Metrics for Evaluating Explainable Recommender Systems

Hulstijn, Joris; Tchappi, Igor; Najjar, Amro; Aydoğan, Reyhan

DOI
10.1007/978-3-031-40878-6_12

Publication date
2023

Document Version
Final published version

Published in
Explainable and Transparent AI and Multi-Agent Systems - 5th International Workshop, EXTRAAMAS 2023, Revised Selected Papers

Citation (APA)
Hulstijn, J., Tchappi, I., Najjar, A., & Aydoğan, R. (2023). Metrics for Evaluating Explainable Recommender Systems. In D. Calvaresi, A. Najjar, A. Omicini, R. Carli, G. Ciatto, R. Aydogan, Y. Mualla, & K. Främling (Eds.), *Explainable and Transparent AI and Multi-Agent Systems - 5th International Workshop, EXTRAAMAS 2023, Revised Selected Papers* (pp. 212-230). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 14127 LNAI). Springer. https://doi.org/10.1007/978-3-031-40878-6_12

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Metrics for Evaluating Explainable Recommender Systems

Joris Hulstijn[1(✉)], Igor Tchappi[1], Amro Najjar[1,2], and Reyhan Aydoğan[3,4]

[1] University of Luxembourg, Esch-sur-Alzette, Luxembourg
joris.hulstijn@uni.lu
[2] Luxembourg Institute of Science and Technology (LIST), Esch-sur-Alzette, Luxembourg
[3] Computer Science, Özyeğin University, Istanbul, Turkey
[4] Interactive Intelligence, Delft University of Technology, Delft, Netherlands

Abstract. Recommender systems aim to support their users by reducing information overload so that they can make better decisions. Recommender systems must be transparent, so users can form mental models about the system's goals, internal state, and capabilities, that are in line with their actual design. Explanations and transparent behaviour of the system should inspire trust and, ultimately, lead to more persuasive recommendations. Here, explanations convey reasons why a recommendation is given or how the system forms its recommendations. This paper focuses on the question how such claims about effectiveness of explanations can be evaluated. Accordingly, we investigate various models that are used to assess the effects of explanations and recommendations. We discuss objective and subjective measurement and argue that both are needed. We define a set of metrics for measuring the effectiveness of explanations and recommendations. The feasibility of using these metrics is discussed in the context of a specific explainable recommender system in the food and health domain.

Keywords: Evaluation · Metrics · Explainable AI · Recommender systems

1 Introduction

Artificial intelligence is becoming more and more pervasive. However, there are also concerns about bias in the algorithm, bias in the data set, or stereotyping users, to name a few examples [9]. In particular, if autonomous systems take decisions, how can they justify and explain these decisions, to those who are affected? These concerns have led to an increasing interest in *responsible* AI [9] and specifically in *explainable* AI (XAI) [2], witness the special issue [27].

An important application of explainable AI is found in *recommender systems* [1,35]. A recommender system is "any system that guides a user in a personalized way to interesting or useful objects in a large space of possible options or that produces such objects as output" [5, p. 2]. Increasingly, recommender systems also provide explanations [35]. There are two types: explanations that motivate the system's choice of recommendations, and explanations that clarify how the system works, to derive the recommendations. In this paper, we focus on the former type of explanations.

D. Calvaresi et al. (Eds.): EXTRAAMAS 2023, LNAI 14127, pp. 212–230, 2023.
https://doi.org/10.1007/978-3-031-40878-6_12

If the purpose is to persuade users to change their behaviour, only predicting which recommendation best fits a user profile, is not enough. Users expect reasons that motivate why this recommendation was given, and not another. That is why researchers now aim to build systems that provide an explanation, personalized to the user's preferences and to the context. In addition, recommender systems are becoming more interactive. A recommendation must be followed by an opportunity for feedback [16]. This allows users to correct misunderstandings and ask follow-up questions.

Designing interactive systems is a complex task. Unlike graphical user interfaces, dialogue systems have no visible menu-structure to display next possible moves [32]. The expectations that users do have about artificial intelligence are often wrong [17]. So the design should guide the user on how to control the interaction. The aim is to make the system transparent to the user. A system is called *transparent* when the user's mental model of the system's intent (purpose), beliefs (current internal state) and capabilities (way of working), corresponds to its actual purpose, state and capabilities [22]. A transparent system should inspire trust [39]. Note that transparency is part of many frameworks for ethical AI e.g. [14] and of the AI legislation proposed by the EU [10].

Consider for example an interactive recommender system in the food and health domain: it selects a recipe on the basis of user preferences and general knowledge about food and health. After that, the system allows feedback and provides explanations in an interactive manner [6]. Explainable recommender systems such as these, need to be *evaluated*. Claims about the usefulness of the recommendations and about the relevance and comprehension of explanations, and the overall effect on the transparency of the system and ultimately on the trust that users have in the system and its recommendations, must be measured. That suggests the following research question:

Can we define a system of metrics to evaluate explainable recommender systems, in the food and health domain?

The research method for this paper is conceptual and is mostly based on a literature study. Currently, there is no consensus in the literature on how to evaluate effectiveness of explainable AI [15,36]. For instance, there is a debate whether one should use *subjective measures* [8,37], or *objective measures*. There is not even consensus on the main concepts, such as explainability, transparency, or trust [43]. So before we can define metrics for evaluation, we must first analyze these concepts and how they relate. So we will discuss several conceptual models and define metrics for the main concepts.

The intended research contribution of this paper, is twofold: (i) to provide clarity on the main concepts used in explainable recommender systems, in particular explainability, transparency, and trust, and (ii) to define metrics, that can precisely and reliable measure these concepts, so explainable recommender systems can be evaluated.

We realize that the context in which a system is used, determines the way a system must be evaluated. In order to illustrate and guide our definitions for a specific context, we will use an example of a specific explainable food recommender system [6].

The remainder of the paper is structured as follows. Section 2 starts with a review of evaluation methods of interactive systems in general, and about explainable recommender systems in particular. After that, we will briefly detail the case in Sect. 3. In Sect. 4, we specify a series of conceptual models, and define the required set of metrics. The paper ends with a list of challenges and a number of recommendations.

Table 1. Comparing objective and subjective system evaluation

	Objective measurement	Subjective measurement
Purpose	measure *task success* of interaction with the system on the basis of observation and log-files	measure *perceived success* of interaction with the system, on the basis of user studies and questionnaires
Way of working	Annotators assess interaction behaviour according to definitions.	Users fill in questionnaires with Likert scales, open or closed questions or card sorting tasks
Metrics	task completion rate, comprehension, duration, misunderstandings	perceived usefulness, perceived ease of use, user satisfaction, trust, transparency

2 Overview

In the following sections, we review some of the literature on evaluating interactive systems in general, and explainable recommender systems in particular. We discuss a number of issues and dilemmas. The argument is largely based on Hoffman et al. [15], and Vorm and Combs [39]. The Q-methodology [29] is also discussed.

2.1 Subjective or Objective Evaluation

Suppose we want to evaluate the effectiveness of a system design in context. To evaluate effectiveness, we first need to define the objectives of the system. Given the objectives, there are two ways in which we can collect evidence of the effectiveness of a system: *subjective*, by asking end-users about their experiences in interviews or questionnaires [8,37], or *objective*, by having developers observing functionalities and system behaviour directly, or from log-files [12,42]. Table 1 lists examples of both. Observe that objective measures test specific features in development, whereas subjective measures look at the total impression of the system on the user.

In general, we believe that we need both perspectives: they have separate purposes and strengthen each other. For example, suppose subjective evaluation reveals that most users like the system (high user satisfaction), but some group of users do not. Analysis of the log-files may show, that the dialogue duration for the users who did not like the system, is longer than for those who liked it. In that case we found that satisfaction (subjective) depends on duration (objective). We can even go further and analyze the longer dialogues in detail. Perhaps, a specific type of misunderstanding causes delays. In that case, the system can be re-designed to avoid such misunderstandings. It is also possible that the objective and subjective measures diverge. In that case, it depends on the purpose of the evaluation, which type of measure takes precedence. For testing individual system features, objective measures remain useful, even if end-users do not perceive the differences. But for over-all system success, it is user satisfaction that counts. For more on this discussion in the context of evaluating explainable AI, see [15] and [36, p3].

2.2 Technology Acceptance

One of the most influential models for evaluating information systems is the Technology Acceptance Model (TAM) [8], and later adjustments [37]. Note that often, the

terms 'adoption', 'acceptance' and 'use', are used interchangeably, although they are not completely equivalent. We start from a simple model of psychology: when people make a decision to perform some action, they first form an attitude towards that action. If that attitude is positive, or more positive than for alternative actions, they will form the intention to pursue the action. That intention will then produce that action.

Which attitudes affect the intention to use a system? The original idea is very simple, which also explains its attractiveness (Fig. 1(a)). In deciding to use or adopt a system, people make a trade-off between expected benefits and expected costs or efforts in using it. If the system is helpful in performing the task (useful), the user is more likely to consider using it. However, using the system will also take some effort. One has to learn to use the system and there may be misunderstandings and delays (ease of use). However, when considering to use a system, the user does not yet know the system. That means, that the intention to adopt a system is usually based on a system description. Therefore, the model uses 'perceived ease of use' and 'perceived usefulness' as the main variables, and not the actual usefulness, or actual ease of use. Subjective judgements like 'perceived usefulness' can be measured using Likert-scales. Well-tested and practical questionnaires exist, and results can be statistically analyzed, for example using regression models. There are no time-series, for example, or feedback loops. This simplicity partly explains the popularity of TAM models.

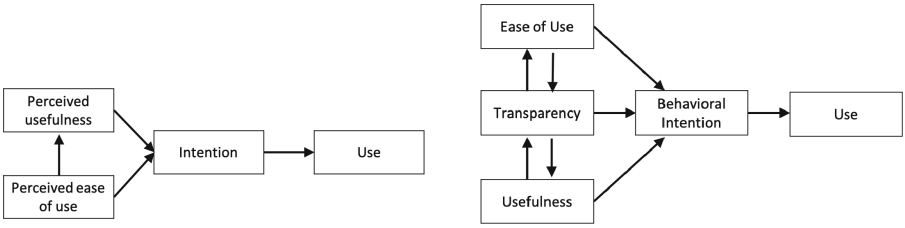

Fig. 1. (a) Technology Acceptance Model [8] and (b) ISTAM Model [39]

The Technology Acceptance Model also has clear disadvantages. The model only looks at the individual user, not at the corporate or social environment in which the system will be used. The model is about the decision to use a system, beforehand. It does not evaluate actual usage, afterwards. Moreover, the model suggests that intentions always lead to successful action; it doesn't look at feasibility. In the TAM model, technology is seen as a black-box. There is no evaluation of the effect of design choices and specific functionalities. Furthermore, the model is psychologically too simple. For example, it does not cover learning effects, habits, or previous experience.

Some of these disadvantages have been addressed in later adjustments and improvements to the model. In particular, the unified model of Venkatesh et al. [37] adds variables for social influence, and facilitating conditions. In addition, control variables for gender, age, experience and voluntariness of use, are taken into account.

It is relatively easy to add additional variables to TAM models. For example, trust has been added in the context of e-commerce systems [30]. System Usability Scale (SUS) [20] is a well-known alternative for TAM-models. It measures usability, which combines both ease of use, and usefulness in a single scale of 10 questions.

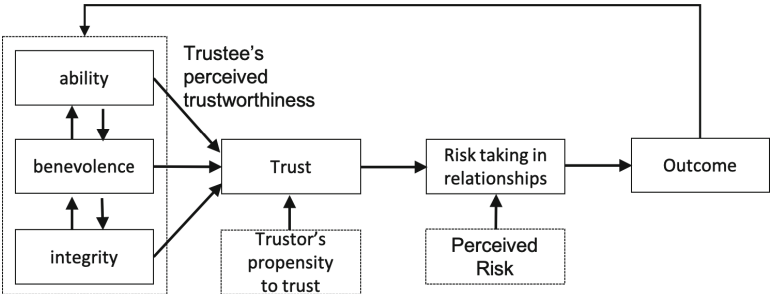

Fig. 2. Trust and Feedback [25]

In the same research tradition, Vorm and Combs [39] extend the TAM model, but now for evaluating intelligent systems (Fig. 1b). The notion of transparency is added as the key intermediate variable, that influences the behavioural intention. Vorm and Combs discuss various conceptions of transparency. Based on earlier work [4], they make a distinction between transparency for monitoring (i.e., what is happening?), transparency for process visibility (i.e., how does it work?), transparency for surveillance (i.e., interactivity and user control) and transparency for disclosure (i.e., opening up secrets regarding purpose). The relation to trust is also discussed. Vorm and Combs [39] state that the role of trust in the model more or less overlaps with transparency: "Transparency factors play moderating and supporting roles that combine to influence trust, and ultimately acceptance" (page 14). The resulting model is what they call the Intelligent Systems Technology Model (ISTAM), see Fig. 1(b).

What we gather from these discussions [22,23], is that transparency involves at least three aspects: (i) the purpose or goal of the system and the specific interaction, (ii) the current internal state of the system, and state of the interaction, and (iii) how the system works, and ways for the user to control the interaction.

2.3 Trust

Trust has been discussed in many disciplines and fields. Here, we will follow the tradition in economics, that relates trust to the willingness to take risk in collaborating with another person, but without additional guarantees or controls over that other person's behaviour. The propensity to take risks, is part of the character of the trustor. We can also look at trustworthiness, the properties that are needed for the trustee to be trusted. Mayer et al. [25] define three properties of trustworthiness: (i) ability (or competence): can the trustee perform the task, (ii) benevolence: does the trustee want to do good to the trustor, and (iii) integrity: does the trustee follows a set of personal principles?

Trust is a relationship, so it depends both on aspects of the trustor and the trustee. In general, the likelihood that a trustor will trust a trustee will depend on (i) the trustor's propensity to trust, and (ii) the trustee's perceived trustworthiness (ability, benevolence and integrity). This is a nice definition, but it doesn't tell us how trust is won or lost. Which signals inspire trust in a person? What is the effect of repeated interactions? Mayer et al. [25] show an interactive model, that allows feedback (Fig. 2). The outcome

of a (repeated) event, will influence the future trustor's assessment of the trustee' trustworthiness. In general, when the outcome is positive, this will increase trust; when the outcome is negative, this will reduce trust, for the next time around.

Lewicki and Bunker [18] study trust in work relationships. Based on older models of trust in personal relationships, they conclude that trust develops in three stages: calculus-based trust, knowledge-based trust, and identification-based trust (Table 2).

Now we need to map these models of inter-personal trust, to trust in machines. The regularity that underlies calculus-based trust is also the main source for trust in a machine [38]. For example, I trust a coffee machine to give me coffee, based on previous experiences, or on testimonies from other people. Such trust based on testimonies of a group of people is often called *reputation* [11]. It is also possible to use knowledge in trusting a machine. For example, I have a naive mental model: the weight of the coin will tip a leaver, that triggers release of a paper cup, coffee powder and hot water. True or not, that mental model allows me to operate the machine. I also have knowledge about the purpose. I trust the machine will give me coffee, because I know that is the vendor's business model. Moreover, I trust that some regulator has put safety regulations into place. We do not believe it is possible to use identification-based trust in the case of machines, at least not with current state of the art in artificial intelligence.

This example shows that theories about trust, especially calculus-based trust (regularities) and knowledge-based trust (mental model), are similar to theories about transparency [22]. Previous experience, as well as knowledge about the design, about the internal state, and about the purpose of the machine will induce trust in the machine.

That ends our discussion of trust. We may conclude that trust is an important factor that influences the intention to use a system or continue to use a system. We distinguish trust in the machine, mediated by knowledge of the design, the internal state, and the purpose of the machine, and institutional trust in the organizations that developed the machine, and that now operate the machine. We can also conclude that there are many parallels between trust and transparency, and that trust in a machine, depends on transparency of the system design. However, unlike Vorm and Combs [39], we do not believe we can reduce trust to transparency. Transparency is a system property, a requirement, that can be designed and tested for, whereas trust is a user attitude, but also an objective to achieve by designing the system in a certain way. That means, that to evaluate effectiveness of the design, these variables should be measured independently.

Table 2. Trust develops in stages [18]

	Calculus-based trust	Knowledge-based trust	Identification-based trust
based on	consistency of behaviour; repeated observations	knowledge of beliefs and goals that underlie behaviour	identification with values and background
example	coffee-machine	chess opponent	former classmate
usage	allow users to inspect what happened (trace)	help users build a mental model; explain	build a relationship

2.4 On Evaluation

We discussed models of trust and transparency, and ideas about evaluation. How can all of this be put together? In this section we discuss the model of Hoffman et al. [15], that was influential in the discussion on evaluation of explainable AI systems (Fig. 3).

In yellow, the model shows a flow. The user receives an initial instruction. This affects the user's initial mental model. The instruction is followed by an explanation, which revises the user's mental model, and subsequently enables better performance. Can we adjust the model Hoffman et al. [15] for recommendation? Yes. As we have seen, explainable recommendation dialogues proceed in three stages: (i) collection of user preferences, (ii) recommendation, and (iii) feedback and explanation. Therefore we have added a step, shown in dark yellow, for recommendation. That means that implicitly, the evaluation of the first step, elicitation of user preferences, is part of the evaluation of the second step, recommendation.

In green, the model shows how to measure these variables. In particular, effectiveness of the explanation is tested by *goodness criteria*, which can be assessed by developers, on the basis of log-files, and a *test of satisfaction*, a subjective measure, asking end-users whether they are satisfied with the explanation. The effect of an explanation on a user's mental model is tested by a *test of comprehension*, similar to an exam question: is the user's mental model in line with reality? Finally, the effect on performance can be a tested by a *test of performance*, related to the task. A recommendation is also evaluated by a goodness criteria, and by user satisfaction, shown here in dark green.

In grey, the model shows the expected effect of explanations on trust. Initially, the user may have inappropriate trust or mistrust, based on previous conceptions. After recommendation and explanation, the user's mental model changes, and leads to more appropriate trust, which enables more appropriate use of the system.

Goodness criteria measure the success conditions, in this case of an explanation. For example: the given explanation must match the type of explanation that was asked for. If the user wants to know how the system works, it should not the purpose. These criteria can be assessed relatively objectively by comparing specified functionality with the behaviour shown on the log-files. At least two coders should verify inter-coder

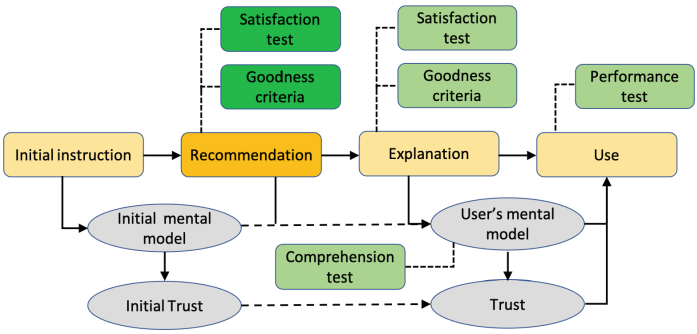

Fig. 3. Evaluating explainable AI in various stages, adjusted from [15]. Components in dark yellow and dark green are added here for explainable recommendation (Color figure online)

Table 3. Goodness criteria for evaluating an explanation [15]

The explanation helps me **understand** how the [software, algorithm, tool] works	Y/N
The explanation of how the [software, algorithm, tool] works is **satisfying**	Y/N
The explanation of the [software, algorithm, tool] sufficiently **detailed**	Y/N
The explanation of how the [software, algorithm, tool] works is sufficiently **complete**	Y/N
The explanation is **actionable**, that is, it helps me know how to use the [software, algorithm, tool]	Y/N
The explanation lets me know how **accurate** or **reliable** the [software, algorithm] is	Y/N
The explanation lets me know how **trustworthy** the [software, algorithm, tool] is	Y/N

agreement on the scores [42]. Hoffman et al. provide a list of goodness criteria for explanations (Table 3). This is meant as an objective measure. However, we can see the list is written from the perspective of an end-user, so it looks like a subjective measure. What to do? First, these criteria can be re-used in user satisfaction test. Second, we can in fact define objective criteria. We will show that in Sect. 4.

A *test of satisfaction* for an explanation aims to test "the degree to which users feel that they understand the AI system or process being explained to them." According to Hoffman et al., user satisfaction is measured by a series of Likert scales for key attributes of explanations: understandability, feeling of satisfaction, sufficiency of detail, completeness, usefulness, accuracy, and trustworthiness. As discussed above, satisfaction seems to overlap with the goodness criteria. Hoffman et al. explain the difference as follows. Relative to goodness, satisfaction is contextualized. It is measured after the interaction, all factors included. The measurements are meant for a different audience. The goodness test is meant for developers and the satisfaction test is for end-users.

A *test of comprehension* aims to test the effectiveness of the explanation on the mental model. Similar to an exam question: is the user able to remember and reproduce elements of the explanation? For example, users can be asked to reproduce how a particular part of the system works, reflect on the task, or be asked to make predictions, for which knowledge of the system is needed. There are many ways in which mental models can be elicited [15, Table 4]. Consider methods like think aloud protocols, task reflection (how did it go, what went wrong?), card sorting tasks (which questions are most relevant at this stage?), selection tasks (identifying the best representation of the mental model), glitch detector tasks (identifying what is wrong with an explanation), prediction tasks, diagramming tasks (drawing a diagram of processes, events and concepts), and a shadow box task (users compare their understanding to that of a domain expert). Various methods have to be combined, to make tests more reliable.

Finally, a *test of performance* aims to objectively test over-all effectiveness of a system. One could take the *success rate*: count the number of successfully completed dialogues, relative to the total number of dialogues. For goal-directed dialogue, progress towards the goal can be measured objectively. Consider for example a system applied in retail [33]. Here, the conversion rate is a measure of success: how many potential customers end up buying a product. We can also try to evaluate *communicative success*. The effectiveness of an explanation is inversely proportional to the number of misunderstandings. Thus, one could identify indicators of misunderstanding (e.g. overly long duration, signs of frustration, aborted dialogues), and count the relative number of such

Table 4. Trigger questions [15]

	Triggers	User/Learner's Goal
1.	How do I use it?	Achieve the primary ask goals
2.	How does it work?	Feeling of satisfaction at having achieved an understanding of the system, in general (global understanding)
3.	What did it just do?	Feeling of satisfaction at having achieved an understanding of the system, in general (local understanding)
4.	What does it achieve?	Understanding of the system's functions and uses
5.	What will it do next?	Feeling of trust based on the observability and predictability of the system
6.	How much effort will this take?	Feeling of effectiveness and achievement of primary task
7.	What do I do if it gets it wrong?	Desire to avoid mistakes
8.	How do I avoid the failure modes?	Desire to mitigate errors
9.	What would it have done if x were different?	Resolution of curiosity at having achieved an understanding of the system
10.	Why didn't it do z?	Resolution of curiosity at having achieved an understanding of the local decision

misunderstandings. The over-all purpose of a recommendation system is to *convince* the user, and perhaps even to induce them to change behaviour. Objectively establishing such a change of behaviour is the ultimate test of success. That concludes our discussion of Hoffman et al. [15]. It serves as a good basis for designing evaluation experiments.

In a more recent paper, Van der Waa et al [36] discuss how to evaluate explainable AI. They conduct experiments, comparing two types of explanations: rule-based and example-based. These explanations are compared on user satisfaction and general system performance. They also discuss the advantages of combining subjective measures with more detailed behavioural analysis, on the basis of observable behaviour.

Another technique for subjective measurement is the *Q-methodology*, from HCI [29]. Using the trigger-questions in Table 4 from [15], Vorm and Miller [40] suggest to evaluate explainable systems by having the user select the question, which they would like to ask at that point in the interaction. Users are asked to sort 36 cards with questions. Vorm and Miller carefully developed the question bank. For example: "How current is the data used in making this recommendation?" or "Precisely what information about me does the system know?". Factor analysis determines specific groups of users with similar preferences. In this way, four groups of users are found [40]: 1: Interested and Independent, 2: Cautious and Reluctant, 3: Socially Influenced, and 4: Egocentric. This shows that different types of users have various needs for explanations. A system should be flexible enough to handle these needs.

3 Application

In this section, we discuss a specific system, that is currently being developed [6]. The system is an explainable recommendation system, for the food and health domain [35]. The system is interactive and provides personalized recommendations and explanations. The system is developed in two versions: a web-based platform allowing the users to experience both the explanation-based interactive recommender and its replica without the explanations and critiques component (i.e., a regular recommender). This allows us to assess the effectiveness of explanation and of interaction.

A user interaction involves three stages, with the following success conditions.

- Stage 1. *User preference elicitation*. Ask user about preferences. Afterwards, the system must know enough user preferences to select a recipe; preferences are consistent and are correctly understood.
- stage 2. *Recommendation of a recipe.* The recipe must fit the user preferences, and follow from knowledge about food, recipes, and healthy lifestyles.
- Stage 3. *Explanation and interaction*. The explanation must fit the user's request. The explanation must be personalized to the user's preferences and be relevant in context, and the subsequent interaction must be coherent.

If we classify the system, we can say that the *application domain* is food and health. The *task* is recommendation, but it also involves elements of *persuasion*, as we intend users to follow a healthy lifestyle. In some cases, that means convincing the user and making them change behaviour. In other words, the system is intended as a nutrition virtual coach (NVC) [35]. In order to persuade the user, trust and transparency are crucial.

Persuasion is often about breaking a habit. What seems to work, based on conversations with nutritionists, is to set personal goals, and help users attain those goals, by measuring the current state, the distance to the goal, and suggesting ways of getting closer. Measuring weight can be used to quantify progress towards the goal, and calories are used to quantify the required energy intake of a meal. Long-term relationship building, as required for a nutrition virtual coach, is out of scope for this research prototype, but it does play a role in the over-all design of the system, and in future research.

In this domain, generally we find that explanations are of two types: preference related explanations, which are based on the user preferences which were inferred or stated just before, or health related explanations, which are based on general knowledge about food an health [6]. Here we show an example of each.

- *Health-related*: Protein amount covers user needs for a meal.
  "This recipe contains X grams of protein, which is about Y % of your daily requirement. Your body needs proteins. Consuming the necessary amount is important!"
- *Preference-related*: User's chosen cuisine matches the recipe.
  "This recipe is a typical part of the cuisine you like: Z."

4 Towards Metrics

In this section, we specify the metrics to evaluate claims about effectiveness of an explainable recommender system, in the context of the food and health domain, as detailed in Sect. 3. Consider the research model in Fig. 4.

On the right, interpret effectiveness as the direct effect of an interaction on the user, in terms of user satisfaction (performance), transparency and trust. That aspect refers to the recommendation part of the task, and also the explanation. In addition, repeated interactions should have an indirect effect on the user, in terms of a change of behaviour, for instance a healthier choice of food. That aspect refers to the persuasion task.

On the left, the system design is detailed. We see a system as a white-box, with separate functionalities. Each of the modules may have an effect on the user interaction.

These modules are: the algorithm for generating a recommendation, the knowledge base about health and food, the goals and plans of the system during interaction, the user interface, the user model that represents user preferences, and the data set with all the recipes that can be recommended. Each of these modules must be evaluated separately and as part of the system (unit test; integration test).

In the middle, we discuss two moderator variables, that may strengthen or weaken the effect of the system design on the success variables. First, *explanation*, whether the system is able to provide explanations about its recommendations. Second, *interaction*, whether the systems allows feedback and interaction about recommendations and explanations that fit the context. These are the interventions that we want to study.

This model can test the over-all effect of recommendations, and the specific effect of explanations and interactive dialogues on user satisfaction, transparency and trust, and ultimately, on behavioural change. However, the model also has some disadvantages. The model disregards several variables that are familiar from TAM, in particular perceived usefulness and perceived ease of use and the intention to use a system. Model (1) focuses not so much on the decision to start using a system (as in TAM), but rather on evaluating actual use of a system. In addition, the 'effectiveness' variables (user satisfaction, transparency, trust) need to be worked out in more detail.

Therefore, we developed the model in Fig. 5. On the left, we see the various modules that make up the recommender system design. If these modules function effectively, they have a positive influence on transparency. In addition, the system design has a direct effect on the use of the system (long red arrow). Transparency in turn has an effect on the perceived usefulness and perceived ease of use, which in turn affect the intention to use, and usage itself, as in the ISTAM model [39]. It is also possible that the system design has a direct effect on perceived usefulness and perceived ease of use. Transparency is expected to have an effect on trust. After all, the perceived competence, benevolence and integrity of the system and organization that deploys the system, are mediated by the interface. Trust, in turn, has an effect on the intention to use, and ultimately, on the suggested behaviour change. Finally, we also consider a feedback loop, back from usage to trust. However, such feedback loops are difficult to test for.

Like in Model (1) we test for two moderator variables: explanation and interaction. These features are expected to affect transparency, and indirectly affect perceived ease of use and perceived usefulness, as well as trust. Moreover, they are expected to have a direct effect on use (success rate and failure rate).

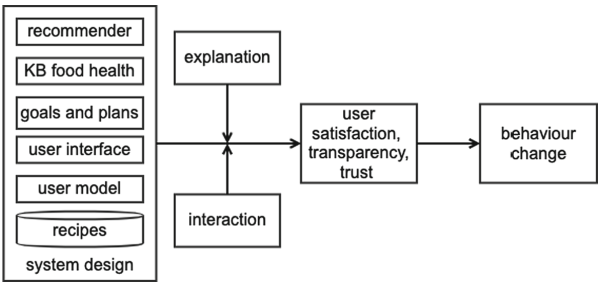

Fig. 4. Towards a Model for Evaluation (1)

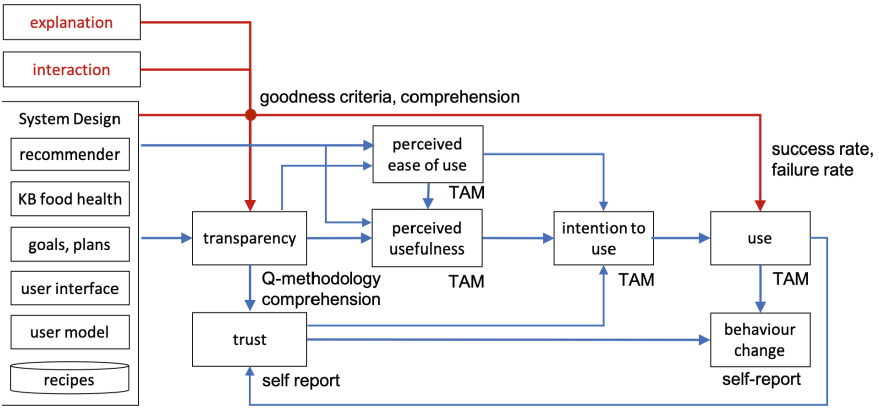

Fig. 5. Towards a Model for Evaluation (2)

Accordingly, we propose a table of metrics for each these variables (Table 5). Symbol '–' means that the variable is inversely related to success. 'Goodness criteria' refers to the tables with specific goodness criteria per functionality, as discussed in Sect. 4.1. TAM instruments refers to established questionnaires.

Table 5. Metrics to measure variables in Model for Evaluation (2)

Variable	Measures
system design	goodness criteria, comprehension
	ease of use: – duration, – number of misunderstandings, – time to learn
	usefulness: success rate, – failure rate
transparency	Q-methodology
	comprehension
	goodness criteria: shows purpose, internal state, and how it works
trust	self-report, willingness to recommend
perceived ease of use	TAM instruments
perceived usefulness	TAM instruments
intention to use	TAM instruments
use	counting, TAM instruments
behavioural change	self-report

4.1 Goodness Criteria for Recommendation and Explanation

An important part of the evaluation methods depend on requirements or goodness criteria for the various components and functionalities. The most important functionalities are the ability to provide a recommendation, and the ability to provide an explanation, in

an interactive manner. Under what conditions can we say that a system has successfully achieved these objectives?

Table 6. Gricean maxims for cooperative communication [13, p 46]

Quantity	1. Make your contribution as informative as is required (for current purposes)
	2. Do not make your contribution more informative than is required
Quality	Try to make your contribution one that is true
	1. Do not say what you believe to be false
	2. Do not say that for which you lack adequate evidence
Relation	Be relevant
Manner	Be perspicuous
	1. Avoid obscurity of expression
	2. Avoid ambiguity
	3. Be brief (avoid unnecessary prolixity)
	4. Be orderly

The following example shows the type of functionality that we develop. How suitable is the explanation in this dialogue?

U. I'd like some toasted white bread for breakfast.
S. You should eat whole meal bread.
U. Why?
S. Because you said you wanted to lose weight, and eating whole meal bread instead of white bread is a good way to reduce the number of quick calories per meal, and it is well known that reducing the number of quick calories per meal will help you lose weight.

We start by clarifying the relation between explanations and transparency. Transparency is a property of a system. The system must reveal its purpose, inner state, and how it works [22]. Transparency is not a property of an explanation, except in the sense of 'clarity' or 'being based on evidence'. Instead, part of the purpose of having explanations, is for the system to be more transparent. There are other methods to make a system more transparent too, such as a user manual, a suitable persona, etc.

4.2 Good Explanation

What makes a good explanation? An explanation is a form of assertion. That means, that we can follow Grice's maxims for cooperative communication [13, p. 46]: quantity, quality, relation and manner (see Table 6). The point about manner, specifically to be brief, is also made by Mualla et al [28], who advocate parsimonious explanations.

There is a lot of research on what makes a good explanation, in various fields. Properties of everyday explanations are summarized in a survey paper by Miller [26]:

1. Explanations are *contrastive*: distinguish an outcome from counterfactual outcomes.
2. Explanations are *selected* from a range of possible reasons.
3. Explanations are not necessarily based on probabilities, but rather on *narratives*.
4. Explanations are social, and are usually part of *interactions*.

These four characteristics can be summarized by stating that explanations are inherently *contextual*. We will discuss them one by one. Ad 1. An explanation must not only be generic (e.g. based on laws of nature), but also involve specific facts of the case of the user, that show why other alternative advice is not given [41]. Ad 2. What are reasons? For natural events, they are causal histories built from facts and natural laws [19]. For human behaviour, they are goals which can be inferred by abduction, because they most likely motivate those actions in that context [24]. In our case, the reasons are ingredients which match user preferences. Other reasons are natural laws of nutrition (vegetables have low calories; pasta has high calories), and motivational goals of the user, to maintain a certain weight, for example. Ad 3. Miller [26] criticizes some technical research on intelligible algorithms, which focuses scientific explanations. A doctor would justify a treatment to a colleague using probabilities, but for lay people, stories often work better. Ad 4. Interactive dialogues with explanations are preferred, because they give the user more control. In case of a problem, the user can just ask.

Generally, there are several levels or successive rounds of explanation.

- Level 1. Why this recommendation? Because of *facts* (user preferences, ingredients, recipes) and a *rule* (knowledge about food and health)
- Level 2. Why that rule? Because the rule is true and relevant relative to a *goal*. Why those facts? Because the procedure for selecting these facts is *valid*.
- Level 3. Why that goal? Because the goal (promote healthy choices) helps to promote social values (health), which represent who we are (virtual nutritionist).

This example of explanation levels is based on value-based argumentation [3], which also has three levels: (1) actions, facts and rules, (2) goals, and (3) social values.

The properties of explanations discussed so far, are relatively abstract. How can they be built into algorithms? Rosenfeld [31] presents four metrics for evaluating explainable artificial intelligent systems: D, R, F, and S. Here D stands for the performance difference between the black-box model and a transparent model, R considers the size of the explanation (i.e., number of rules involved in given explanations), F takes the relative complexity into consideration, by counting the number of features to construct an explanation, and D measures the stability of the explanations (i.e., ability to handle noise perturbations) [31]. In the context of explaining recommendations, we can measure the following aspects:

- *Improvement Effect:* Test the system with and without explanations and observe the effect of explanations on system performance. We can list a number of performance metrics such as acceptance rate, average acceptance duration, and average number of interactions spent for acceptance of the recommendation.
- *Simplicity of Explanations:* This can be measured with the length of the explanations and to what extent that can be grasped by the user (comprehension).

- *Cognitive Effort Required:* An explanation may focus on a single decision criterion (e.g., only nutrition levels) to reduce the user's cognitive effort. Some explanations may point out several criteria (e.g., nutrition levels, user's goals such as losing weights, and their preferences on ingredients) at one time, which may increase the cognitive load. We can count the number of criteria captured in a given explanation.
- *Accuracy of Explanations:* The system generates recommendations based on its objectives and its beliefs about the user's goal and preferences. What if it is wrong? Then, the system may generate explanations which conflict with actual preferences. In such case, users give feedback on the explanations by pointing out their mistakes. We can analyze the given feedback to determine the accuracy of the explanations.

Table 7. Goodness criteria for recommendation, adjusted from criteria in Table 3 [15].

The recommendation **helps me to decide** [what action to do/which recipe to cook]	Y/N
The recommendation [what action to do/which recipe to cook] is **satisfying**	Y/N
The recommendation [what action to do/which recipe to cook] is **sufficiently detailed**	Y/N
The recommendation [what action to do/which recipe to cook] is **sufficiently complete**	Y/N
The recommendation is **actionable**, that is, it helps me to carry out my decision	Y/N
The recommendation lets me know how **accurate** or **reliable** the [action/recipe] is	Y/N
The recommendation lets me know how **trustworthy** the [action/recipe] is	Y/N

4.3 Good Recommendation

In recommendation systems, performance metrics are often borrowed from information retrieval: *precision* (fraction of given recommendations that are relevant) and *recall* (fraction of potentially relevant recommendations that are given). One can balance precision and recall, by means of the F-measure. Alternatively, people use the area under the Receiver Operator Characteristic (ROC) curve, to measure how well the algorithm scores on this trade-off between precsion and recall, see [34].

What makes a good recommendation? We can see a recommendation as a response to a request for advice. The same Gricean maxims apply (Table 6). In the context of our application that suggest the following requirements.

- *Quality*: the recipe must be an existing recipe, and fit the agreed dietary goals.
- *Quantity*: the recipe must be detailed enough to be able to make it. All ingredients and quantities must be listed and clear.
- *Relation*: the recipe must respond to the request of the user and fit the context. Specifically, the recipe must match the user preferences, if such recipes exist. If no such recipes exist, a clear no-message must be given.
- *Manner*: the recipe is presented clearly and with diagrams or photographs to illustrate. The recipe must not be too long or detailed [28].

The goodness criteria for recommendations are similar to those for explanations in Table 3. For comparison, we have adjusted them to fit recommendations (Table 7).

5 Discussion

Building explainable recommender systems in the food and health domain, has ethical consequences. This is why we care about explainability, transparency and trust. In a previous paper, we have given a survey of ethical considerations for nutritional virtual coaches [7]. Here, we will discuss a few examples.

First, the factual *information* about food and ingredients must be true, informative, and relevant (Gricean Maxims). The data set and knowledge bases used must be fair and present a representative coverage of foods and tastes. This is not trivial, as food is related to culture and identity. The system will collect personal data from the user, namely food preferences and health related data. These data are sensitive, and must be adequately protected. We observe a trade-off between privacy and relevance. If more detailed personal data is collected, a better recommendation can be made. A metric to test this balance, is to check how many requested data items, are actually used.

Second, the system makes recommendations and provides explanations. A related ethical issue is *control*: in case of a conflict between user preferences and healthy choices, who determines the final recommendation? Suppose the user asks for a hamburger? Suggesting a more healthy alternative may be seen as patronizing. Here, we believe the solution is to be transparent: whenever a requested unhealthy choice is *not* recommended, this must always be explained. The explanation is contrastive. Moreover, the recommendation must be in line with the stated purpose and 'persona' of the system: chef (good food) or nutritionist (advice).

Another ethical issue is *sincerity*. A recommendation or explanation must be trusted not to have a hidden purpose, like commercial gain [21]. If a system provides a clear explanation, a user can verify that reason. Moreover, if the explanation is contrastive [26], indicating why some recipes are not shown, and interactive, allowing users to vary the request to see how that affects the recommendation, this will make the reasoning mechanism transparent, and make it easier to detect a hidden purpose [41].

Third, the systems aims to *persuade* the user, for instance to make healthier choices for food. To do so, the system makes use of argumentation techniques. An ethical issue is how much persuasion we accept from a machine. Again, this depends on the stated purpose and persona of the system (e.g. chef or nutritionist). Who is ultimately responsible? Here, the answer is to develop the system as a tool to support a nutritionist in coaching a large group of clients. After deployment a qualified nutritionist should remain as *human-in-the-loop*, with meaningful control over the persuasion process.

To summarize, an explainable recommender system offers many opportunities for manipulation [21]. Manipulation is harder to achieve, if the system is transparent: the user can verify and compare the stated purpose with actual behavior.

6 Conclusions

In this paper, we have discussed models and metrics for evaluation interactive explainable recommender systems. We pointed out the debate between subjective measurement (perceived ease of use, perceived usefulness, user satisfaction) and objective measures (goodness criteria, task success, misunderstandings). We argue that subjective and

objective evaluation strengthen each other. For example, consider the following design principle: users who experience a misunderstanding are less likely to be satisfied. Misunderstandings might be clear from the log-files. Satisfaction depends on users. So, in order to test this design principle, one needs to compare both objective and subjective evaluation metrics.

The model of [15] forms a good basis to develop evaluation metrics for explainable recommender systems, except that the notion of 'goodness criteria' needs to be worked out, and must be more clearly separated from user satisfaction. For acceptance testing and user evaluations, the famous TAM model is relevant [8,37]. Trust can be added. Following Vorm and Combs [39] we believe that transparency is crucial, and that trust is largely influenced by transparency. However, unlike [39] we believe trust is a separate notion, that can be measured, by subjective measures.

So, our evaluation model is based on three components: (i) the ISTAM model [39], which combines TAM and Transparency, (ii) trust [25] and specifically trust in machines [38], and (iii) various objective measures, such as goodness criteria and fit to the context, success rate, number of misunderstandings, and over-all performance [15].

Especially for applications in the food and health domain, building an explainable recommender system has important ethical considerations [7]. An important part of the solution is to provide explanations and be transparent about the system's purpose and way of working. This should allow the user to verify the behaviour of the system and decide if this forms a basis to trust the system and the recommendations it makes.

Acknowledgments. This work has been supported by CHIST-ERA grant CHIST-ERA19-XAI-005, and by (i) the Swiss National Science Foundation (G.A. 20CH21_195530), (ii) the Italian Ministry for Universities and Research, (iii) the Luxembourg National Research Fund (G.A. INTER/CHIST/19/14589586), (iv) the Scientific and Research Council of Turkey (TÜBİTAK, G.A. 120N680).

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. **17**, 734–749 (2005)
2. Anjomshoae, S., Calvaresi, D., Najjar, A., Främling, K.: Explainable agents and robots: results from a systematic literature review. In: Autonomous Agents and Multi Agent Systems (AAMAS 2019), pp. 1078–1088 (2019)
3. Atkinson, K., Bench-Capon, T., McBurney, P.: Computational representation of practical argument. Synthese **152**(2), 157–206 (2006)
4. Bernstein, E.: Making transparency transparent: the evolution of observation in management theory. Acad. Manag. Ann. **11**(1), 217–266 (2017)
5. Burke, R., Felfernig, A., Göker, M.H.: Recommender systems: an overview. AI Mag. **32**, 13–18 (2011)
6. Buzcu, B., Varadhajaran, V., Tchappi, I.H., Najjar, A., Calvaresi, D., Aydoğan, R.: Explanation-based negotiation protocol for nutrition virtual coaching. In: PRIMA 2022. LNCS, vol. 13753, pp. 20–36. Springer (2022). https://doi.org/10.1007/978-3-031-21203-1_2

7. Calvaresi, D.: Ethical and legal considerations for nutrition virtual coaches. In: AI and Ethics, pp. 1–28 (2022)
8. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. **13**(3), 319–340 (1989)
9. V. Dignum. Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer (2019). https://doi.org/10.1007/978-3-030-30371-6
10. European Commission. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts (2021)
11. Falcone, R., Castelfranchi, C.: Trust and relational capital. Comput. Math. Organ. Theory **17**(2), 179–195 (2011)
12. Goodhue, D.L.: Understanding user evaluations of information systems. Manage. Sci. **41**(12), 1827–1844 (1995)
13. Grice, H.P.: Logic and conversation. In: Cole, P., Morgan, J.L. (eds.) Syntax and Semantics, vol. 3, pp. 41–58. Academic Press, New York (1975)
14. HLEG. Ethics guidelines for trustworthy AI (2019)
15. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, O.: Metrics for explainable ai: challenges and prospects. arXiv:1812.04608 [cs.AI] (2018)
16. Jannach, D., Pearl, P., Ricci, F., Zanker, M.: Recommender systems: past, present, future. AI Mag. **42**, 3–6 (2021)
17. Kriz, S., Ferro, T.D., Damera, P., Porter, J.R.: Fictional Robots as a Data Source in HRI Research, pp. 458–463. IEEE (2010)
18. Lewicki, R.J., Bunker, B.B.: Developing and maintaining trust in work relationships. In: Trust in Organizations, pp. 114–139. Sage Publications (1996)
19. Lewis, D.: Causal explanation, pp. 214–240. Oxford University Press, Oxford (1986)
20. Lewis, J.R., Sauro, J.: Item benchmarks for the system usability scale. J. Usability Stud. **13**(3), 158–167 (2018)
21. Lima, G., Grgić-Hlača, N., Jeong, J.K., Cha, M.: The conflict between explainable and accountable decision-making algorithms. In: FACCT, pp. 2103–2113. ACM, Seoul, Republic of Korea (2022)
22. Lyons, J.B.: Being transparent about transparency: A model for human-robot interaction, pp. 48–53. AAAI (2013)
23. Lyons, J.B., Havig, P.R.: Transparency in a human-machine context: approaches for fostering shared awareness/intent. In: Shumaker, R., Lackey, S. (eds.) VAMR 2014. LNCS, vol. 8525, pp. 181–190. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07458-0_18
24. Malle, B.F.: How people explain behavior: a new theoretical framework. Pers. Soc. Psychol. Rev. **3**(1), 23–48 (1999)
25. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. Acad. Manag. Rev. **20**(3), 709–734 (1995)
26. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
27. Miller, T., Hoffman, R., Amir, O., Holzinger, A.: Special issue on explainable artificial intelligence. Artif. Intell. **307**, 103705 (2022)
28. Mualla, Y., et al.: The quest of parsimonious XAI: a human-agent architecture for explanation formulation. Artif. Intell. **302**, 103573 (2022)
29. O'Leary, K., Wobbrock, J.O., Riskin, E.A.: Q-methodology as a research and design tool for HCI, pp. 1941–1950. ACM, Paris (2013)
30. Pavlou, P.A., Gefen, D.: Building effective online marketplaces with institution-based trust. Inf. Syst. Res. **15**(1), 37–59 (2004)
31. Rosenfeld, A.: Better metrics for evaluating explainable artificial intelligence. In: AAMAS, pp. 45–50, Richland, SC (2021)

32. Smith, R.W., Hipp, D.R.: Spoken Language Dialog Systems: A Practical Approach. Oxford University Press, Oxford (1994)
33. Christina Soyoung Song and Youn-Kyung Kim: The role of the human-robot interaction in consumers' acceptance of humanoid retail service robots. J. Bus. Res. **146**, 489–503 (2022)
34. Tintarev, N., Masthoff, J.: Explaining recommendations: design and evaluation. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 353–382. Springer, Boston, MA (2015). https://doi.org/10.1007/978-1-4899-7637-6_10
35. Trang Tran, T.N., Atas, M., Felfernig, A., Stettinger, M.: An overview of recommender systems in the healthy food domain. J. Intell. Inform. Syst. **50**(3), 501–526 (2018)
36. van der Waa, J., Nieuwburg, E., Cremers, A., Neerincx, M.: Evaluating XAI: A comparison of rule-based and example-based explanations. Artif. Intell. **291**, 103404 (2023)
37. Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D.: User acceptance of information technology: toward a unified view. MIS Q. **27**(3), 425–478 (2003)
38. Vermaas, P.E., Tan, Y.-H., van den Hoven, J., Burgemeestre, B., Hulstijn, J.: Designing for trust: a case of value-sensitive design. Knowl. Technol. Policy **23**(3–4), 491–505 (2010)
39. Vorm, E.S., Combs, D.J.Y.: Integrating transparency, trust, and acceptance: The intelligent systems technology model (ISTAM). Int. J. Hum.-Comput. Interact., 1–19 (2022)
40. Vorm, E.S., Miller, A.D.: Modeling user information needs to enable successful human-machine teams: designing transparency for autonomous systems. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) HCII 2020. LNCS (LNAI), vol. 12197, pp. 445–465. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50439-7_31
41. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harvard J. Law Technol. **31**(2), 841–887 (2018)
42. Walker, M.A., Litman, D.J., Kamm, A., Abella, A.: PARADISE: A framework for evaluating spoken dialogue agents. In: Proceedings of the 35th Annual meeting of the ACL/EACL, pp. 271–280, Madrid (1997)
43. Wallkötter, S., Tulli, S., Castellano, G., Paiva, A., Chetouani, M.: Explainable embodied agents through social cues: a review. ACM Trans. Hum.-Robot Interact. **10**(3), 27:2–27:24 (2021)