate decision and serves to control the response bias of a subject, which was envisaged during the requisite definition stage. This is particularly pertinent for dealing with acquiescence bias, which is the tendency of an individual to respond affirmatively to a statement regardless of its content and is a useful indicator of response bias.

As previously stated, following the literature review, no evaluation tools that could be directly applied were identified. However:

- The elementary TAM model [23], [24] was deemed an appropriate basis for the study, as it allowed for the extraction of information reflected in requirement 5 while also satisfying requirements 1 and 2.
- A challenge was identified in filling the gap highlighted by Vorm and Combs [25], who examined new theoretical frameworks that are used in HCI and can help in the development of systems. They concluded that it is timely for the elementary factors of the TAM model to be extended to include issues related to trust and user acceptance of systems employing artificial intelligence.
- We gathered data on the development of other AI trust assessment tools, which were useful for defining a methodology to guide the development of a tool that meets specific requirements.

In light of the above, it was determined that the TAM would serve as the foundation, with the inclusion of questions related to trust in AI to address the requirements set out at points 3 and 4. To meet requirements R1 and R2, while maintaining the structural integrity of the TAM, it was decided that an additional set of six questions should be included, bringing the total number of questions to 18.

At the outset, there was some doubt among the experts as to whether the proposed six-question approach would be sufficient to meet expectations. During the discussion, it was noted that if the TAM model could provide a useful assessment of perceived usefulness and perceived ease of use with six questions each, it was reasonable to assume that something similar could be achieved for the perceived degree of trust.

Additionally, other scales were discussed, such as the System Usability Scale (SUS), which assesses usability using 10 questions. Even simpler assessment tools, such as those used by Guo [19], Schreibelmayr et al. [37], and Sebastian et al. [31], were considered. Consequently, the experts focused on trust-related requirements (|R3| and |R4|), implementing the card sorting experiment as described in the methodology section. The proposed questions were refined through successive rounds of discussion, focusing primarily on the following:

- Repetition: The same aspect was queried on two occasions, albeit in different ways. For instance: *"I am concerned that I will not be able to discern if the [AI TECHNOLOGY] is giving me an inappropriate response"* and *"I think I could easily detect when [THE AI TECHNOLOGY] is not giving me the right indications and redirect the situation"*

- Grouping: questions posed regarding different factors that could be grouped together to achieve a similar operational outcome. For instance *"I think I will feel that I will be observed, watched or judged by [THE AI TECHNOLOGY] during the course of my work."* was composed by grouping different questions asking just about ''being judged'' or ''being watched,'' etc.

- Out of context: The question was not directly related to the assessment of ''overtrust/mistrust,'' but rather to a factor derived from these or that is not controllable by the user. For example, *"I believe that the use of [AI TECHNOLOGY] will be excessively expensive."*

- Generalization: Questions pertaining to trust in AI in general rather than in the specific technology being tested. For instance, the question *"I believe that AI algorithms will steal my data"* is an example of a question that refers to any type of AI.

- Technical Expertise: The question can only be answered by an AI subject matter expert and does not have an operational focus. For example, the question *"I consider that this AI does not meet the requirements set by the European Commission regarding Trustworthy AI*."

This process not only served to define the questions, but also to make other decisions affecting the overall questionnaire. These included:

- Reverse scores were used to detect inconsistencies during the completion of the questionnaire.
- Observations were made on the possible relationships between the original TAM questions and new questions on the perception of operational trust.
- It was determined that the selected questions could be employed independently from the TAM to gather data on the extent of operational trust. Furthermore, alternative methodologies could be employed to assess other aspects related to usability, usefulness, and UX.
- It was also concluded that the evaluation tool would be beneficial in most cases where the perception of the degree of trust in a system that employs AI is to be evaluated, in addition to what is expressed in requirement 1.

Table 8 lists the final wording of the six questions. This indicates whether the value was normal (N) or inverted (I). A normal value is associated with a higher Likert score, indicating a more favorable assessment of the technology using AI. It also included comments on the fulfilment of requirements 3 and/or 4, their relation to other questions in the questionnaire, and the concepts of trustworthy AI.

As can be seen, there were two questions from a distrust perspective, two from an overtrust perspective, and two on a mistrust-overtrust scale. Of the six questions, three are related to requirement 4, one of which (question 4) is a direct question.

TABLE 8. Final assessment questions.

QUESTION	REQUISITES AND OBSERVATIONS
Q1. (I) I believe that [THE AI TECHNOLOGY] will be prone to not give me the most convenient indications	|R3| Mistrust: The operator perceives that an AI is not mature, does not have the ability to perform the task, or simply distrusts AIs in general.
Q2. (N) I think I could easily detect when [THE AI TECHNOLOGY] is not giving me the right indications and redirect the situation.	|R3| Overtrust: The operator feels that his knowledge is sufficient to avoid over-reliance on what the AI dictates.
Q3. (I) I think that working with [THE AI TECHNOLOGY] could lead to overconfidence and dependency.	|R3| Overtrust: The operator feels that the technology may create an excessive/constant need to consult the AI to validate its decisions. This constant need could be associated with |R4| because the operator feels that performing a task contrary to what the AI dictates could cause him harm.
Q4. (I) I think I will feel that I will be observed, watched or judged by [THE AI TECHNOLOGY] during the course of my work.	This question is directly related to |R4| and was an aspect highlighted by security experts as a common concern for agents. Doubts arise about the consequences of following or not following AI recommendations (judged), or that the AI is watching while the user performing a task, or if the user is constantly being observed by the AI. As for |R3|, it would be associated with Mistrust.
Q5. (N) In cases where I am not sure how to proceed, I would normally follow the recommendations provided by [THE AI TECHNOLOGY].	|R3| Mistrust-Overtrust: Global scale on whether the user would follow the AI's instructions operationally when the user feels that AI support is truly needed. This question should be consistent with the previous answers. The operator may be inclined to follow the recommendations because he really trusts the system (questions 1 and 2) or because he perceives that not doing so could have consequences (question 4). It may also indicate how much weight the above questions carry for the operator in the final operational trust decision.
Q6. (N) I believe that [THE AI TECHNOLOGY] will allow me to do my job more safely.	|R3| Mistrust-Overtrust: Especially in cases where decisions can be critical and need to be made quickly, the ability of AI to provide greater certainty in decisions is a key factor. The feeling of being watched or judged using AI can make the operator feel less trust and/or distract them from their work. It is therefore interesting to note the relationship with the result of question 4.

E. RELATION WITH ISTAM APPROACH AND RECOMMENDATIONS FOR THE USE OF POTDAI TOOL

In their article, Vorm and Combs [25] emphasize the importance of designing more transparent systems to foster trust and acceptance of intelligent systems. They examine several transparency frameworks and present a model where transparency is added to "Perceived ease of use" (PEOU) and "Perceived usefulness" (PU). This model identifies different transparency factors and shows how those factors play moderating and supporting roles that combine to influence trust and, ultimately, acceptance. Their research outlines the significant impact that transparency has on trust and presents a comprehensive approach to evaluating transparency based on their findings. However, they do not provide a specific set of questions or an evaluation tool. In this sense, their work is linked to the extended versions of the TAM model, where more variables/factors and relations have been established among PU and PEOU, with the aim of increasing the robustness in explaining the use intention. This can be seen in figure 4, extracted from Vorm and Combs [25] paper, where they include "Transparency" at the same level as PU and PEOU, as described in TAM3 version.

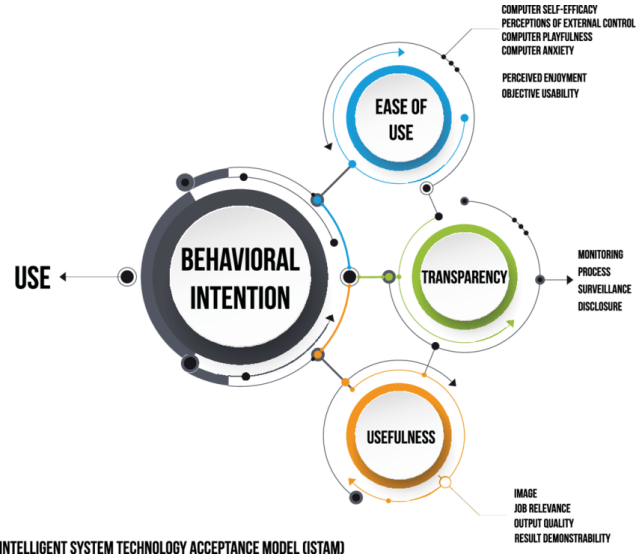


FIGURE 4. ISTAM model as depicted in Vorm and Combs [25].

Our work is based on the fundamental tenets of the Technology Acceptance Model (TAM). It has a clear objective: to assess user confidence in complying with the directives of an intelligent system. We concur that transparency is a factor that affects trust and, as a result, the final decision to follow the instructions of an intelligent system. However, transparency is also influenced by several other factors, as Vorn and Combs have highlighted. One such factor is the user's technical knowledge, which is notably sophisticated in this field. While organizations or states may offer certifications to endorse different aspects of AI transparency, if the users lacks the expertise, it will be challenging or impossible for them to verify those claims independently. Consequently, even if an intelligent system presents technically verifiable transparency, the user may perceive the opposite.

We recognize that there are multiple factors that influence trust, and that the information provided by POTDAI has its limitations. POTDAI is designed to assess the user's trust perception in intelligent systems and their willingness to follow recommendations from those systems, based on criteria related to trust and mistrust. POTDAI is not designed to provide specific recommendations about how to increase the trustworthiness of an intelligent system or how to implement those recommendations. However, it can provide valuable insights into how to approach these questions. As previously stated, we believe that attempting to expand TAM to encompass all the necessary information is unfeasible and impractical. Instead, we recommend leveraging alternative HCI techniques tailored to the specific objectives at hand.

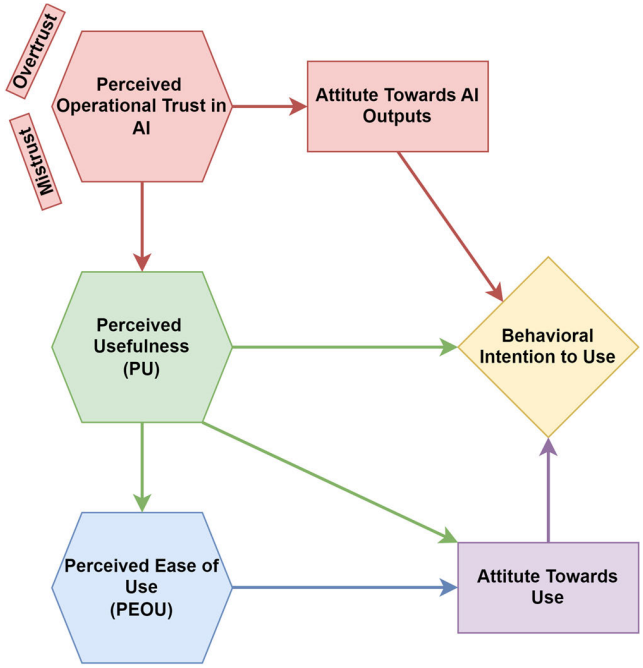


FIGURE 5. Conceptualization of POTDAI integrated into TAM.

We propose a set of 6 questions inspired by the 12 items TAM [23] – 6 for PU and 6 for PEOU. These can be used alone to make a quick assessment of the user's inclination to follow the recommendations provided by an intelligent system. Our conceptualization is depicted in figure 5 – we used the hexagon shape to represent the 6 questions. We encourage researchers interested in using POTDAI to integrate it with the basic version of TAM. This will allow us to gather evidence regarding the interrelationships between POTDAI and TAM's PU and PEOU questions, as outlined in the conclusions and future work section.

IV. CONCLUSION AND FUTURE WORK

There is no single definition for trust in AI. Different authors have defined and assessed trust according to the factors they consider relevant, derived from their object of study. The objective of this study is to evaluate the final cause: *whether or not the user will tend to follow the AI instructions*. For this purpose, we consider the excess or lack of trust (overtrust/mistrust) due to the user's perception of the AI system as key factors, and whether this is influenced by the user's perception of whether it is being monitored by the AI system.

Our assessment tool does not delve into the motivation for this perception but simply into its existence. By focusing on perception and not delving into the reasons, it limits the usefulness of the tool, which is also a common criticism of the TAM model [54]. However, knowing the user's perception is of great importance, and other specific techniques can be used to go deeper into the whys and wherefores, and how to deal with the issues to be addressed.

For example, in the case of Question 2 of our tool (see Table 8), the user could overestimate their knowledge. To verify this, it is necessary to carry out some tests. If this test showed that the user did indeed overvalue their knowledge, then another type of technique would be needed to determine why and then consider how to remedy this. Thus, this could be a limitation but not a disadvantage. As explained in the results section, there is a huge variety of evaluation techniques in the field of HCI, and each technique has a purpose and is useful for extracting some kind of information, which can then be used to develop another technique. Therefore, studies in this field are not usually limited to the use of a single technique.

According to the authors, the level of trust perceived by the user is a key element of its profile, which must be taken into account in order to be able to evaluate the results of any other type of test, for example, mistrust in the AI system, may worsen the efficiency scores in a benchmark test or to plan how the user should be trained to use the system.

The literature review also shows how the construction of AI trust assessment tools often starts by referring to one or more of the various definitions of trust or by creating a new one. Once the definition of trust has been established, the authors identify the factors that they consider relevant based on which they proceed to define the evaluation tool, either through their own methods or by adopting some or all existing evaluations.

In our case, we not only follow the approach used by authors such as Park et al. [38] to define their assessment tools but also

- We defined the main object and requirements through expert analysis.
- We conducted a comprehensive literature review.
- We justified the lack of a previously established tool to validate our requirements.
- We then defined the concept of "operational confidence" and identified three key factors: overtrust, mistrust, and monitoring.
- We also generated the questions through a committee of experts, supported by the application of the card sorting technique, and provided details of the process of refining the questions.
- Each question is clearly related to the established requirements. In addition, the expected interrelationships with the other questions are indicated (see Table 8).

This demonstrates the rigor in creating the evaluation tool using a methodology that, as we have commented, exceeds usual practice. POTDAI has been designed with the TAM model as a reference and has been integrated into it, thus aiming to fill the gap described by Vorm and Combs [25]. It should be emphasized that POTDAI can be used by itself to extract information on the level of operational reliability, although its integration with TAM would provide additional information on usability and ease of use. This additional information can be useful for inferring certain aspects or checking the consistency of user responses.

Our forthcoming research will entail the application of POTDAI to different use cases. This will allow to establish relationships between TAM and POTDAI questions. Similarly, the use of the tool would provide valuable information

for its refinement and how to interpret the data gathered through POTDAI. To this end, it would be advisable to employ advanced statistical techniques such as structural equation modeling (SEM), a multivariate statistical analysis technique that is widely utilized in the validation of psychometric instruments. SEM integrates other techniques, including analysis of variance (ANOVA), multiple regression, and factor analysis. This would facilitate the identification of intricate patterns of relationships between test factors, enable comparisons between them, and even allow for the modeling of measurement error.

ACKNOWLEDGMENT

The authors would like to thank the HumanE AI Network team (https://www.humane-ai.eu), and the institutions participating in the "To develop a trustworthy AI model for situation awareness by using mixed reality in police interventions." microproject: Umeå University - Computing Science Department, Umeå University - Police Education Unit, Comet Global Innovation S.L. and Institut de Seguretat Pública de Catalunya -ISPC.

REFERENCES

[1] J. Li, Y. Zhou, J. Yao, and X. Liu, "An empirical investigation of trust in AI in a Chinese petrochemical enterprise based on institutional theory," *Sci. Rep.*, vol. 11, no. 1, p. 13564, Jun. 2021, doi: 10.1038/s41598-021-92904-7.

[2] K. Okamura and S. Yamada, "Empirical evaluations of framework for adaptive trust calibration in human-AI cooperation," *IEEE Access*, vol. 8, pp. 220335–220351, 2020, doi: 10.1109/ACCESS.2020.3042556.

[3] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," *ACM Trans. Interact. Intell. Syst.*, vol. 11, nos. 3–4, pp. 1–45, Sep. 2021, doi: 10.1145/3387166.

[4] S. Sousa, J. Cravino, and P. Martins, "Challenges and trends in user trust discourse in AI popularity," *Multimodal Technol. Interact.*, vol. 7, no. 2, p. 13, Jan. 2023, doi: 10.3390/mti7020013.

[5] A. Papenmeier, D. Kern, G. Englebienne, and C. Seifert, "It's complicated: The relationship between user trust, model accuracy and explanations in AI," *ACM Trans. Comput.-Hum. Interact.*, vol. 29, no. 4, pp. 1–33, Mar. 2022, doi: 10.1145/3495013.

[6] K. M. Stiff, M. J. Franklin, Y. Zhou, A. Madabhushi, and T. J. Knackstedt, "Artificial intelligence and melanoma: A comprehensive review of clinical, dermoscopic, and histologic applications," *Pigment Cell Melanoma Res.*, vol. 35, no. 2, pp. 203–211, Mar. 2022, doi: 10.1111/pcmr.13027.

[7] M. Phillips, H. Marsden, W. Jaffe, R. N. Matin, G. N. Wali, J. Greenhalgh, E. McGrath, R. James, E. Ladoyanni, A. Bewley, G. Argenziano, and I. Palamaras, "Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions," *JAMA Netw. Open*, vol. 2, no. 10, Oct. 2019, Art. no. e1913436, doi: 10.1001/jamanetworkopen.2019.13436.

[8] I. B. Ajenaghughrure, S. C. D. C. Sousa, and D. Lamas, "Psychophysiological modeling of trust in technology: Influence of feature selection methods," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, pp. 1–25, May 2021, doi: 10.1145/3459745.

[9] V. Dominguez, P. Messina, I. Donoso-Guzmán, and D. Parra, "The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images," in *Proc. 24th Int. Conf. Intell. User Interfaces* New York, NY, USA: ACM, Mar. 2019, pp. 408–416, doi: 10.1145/3301275.3302274.

[10] I. Amaro, P. Barra, A. D. Greca, R. Francese, and C. Tucci, "Believe in artificial intelligence? A user study on the ChatGPT's fake information impact," *IEEE Trans. Computat. Social Syst.*, vol. 11, no. 4, pp. 5168–5177, Aug. 2024, doi: 10.1109/TCSS.2023.3291539.

[11] D. Martín-Moncunill, E. García-Barriocanal, M.-A. Sicilia, and A. S. Sánchez-Alonso, "Evaluating the practical applicability of thesaurus-based keyphrase extraction in the agricultural domain: Insights from the VOA3R project," *Knowl. Org.*, vol. 42, no. 2, pp. 76–89, 2015, doi: 10.5771/0943-7444-2015-2-76.

[12] F. Durso, M. S. Raunak, R. Kuhn, and R. Kacker, "Analyzing failures in artificial intelligent learning systems (FAILS)," in *Proc. IEEE 29th Annu. Softw. Technol. Conf. (STC)*, Oct. 2022, pp. 7–8, doi: 10.1109/STC55697.2022.00010.

[13] I. Linkov, S. Galaitsi, B. D. Trump, J. M. Keisler, and A. Kott, "Cybertrust: From explainable to actionable and interpretable artificial intelligence," *Computer*, vol. 53, no. 9, pp. 91–96, Sep. 2020, doi: 10.1109/MC.2020.2993623.

[14] N. E. Boudette. (2021). *Tesla Says Autopilot Makes Its Cars Safer. Crash Victims Say it Kills*. New York Times. [Online]. Available: https://www.nytimes.com/2021/07/05/business/tesla-autopilot-lawsuits-safety.html

[15] T. A. Bach, A. Khan, H. Hallock, G. Beltrão, and S. Sousa, "A systematic literature review of user trust in AI-enabled systems: An HCI perspective," *Int. J. Hum.–Comput. Interact.*, vol. 40, no. 5, pp. 1251–1266, Nov. 2022, doi: 10.1080/10447318.2022.2138826.

[16] A. Sharma, A. Gupta, A. Bhargava, A. Rawat, P. Yadav, and D. Gupta, "From sci-fi to reality: The evolution of human-computer interaction with artificial intelligence," in *Proc. 2nd Int. Conf. Appl. Artif. Intell. Comput. (ICAAIC)*, May 2023, pp. 127–134, doi: 10.1109/ICAAIC56838.2023.10141431.

[17] A. Krausman, C. Neubauer, D. Forster, S. Lakhmani, A. L. Baker, S. M. Fitzhugh, G. Gremillion, J. L. Wright, J. S. Metcalfe, and K. E. Schaefer, "Trust measurement in human-autonomy teams: Development of a conceptual toolkit," *ACM Trans. Hum.-Robot Interact.*, vol. 11, no. 3, pp. 1–58, Sep. 2022, doi: 10.1145/3530874.

[18] R. Zhang, C. Flathmann, G. Musick, B. Schelble, N. J. McNeese, B. Knijnenburg, and W. Duan, "I know this looks bad, but I can explain: Understanding when AI should explain actions in human-AI teams," *ACM Trans. Interact. Intell. Syst.*, vol. 14, no. 1, pp. 1–23, Mar. 2024, doi: 10.1145/3635474.

[19] Y. Guo, "Digital trust and the reconstruction of trust in the digital society: An integrated model based on trust theory and expectation confirmation theory," *Digit. Government, Res. Pract.*, vol. 3, no. 4, pp. 1–19, Oct. 2022, doi: 10.1145/3543860.

[20] B. Gebru, L. Zeleke, D. Blankson, M. Nabil, S. Nateghi, A. Homaifar, and E. Tunstel, "A review on human–machine trust evaluation: Human-centric and machine-centric perspectives," *IEEE Trans. Human-Mach. Syst.*, vol. 52, no. 5, pp. 952–962, Oct. 2022, doi: 10.1109/THMS.2022.3144956.

[21] A. Duenser and D. M. Douglas, "Whom to trust, how and why: Untangling artificial intelligence ethics principles, trustworthiness, and trust," *IEEE Intell. Syst.*, vol. 38, no. 6, pp. 19–26, Nov. 2023, doi: 10.1109/MIS.2023.3322586.

[22] P. Purwanto, K. Kuswandi, and F. Fatmah, "Interactive applications with artificial intelligence: The role of trust among digital assistant users," *Foresight STI Governance*, vol. 14, no. 2, pp. 64–75, Jun. 2020, doi: 10.17323/2500-2597.2020.2.64.75.

[23] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quart.*, vol. 13, no. 3, p. 319, Sep. 1989, doi: 10.2307/249008.

[24] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, "User acceptance of computer technology: A comparison of two theoretical models," *Manage. Sci.*, vol. 35, no. 8, pp. 982–1003, Aug. 1989, doi: 10.1287/mnsc.35.8.982.

[25] E. S. Vorm and D. J. Y. Combs, "Integrating transparency, trust, and acceptance: The intelligent systems technology acceptance model (ISTAM)," *Int. J. Hum.-Comput. Interact.*, vol. 38, nos. 18–20, pp. 1828–1845, Dec. 2022, doi: 10.1080/10447318.2022.2070107.

[26] J. A. Goguen and C. Linde, "Techniques for requirements elicitation," in *Proc. IEEE Int. Symp. Requir. Eng.*, Jun. 1993, pp. 152–164, doi: 10.1109/ISRE.1993.324822.

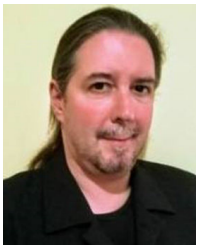
[27] J. R. Wood. (2008). *Card Sorting: Current Practices and Beyond*. [Online]. Available: https://www.semanticscholar.org/paper/Card-Sorting-%3A-Current-Practices-and-Beyond-Wood/334ea2265d294cc40cb850c66360187c96a48a0b

[28] S. Cao and C.-M. Huang, "Understanding user reliance on AI in assisted decision-making," *Proc. ACM Hum.-Comput. Interact.*, vol. 6, pp. 1–23, Nov. 2022, doi: 10.1145/3555572.

[29] R. Larasati, A. De Liddo, and E. Motta, "Meaningful explanation effect on user's trust in an AI medical system: Designing explanations for non-expert users," *ACM Trans. Interact. Intell. Syst.*, vol. 13, no. 4, pp. 1–39, Dec. 2023, doi: 10.1145/3631614.

[30] A. Choudhury and H. Shamszare, "Investigating the impact of user trust on the adoption and use of ChatGPT: Survey analysis," *J. Med. Internet Res.*, vol. 25, Jun. 2023, Art. no. e47184, doi: 10.2196/47184.
[31] G. Sebastian, A. George, and G. Jackson Jr., "Persuading patients using rhetoric to improve artificial intelligence adoption: Experimental study," *J. Med. Internet Res.*, vol. 25, Mar. 2023, Art. no. e41430, doi: 10.2196/41430.
[32] M. H. Lee and C. J. Chew, "Understanding the effect of counterfactual explanations on trust and reliance on AI for human-AI collaborative clinical decision making," *Proc. ACM Hum.-Comput. Interact.*, vol. 7, pp. 1–22, Oct. 2023, doi: 10.1145/3610218.
[33] G. Warren, R. M. J. Byrne, and M. T. Keane, "Categorical and continuous features in counterfactual explanations of AI systems," in *Proc. 28th Int. Conf. Intell. User Interfaces*, New York, NY, USA, Jun. 2023, pp. 171–187, doi: 10.1145/3673907.
[34] L. Sanneman and J. A. Shah, "The situation awareness framework for explainable AI (SAFE-AI) and human factors considerations for XAI systems," *Int. J. Hum.-Comput. Interact.*, vol. 38, nos. 18–20, pp. 1772–1788, Dec. 2022, doi: 10.1080/10447318.2022.2081282.
[35] R. Dvorak, H. Liao, S. Schibel, and B. Tribelhorn, "Towards evaluating ethical accountability and trustworthiness in AI systems," *J. Comput. Sci. Coll.*, vol. 37, no. 2, pp. 11–22, Oct. 2021.
[36] J. Drozdal, J. Weisz, D. Wang, G. Dass, B. Yao, C. Zhao, M. Müller, L. Ju, and H. Su, "Trust in AutoML: Exploring information needs for establishing trust in automated machine learning systems," in *Proc. 25th Int. Conf. Intell. User Interfaces*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 297–307, doi: 10.1145/3377325.3377501.
[37] S. Schreibelmayr, L. Moradbakhti, and M. Mara, "First impressions of a financial AI assistant: Differences between high trust and low trust users," *Frontiers Artif. Intell.*, vol. 6, Oct. 2023, Art. no. 1241290, doi: 10.3389/frai.2023.1241290.
[38] S. Park, H. K. Kim, J. Park, and Y. Lee, "Designing and evaluating user experience of an AI-based defense system," *IEEE Access*, vol. 11, pp. 122045–122056, 2023, doi: 10.1109/ACCESS.2023.3329257.
[39] J. Wanner, L.-V. Herm, K. Heinrich, and C. Janiesch, "The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study," *Electron. Markets*, vol. 32, no. 4, pp. 2079–2102, Dec. 2022, doi: 10.1007/s12525-022-00593-5.
[40] A. A. Tutul, E. H. Nirjhar, and T. Chaspari, "Investigating trust in human-machine learning collaboration: A pilot study on estimating public anxiety from speech," in *Proc. Int. Conf. Multimodal Interact.* New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 288–296, doi: 10.1145/3462244.3479926.
[41] M. Körber. (2024). *Moritzkoerber/TiA_Trust_in_Automation_Questionnaire*. [Online]. Available: https://github.com/moritzkoerber/TiA_Trust_in_Automation_Questionnaire
[42] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *Int. J. Cognit. Ergonom.*, vol. 4, no. 1, pp. 53–71, Mar. 2000, doi: 10.1207/s15327566ijce0401_04.
[43] M. Madsen and S. Gregor. (2000). *Measuring Hum.-Comput. Trust*. [Online]. Available: https://www.semanticscholar.org/paper/Measuring-Human-Computer-Trust-Madsen-Gregor/b8eda9593fbcb63b7ced1866853d9622737533a2
[44] D. Ullman and B. F. Malle, "What does it mean to trust a robot? Steps toward a multidimensional measure of trust," in *Proc. Companion ACM/IEEE Int. Conf. Hum.-Robot Interact.* Chicago, IL, USA: ACM, Mar. 2018, pp. 263–264, doi: 10.1145/3173386.3176991.
[45] B. F. Malle and D. Ullman, "Measuring human-robot trust with the MDMT (multi-dimensional measure of trust)," 2023, *arXiv:2311.14887*.
[46] A. T. Jebb, V. Ng, and L. Tay, "A review of key Likert scale development advances: 1995–2019," *Frontiers Psychol.*, vol. 12, May 2021, Art. no. 637547, doi: 10.3389/fpsyg.2021.637547.
[47] S. C. Kohn, E. J. de Visser, E. Wiese, Y.-C. Lee, and T. H. Shaw, "Measurement of trust in automation: A narrative review and reference guide," *Frontiers Psychol.*, vol. 12, Oct. 2021, Art. no. 604977, doi: 10.3389/fpsyg.2021.604977.
[48] D. Marikyan and S. Papagiannidis. (2023). *Technology Acceptance Model: A Review*. TheoryHub Book. [Online]. Available: https://open.ncl.ac.uk/
[49] I. Benbasat, H. Barki, and H. Montréal, "Quo vadis TAM?" *J. Assoc. Inf. Syst.*, vol. 8, no. 4, pp. 211–218, Apr. 2007, doi: 10.17705/1jais.00126.
[50] S. McDonald, H. M. Edwards, and T. Zhao, "Exploring think-alouds in usability testing: An international survey," *IEEE Trans. Prof. Commun.*, vol. 55, no. 1, pp. 2–19, Mar. 2012, doi: 10.1109/TPC.2011.2182569.
[51] J. Brooke, "SUS: A, 'quick and dirty' usability scale," in *Usability Evaluation In Industry*. Boca Raton, FL, USA: CRC Press, 1996.
[52] B. Myers, N. Altman, K. Amiri, M. Centurion, F. Chang, C. Chen, H. Derby, J. Huebner, R. Kaylor, R. Melton, R. O'Callahan, M. Tarpy, K. Unyelioglu, and Z. Wang. (1997). *Using Benchmarks to Teach and Evaluate User Interface Tools*. [Online]. Available: https://www.semanticscholar.org/paper/Using-Benchmarks-to-Teach-and-Evaluate-User-Tools-Myers-Altman/25aef1a2ca024a6f91d20d6d6622babf3ec97c58
[53] A. Matas, "Diseño del formato de escalas tipo likert: Un estado de la cuestión," *Revista Electrónica de Investigación Educativa*, vol. 20, no. 1, pp. 38–47, Feb. 2018.
[54] R. Bagozzi, "The legacy of the technology acceptance model and a proposal for a paradigm shift," *J. Assoc. Inf. Syst.*, vol. 8, no. 4, pp. 244–254, Apr. 2007, doi: 10.17705/1jais.00122.

DAVID MARTÍN-MONCUNILL (Member, IEEE) was born in Madrid, Spain, in 1984. He received the degree in information systems from Universidad de Alcalá, in 2013, the M.Sc. degree in e-learning from Universidad Internacional de la Rioja, in 2015, and the Ph.D. degree in information and knowledge engineering from Universidad de Alcalá, in 2018. He currently leads the Robotics and Artificial Intelligence Engineering degree. He was a Computer Engineer with Universidad de Alcalá, in 2012. He is a member of the Computing and Artificial Intelligence Laboratory (CAILab), Universidad Camilo José Cela, Madrid, Spain. His professional work, outside the academic world, is focused on research and development management and direction. He has participated in more than 20 research projects, most of them in the European context (FP7, H2020, Horizon, and Europe). He is focused on human–computer and human–robot interaction, trustworthy artificial intelligence, usability, and user experience (UX).



EDUARDO GARCÍA LAREDO (Member, IEEE) was born in Madrid, Spain, in 1976. He received the B.S. degree in psychology from Camilo José Cela University, in 2004, the Master's Diploma degree from the Department of Psychiatry and Medical Psychology, Faculty of Medicine, Complutense University of Madrid, in 2007, and the Ph.D. degree in neurosciences from the Complutense University of Madrid, in 2016, with additional training in neuroimaging by MRI, by the FIDMAG Sisters Hospitallers Research Foundation and the Spanish Society of Neuroimaging, in 2019. He is currently a Professor of neuroscience, methodology, and statistical analysis with the International University of La Rioja (UNIR) and a Researcher in human–computer interaction (HCI) with COMET GLOBAL INNOVATION S. L. His professional career has been focused on teaching and clinical and technological research, especially in severe mental illness and the use of technology to improve the quality of life in patients with psychosis and depression.

JUAN CARLOS NIEVES received the Ph.D. degree from Universitat Politècnica de Catalunya—Barcelona Tech (UPC), Barcelona, Spain, in November 2008. He is currently an Associate Professor (docent) with Umeå universitet, Umeå, Sweden. He has served as an external (ethical) advisor/reviewer for different EU projects. His research interests include trustworthy artificial intelligence (AI), theory and application of nonmonotonic reasoning, and multi-agent systems. He has served as an expert reviewer for various European national research councils. He has also been an AI ethical advisor for European initiatives, such as EU BonAPPS, and for American initiatives, such as fAIr LAC of the Inter-American Development Bank.