User Model User-Adap Inter (2017) 27:393–444
DOI 10.1007/s11257-017-9195-0

A systematic review and taxonomy of explanations in decision support and recommender systems

Ingrid Nunes[1,2] · **Dietmar Jannach**[2]

Received: 27 March 2017 / Accepted in revised form: 19 September 2017 /
Published online: 5 October 2017

Abstract With the recent advances in the field of artificial intelligence, an increasing number of decision-making tasks are delegated to software systems. A key requirement for the success and adoption of such systems is that users must trust system choices or even fully automated decisions. To achieve this, explanation facilities have been widely investigated as a means of establishing trust in these systems since the early years of expert systems. With today's increasingly sophisticated machine learning algorithms, new challenges in the context of explanations, accountability, and trust towards such systems constantly arise. In this work, we systematically review the literature on explanations in advice-giving systems. This is a family of systems that includes recommender systems, which is one of the most successful classes of advice-giving software in practice. We investigate the purposes of explanations as well as how they are generated, presented to users, and evaluated. As a result, we derive a novel comprehensive taxonomy of aspects to be considered when designing explanation facilities for current and future decision support systems. The taxonomy includes a variety of different facets, such as explanation objective, responsiveness, content and presentation. Moreover, we identified several challenges that remain unaddressed so far, for example related to fine-grained issues associated with the presentation of explanations and how explanation facilities are evaluated.

✉ Ingrid Nunes
ingridnunes@inf.ufrgs.br

Dietmar Jannach
dietmar.jannach@tu-dortmund.de

1 Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil

2 TU Dortmund, Dortmund, Germany

Keywords Explanation · Decision support system · Recommender system · Expert system · Knowledge-based system · Systematic review · Machine learning · Trust · Artificial intelligence

1 Introduction

In recent years, significant progress has been made in the field of artificial intelligence and, in particular, in the context of machine learning (ML). Learning-based techniques are now embedded in various types of software systems. Features provided in practical applications range from supporting the user while making decisions, e.g. in the form of a recommender system, to making decisions fully autonomously, e.g. in the form of an automated pricing algorithm. In the future, a constant increase in such intelligent applications is expected, in particular because more types of data become available that can be leveraged by modern ML algorithms. This raises new issues to be taken into account in the development of intelligent systems, such as accountability and ethics (Banavar 2016).

A key requirement for the success and practical adoption of such systems in many domains is that users must have confidence in recommendations and automated decisions made by software systems, or at least trust that the given advice is unbiased. This was in fact acknowledged decades ago, when *expert systems*, mainly those to support medical decisions, were popular. Since the early years of expert systems, automatically generated *explanations* have been considered as a fundamental mechanism to increase user trust in suggestions made by the system (Ye and Johnson 1995). In these systems, provided explanations were often limited to some form of system logging, consisting of a chain of rules that were applied to reach the decision. Nevertheless, such explanations were often hard to understand by non-experts (Guida et al. 1997), thus being used in many cases only to support system debugging.

Today, with modern ML algorithms in place, generating useful or understandable explanations becomes even more challenging, for instance, when the system output is based on a complex artificial neural network (Samarasinghe 2006). One of the most prominent examples of ML-based applications today are the so-called *recommender systems* (Jannach et al. 2010). These systems are employed, e.g., on modern e-commerce sites to help users find relevant items of interest within a larger collection of objects. In the research literature, a number of explanation approaches for recommenders has been proposed (Herlocker et al. 2000; Gedikli et al. 2014; Carenini and Moore 2006; Nunes et al. 2014; Bilgic et al. 2005), and existing work has shown that providing explanations can be beneficial for the success of recommenders in different ways, e.g. by helping users make better or more informed decisions (Tintarev and Masthoff 2011).

Therefore, how to explain to the user recommendations or automated decisions made by a software system has been explored in various classes of systems. These include expert systems, knowledge-based systems, decision support systems, and rec-

ommender systems, which we collectively refer to as *advice-giving systems*.[1] However, despite the considerable amount of research literature in this context, providing adequate explanations remains a challenge. There is, for example, no clear consensus on what constitutes a *good* explanation (Nunes et al. 2012a). In fact, different types of explanations can impact a user's decision making process in many forms. For instance, explanations can help users make better decisions or persuade them to make one particular choice (Tintarev and Masthoff 2011). Finally, deriving a *user-tailored* explanation for the output of an algorithm that learns a complex decision function based on various (latent) patterns in the data is challenging without the use of additional domain knowledge (Zanker and Ninaus 2010).

Given these challenges, it is important to gather and review the variety of existing efforts that were made in the last decades to be able to design explanation facilities for future intelligent advice-giving systems. Next-generation explanation facilities are particularly needed when further critical tasks are delegated to software systems, e.g. in the domain of robotics or autonomous driving (Wang et al. 2016a, b). At the same time, many future advice-giving systems will need more interactive interfaces for users to give feedback to the system about the appropriateness of the advice made, or to overwrite a decision of the system. In both cases, system-provided explanations can represent a starting point for better *user control* (Tintarev and Masthoff 2011; Jannach et al. 2016; Jugovac and Jannach 2017).

In this work, we present the results of a *systematic literature review* (Kitchenham and Charters 2007) on the topic of explanations for advice-giving systems. We discuss in particular how explanations are generated from an algorithmic perspective as well as what kinds of information are used in this process and presented to the user. Furthermore, we review how researchers evaluated their approaches and what conclusions they reached. Based on these results, we propose a new comprehensive taxonomy of aspects to be considered when designing an explanation facility for advice-giving systems. The taxonomy includes a variety of different facets, such as explanation objective, responsiveness, content and presentation. Moreover, we identified several challenges that remain unaddressed so far, for example related to fine-grained issues associated with the presentation of explanations and how explanation facilities are evaluated.

Our work is different from previous overview papers on explanations in many ways.[2] To our knowledge, it is the first *systematic* review on the topic. We initially retrieved 1209 papers in a structured search process and finally included 217 of them in our review based on defined criteria. This systematic approach allowed us to avoid a potential researcher bias, which can be introduced when the selection of the papers that are discussed is not based on a defined and structured process. Moreover, as opposed to some other works, our review does not only focus on one single aspect, such as

[1] In general, *advice-giving system* is a term that can also include other types of systems, such as conversational agents or autonomous systems. In this paper, however, we use the term to refer only to the four types of systems listed above.

[2] There is a large number of surveys that provide an overview of explanations published elsewhere (Tintarev and Masthoff 2007b, 2011; Swartout and Moore 1993; Chandrasekaran et al. 1989; Dhaliwal and Benbasat 1996; Gregor and Benbasat 1999; Moulin et al. 2002; Lacave and Díez 2002, 2004; Sørmo et al. 2005; Nakatsu 2006; Papadimitriou et al. 2012; Scheel et al. 2014).

analysing the different purposes of explanations (Tintarev and Masthoff 2007b). We, in contrast, discuss a variety of aspects, including questions and derived conclusions associated with explanation evaluations. Finally, the comprehensive taxonomy that we put forward at the end of this work is constructed based on the results of a systematic bottom-up approach, i.e. its structure is not determined based solely on the authors' expertise in the topic.

2 Systematic review planning

A systematic review is a type of literature-based research that is characterised by the existence of an exact and transparent specification of a procedure to find, evaluate, and synthesise the results. It includes a careful planning phase, in which goals, research questions, search procedure (including the search string), and inclusion and exclusion criteria are explicitly specified. Key advantages of using such a defined procedure are that it helps to avoid or at least minimise potential researcher biases and supports reproducibility. Such reviews are common in the medical and social sciences and are increasingly adopted in the field of computer science (Kitchenham and Brereton 2013). We next describe the steps that were taken while planning the systematic review. We follow the procedure proposed by Kitchenham and Charters (2007).

The need for a systematic review To our knowledge, no previous work has aimed at providing a comprehensive overview of existing work on explanations in advice-giving systems using a systematic approach. A number of survey papers exists as mentioned above, but they are either limited to a subjective selection of papers or focused on certain aspects of explanations. Often, existing survey papers simply summarise individual papers and do not consider the developments in the field over time.

Research goals The comprehensive review provided in this paper shall help designers of next-generation advice-giving systems understand what has already been explored over the last decades in different subfields of computer science. Specifically, we aim to investigate and classify: (i) which forms of explanations were proposed in the literature; (ii) how explanations are generated from an algorithmic perspective; and (iii) how researchers evaluated their approaches and what conclusions they reached. By aggregating the insights obtained from the review in a new multi-faceted taxonomy, we aim to provide an additional aid for designers to understand the various dimensions that one has to potentially consider when designing an explanation facility.

Research questions The specific research questions of the review are consequently as follows.

RQ-1: What are the characteristics of explanations provided to users, in terms of content and presentation?
RQ-2: How are explanations generated?
RQ-3: How are explanations evaluated?
RQ-4: What are the conclusions of evaluation or foundational studies of explanations?

Table 1 Digital databases used in the search

Source	URL
ACM Digital Library	http://portal.acm.org
IEEE Xplore Digital Library	http://ieeexplore.ieee.org
ScienceDirect	http://www.sciencedirect.com
Springer Link	http://link.springer.com

Search strategies To find primary studies that are relevant for our review, we selected the databases presented in Table 1, against which we ran different search queries. Other databases that automatically gather information from difference sources, such as Google Scholar, or allow the addition of non-reviewed papers by their authors, such as arXiv, were left out of the scope of our review. Generally, we assumed that peer-reviewed papers published in computer science are mostly stored in our selected databases. Therefore, these other additional databases would mostly provide duplicated studies. Furthermore, we focused on peer-reviewed work in order to have some evidence regarding the quality of the selected studies.

Selection criteria Primary studies retrieved from the databases are filtered using a set of criteria. We consider four inclusion criteria (IC) and five exclusion criteria (EC) to select papers associated with the primary studies to be analysed in our review. The inclusion and exclusion criteria are summarised in Table 2. We are interested in four general *types* of studies. Studies involving: (i) the proposal of a TECHNIQUE to generate new forms of explanations (IC-1); (ii) the description of a TOOL that includes an explanation facility (IC-2); (iii) an EVALUATION or comparison of one or more forms of explanations (IC-3); or (iv) a discussion of FOUNDATIONAL aspects of explanations (IC-4). Throughout the paper, terms highlighted in small caps are used to refer to these study types.

The following additional considerations apply regarding the satisfaction of our inclusion criteria. First, in our work, we are only interested in systems in which explanations are provided for *end users*, usually someone that is actually responsible for a final decision. In some studies, visualisations of the outcomes of a data mining process are called explanations as well. Given that such visualisations are designed for data scientists to understand the outputs of the decision algorithms, such works are an example of studies that do not satisfy our inclusion criteria.

Second, we are only interested in studies involving explanations that are related to a specific decision making instance. Generally, an explanation provided by the system can be any form of visual, textual, or multi-modal means to convey additional information to the decision maker *regarding the specific decision to be made*, e.g. about the system's reasons to recommend a certain alternative. However, in our review, we do *not* consider approaches in which the system merely provides background knowledge about the domain and this knowledge is independent of the current decision making problem instance. Approaches in which the system displays details about how to interact with user interfaces are also not in the scope of our work.

Table 2 Inclusion and exclusion criteria

Inclusion criteria	
IC-1	The paper proposes an explanation generation technique
IC-2	The paper presents a software application that includes an explanation facility
IC-3	The paper presents an evaluation of forms of explanations
IC-4	The paper presents a study that investigates foundations of explanations in advice-giving systems
Exclusion criteria	
EC-1	The paper is not written in English
EC-2	The content of the paper was also published in another, more complete, paper that is already included
EC-3	The content is not a scientific paper, but an introduction, glossary, etc.
EC-4	We have no access to the full paper
EC-5	There is a statement in the abstract or content of the paper that explanations are provided, but they are not detailed

Third, we consider only scenarios in which there is a limited number of *alternatives*, and the system task is to select one or more of the available choices. In case of multiple selected alternatives, in many cases a ranking is determined by the system. However, a number of papers on decision support systems exist in which the algorithmic task of the system is to compute the single mathematically optimal solution to a given problem. Work on such scenarios, in which there is no set of alternative choices, are also not included in our review. Similarly, studies of explanations regarding the outcomes of mathematical simulations are excluded as well.

With respect to exclusion criteria, we proceeded as follows. In order to obtain the papers in which the studies were published, we first tried to access them through the TU Dortmund network and the Portal de Periódicos CAPES[3]. If the full text was not available, we searched for the paper on the web using: (i) author websites; (ii) Googles search website; and (iii) repositories of scientific papers, such as Google Scholar and ResearchGate. At the end, only eight studies were excluded because we had no access to the full papers (EC-4).

Finally, there are a number of studies excluded by EC-5 that state that the proposed system has the *potential* to provide explanations, but do not concretely describe how it is done. Examples of studies excluded due to this reason are mainly those in the context of argumentation (Rahwan and Simari 2009) and studies that focus on transforming a particular output of a decision inference method to a data structure, which is assumed to be easier to explain. For instance, there are approaches that transform artificial neural networks into rules, without detailing how such rules are used to provide explanations to end users.

[3] http://www.periodicos.capes.gov.br/.

3 Systematic review execution

The next step in the systematic review process is to execute an appropriate search query against the literature databases. In this section, we provide details of how we constructed the search query and how we ended up with the set of *primary studies* that are considered relevant for our work. The results and insights are then discussed in the remaining sections.

3.1 Search string construction

String construction in systematic reviews is based on a set of terms of interest and their synonyms. Our search string covers two main terms, which both have to appear in the potentially relevant papers. The first term is *explanations*, which is the main topic of our study. In different communities, researchers use alternative terms to refer to explanations, and those were added as synonyms of the term *explanation*. The second term is *decision support system*, which is one of the classes of system that are targets of our review. As synonyms, we consider alternative classes of advice-giving systems that exist, including in particular knowledge-based and expert systems, as well as recommender systems. For some of them, we included subsets of typically used expressions to refer to such system classes, because they are used with different complementary terms, such as recommendation system, recommendation provider, etc. Therefore, in such cases, we included only the main words as synonyms, because they cover all of the alternatives for the term. The resulting set of synonyms used in our search string is shown in Table 3. The final search string is consequently as follows.

```
(explanation OR justification OR argumentation) AND (decision support system OR decision making OR expert system OR recommender OR recommendation OR knowledge-based system OR knowledge based system)
```

The search string was customised to the specific syntax of each of our target databases. In all but one of the cases, we were able to search for the terms in the abstracts of the papers. In the case of the Springer Link database, searching within abstracts was not possible due to API limitations. We thus searched for our terms in the keywords of the papers in this case.

Table 3 Terms and their synonyms

Term	Synonyms
Explanation	Justification, argumentation
Decision support system	Decision making, expert system, recommender, recommendation, knowledge-based system, knowledge based system

Table 4 Search results by source

Source	Number of studies
ACM Digital Library	278
IEEE Xplore Digital Library	469
ScienceDirect	385
Springer Link	92
Duplicates	15
Total (including duplicates)	1224
Total (excluding duplicates)	1209

Table 5 Selection of primary studies according to inclusion and exclusion criteria

Step/criterion	Technique	Tool	Evaluation	Foundational	Total
Abstract analysis	192	161	30	3	386
Full analysis	135	139	23	5	302
EC-01		1			1
EC-02	10	2	1		13
EC-03	2	5			7
EC-04	3	5			8
EC-05	20	37			57
Selected studies	101	89	22	5	217

One TECHNIQUE paper satisfies both EC-02 and EC-04

3.2 Selection of primary studies

After querying the four selected databases on August 12, 2016, we obtained 1209 papers (excluding duplicates) as result. The detailed statistics for each database are given in Table 4.

The relevant primary studies were then selected using a two-step procedure. In the first step, we analysed the title and abstract of each of the 1209 papers. If the title or abstract provided any sort of indication that the paper matched one of the inclusion criteria, the paper was selected to be subsequently analysed in detail. We also pre-classified each paper according to one of the inclusion criteria, i.e. we identified a preliminary study type.

In the second step, the full text of each paper was retrieved, if available, and analysed. Then, we checked each of the exclusion criteria and re-evaluated whether the paper truly satisfied one of the inclusion criteria. As result, some papers were discarded or associated with a different inclusion criterion.

Each abstract and paper was analysed by a single researcher, strictly following the specified protocol. When it was not fully clear whether a certain criterion was satisfied, i.e. border cases, the opinion of a second researcher was requested in order to minimise the potential researcher bias.

Following this procedure, we ended up with a total of *217 primary studies*. The detailed statistics about the paper selection process are presented in Table 5. The column headings correspond to the applicable inclusion criteria, and rows with labels starting with *EC* show how many papers satisfied each exclusion criterion. The full list of papers that was finally considered is shown in Table 6.[4]

4 Results

In this section, we summarise the insights that we obtained by analysing the 217 selected primary studies with respect to our research questions.[5] These insights are used later on to develop a comprehensive taxonomy of aspects to consider when designing future explanation facilities.

4.1 Historical developments

Before focusing on our research questions, we provide an overview of the historical developments in the research field of explanations by analysing the papers associated with the examined studies in terms of type and publication date. Figure 1a shows how many papers of each type were published *per decade* in absolute numbers. Figure 1b is based on the same data, but it reports percentages within the corresponding decade instead of absolute numbers. The following main observations can be made.

- Generally, the total number of published papers in the field increases. We emphasise that the last decade (2010–present) corresponds to only about 6.5 years.[6]
- Papers on TOOLS with explanation facilities were much more common in the past, perhaps because such papers were more often considered as research contributions at that time. In exchange, an increase over time can be observed in the number of papers related to the proposal of new TECHNIQUES.
- The empirical EVALUATION of explanations received much more attention in the recent past. This is an indication that the field achieved a higher level of maturity due to the progress made in terms of research methodology.
- Papers on FOUNDATIONAL aspects of explanations are very scarce.

From a historical perspective, it seems that research on the topic of explanations reached some plateau in the 1990s. In the 2000s, we see a stagnation, but a considerable increase again in the past few years. We attribute the observed stagnation in the 1990s to the declining role of knowledge-based systems at that time. In some areas of decision support systems, and particularly in the field of recommender systems, ML-based approaches became predominant in the 2000s and researchers focused more on determining the *right* recommendations than on the provision of explanations.

[4] Further details about which inclusion and exclusion criterion was fulfilled by each individual paper, along with the detailed results of our analysis, can be found online at: http://inf.ufrgs.br/prosoft/resources/2017/umuai-sr-explanations.

[5] The interested reader can find the detailed analysis of each of the studies on the following link: http://inf.ufrgs.br/prosoft/resources/sr-explanations.

[6] Note that more papers were published in computer science in general over time.

Table 6 Selected primary studies

Approach

Norton (1988), Burattini et al. (2002), Deep et al. (1988), Slagle (1988), Abu-Hakima and Oppacher (1988), Koussev et al. (1989), David and Krivine (1989), Maybury (1989), Saunders and Dobbs (1990), Riordan and Carden (1990), Lambert and Ringland (1990), Lee and Hsu (1992), Hair et al. (1992), Machado and da Rocha (1993), Guida et al. (1997), Liu et al. (1998), Bofeng et al. (2004), Hu et al. (2008), Amer-Yahia et al. (2008), O'Donovan et al. (2009), Yu et al. (2009), Guy et al. (2009, 2010), Vig et al. (2009), Horan and O'Sullivan (2009), Hussein and Neuhaus (2010), Marx et al. (2010), Zanker and Ninaus (2010), Mejia-Lavalle (2010),, Song et al. (2010), Grando et al. (2011), Reyes et al. (2011), Thirumuruganathan and Huber (2011), Gedikli et al. (2011), Fong et al. (2012), Bader et al. (2012), Cleger-Tamayo et al. (2012), Bostandjiev et al. (2012), Nunes et al. (2012b), Vashisth et al. (2012), Blanco et al. (2012), Hornung et al. (2013), Chen et al. (2013a), Widyantoro and Baizal (2014), Katarya et al. (2014), Zhang et al. (2014), Bedi et al. (2014), Chen and Wang (2014), Barbieri et al. (2014), Muhammad et al. (2015), Charissiadis and Karacapilidis (2015), Wang et al. (2016a), Hasling et al. (1984), Reggia et al. (1985), Basu and Dutta (1986), Hunt and Price (1988), Yasdi (1989), Strat and Lowrance (1989), Yen (1989), Wick and Slagle (1989b), Tanner and Keuneke (1991), Diederich (1992), Basu and Ahad (1992), Klein and Shortliffe (1994), Lopez-Suarez and Kamel (1994), Yoon et al. (1994), Tong and Ang (1995), Slotnick and Moore (1995), Mitra and Pal (1995), Kim and Park (1996), Benaroch (1996), Shaalan et al. (1998), Metzler and Martincic (1998), Pal (1999), Chandrasekaran and Mittal (1999), Bohanec et al. (2000), Richards (2000), Mao and Benbasat (2001), Papamichail and French (2003), Wall et al. (2003), Martincic (2003), Reilly et al. (2005), Mcsherry (2005), Lacave et al. (2006), Pu and Chen (2007), Sherchan et al. (2008), Symeonidis et al. (2008), Štrumbelj et al. (2009), Kagal and Pato (2010), Labreuche (2011), Bedi and Sharma (2012), Chen et al. (2013b), Garca et al. (2013), Briguez et al. (2014), Bavota et al. (2014), Du and Ruhe (2014), Hatzilygeroudis and Prentzas (2015), Hanshi et al. (2016), Belahcene et al. (2017), Oramas et al. (2016), Ji and Shen (2016)

Tool

Beiley and Duban (1990), Holman and Wolff (1988), Chelsom et al. (1988), Washington and Ali (1998), Terano et al. (1989), Popchev et al. (1989), Hudson and Cohen (1989), Perlin et al. (1990), Tong (1990), Sarkar et al. (1990), Cagnoni et al. (1991), Cheng et al. (1991), Wang et al. (1992), Srivastava (1992), Gallagher et al. (1995), Guida and Zanella (1995), de Braal et al. (1996), Schröder et al. (1996), Pazzani et al. (1997), Malheiro et al. (1999), Santoso et al. (1999), Ng and Ong (2000), Bohnenberger et al. (2005), Borlea et al. (2005), Strachan et al. (2005), Libório et al. (2005), Felfernig (2005), Chiou and Yu (2007), Hussain and Abidi (2007), Mahmoud et al. (2008), Narayanan and McGuinness (2008), Chang and Hsieh (2010), Roitman et al. (2010), Janjua and Hussain (2011), Kadhim et al. (2013), Balleda et al. (2014), Buschner et al. (2014), Mocanu (2015), Shoval (1985), Davis (1986), Karwowski et al. (1987), Overby (1987), Grierson and Cameron (1988), Basu et al. (1988), Wick and Slagle (1989a), Wong and Cheung (1989), Jabri (1989), Levy et al. (1989), Allgayer et al. (1989), Tjahjadi et al. (1990), Helms et al. (1990), Mller-Kolck (1990), Jamieson (1991), Ringer et al. (1991), Davey-Wilson (1991), Ray (1991), Gowri et al. (1991), Tzafestas and Konstantinidis (1992), Bau and Brezillon (1992), Aarle and Bercken (1992), Jung and Burns (1993), Chouicha and Siller (1994), Mitra (1994), Kim and Lee (1995), Buchanan et al. (1995), Zeleznikow et al. (1995), Artioli et al. (1996), Castro et al. (1996), Nuthall and Bishop-Hurley (1996), Matsatsinis et al. (1997), Horn et al. (1998), Ezquerra et al. (1999), Lieberman et al. (1999), Pal and Palmer (2000), Gvenir and Emeksiz (2000), Bielza et al. (2000), Joch and Dudeck (2001), Hodgkinson and Walker (2003), Zain et al. (2005), Goud et al. (2008), Matelli et al. (2009), Vogiatzis and Karkaletsis (2011), Bosnić et al. (2012), Mendes et al. (2013), Nart and Tasso (2014), Omran and Khorshid (2014b), Omran and Khorshid (2014a), Gómez-Vallejo et al. (2016), Blake et al. (2016)

Table 6 continued

Evaluation
Herlocker et al. (2000), Felfernig and Gula (2006), Ehrlich et al. (2011), Zanker (2012), Sharma and Cosley (2013), Gkika and Lekakos (2014), Schaffer et al. (2015), Bussone et al. (2015), Muhammad et al. (2016), Wang et al. (2016b), Murphy and Phillips (1991), Suermondt and Cooper (1993), Swinney (1995), Ye (1995), Ramberg (1996), Nakatsu and Benbasat (2003), Gönül et al. (2006), Li and Gregor (2011), Tan et al. (2012), Tintarev and Masthoff (2012), Gedikli et al. (2014), Wang et al. (2016c)
Foundation
Tintarev and Masthoff (2007a), Nunes et al. (2012a), Rook and Donnell (1993), Giboney et al. (2015), Gregor (2001)

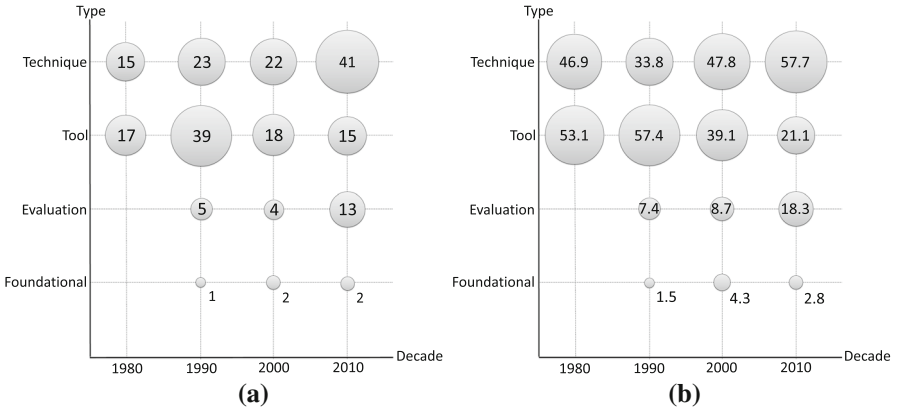

Fig. 1 Primary studies: number of studies by type per decade. **a** Number of primary studies. **b** Percentage of primary studies

A measurable increase can be seen in the number of published research papers on the topic in the past few years, indicating the importance and timeliness of the topic. Furthermore, we expect more works in the future caused by two recent trends in computer science. First, explanations of the behaviour of a software system that are directed to the end user are clearly a key ingredient for modern *human-centric computing* approaches (Jaimes et al. 2007). Second, as discussed, an increasing number of tasks are expected to be delegated to automated software systems that are based on modern ML technology in the future. However, to be accepted by end users, the suggestions made by advice-giving systems must be perceived to be fair and transparent in many application domain, and explanations are key to this.[7]

[7] See http://www.fatml.org for a recent workshop series on fair, accountable, and transparent machine learning approaches.

4.2 RQ-1: What are the characteristics of explanations provided to users, in terms of content and presentation?

To answer our first research question, we used the analysis method described next. Primary studies analysed in the context of this research question are those that proposed forms of explanations, consisting of TECHNIQUES and TOOLS. Together, such types of studies are referred to as (explanation generation) approaches. We discuss the obtained results in terms of the content and presentation of explanations.

4.2.1 Analysis method

We followed principles from *grounded theory* (Glaser 1992) to investigate this question in a systematic and unbiased way. Following these principles, we iterated through all primary studies and labelled their proposed explanations with *codes* (in grounded theory terminology) that captured key ideas associated with the explanation content. For example, consider an explanation approach that has the following system output: "The recommended alternative has A and B as positives aspects, even though it has C as a negative aspect." Explanations of this type, in which the features of different alternatives are contrasted, were labelled with the code *Pros and Cons*.

The inspection of all studies led us to a first set of codes. This preliminary set was then analysed, in order to merge codes that represent the same underlying idea. For instance, there is an explanation that organises the suggested alternatives in groups to highlight the trade-off relationships between certain features (Pu and Chen 2007). This and similar approaches were initially labelled with the code *Trade-off*. As both *Pros and Cons* and *Trade-off* represent the same underlying idea of explaining the alternative options, these codes were merged.

At the end of the process, each form of explanation proposed in the considered studies was labelled with one or more codes related to *what* kind of information is presented. After merging the codes, we ended up with 26 labels. A similar method was adopted regarding *how* the information is presented.

4.2.2 Explanation content

From the 26 codes, 17 refer to the type of information that was displayed, while the 9 remaining codes are about general observations regarding content, such as cases where explanations are context-tailored. The content-related codes are described in detail in Table 7, while their occurrence frequencies are shown in Fig. 2. The different types of information presented in the explanations can be organised in four main groups, as indicated in Table 7 and detailed as follows.

User preferences and inputs A possible way to explain a system's suggestion is to provide users with explanatory information that is related to the provided inputs. The explanation can, for example, indicate (i) which of the user preferences and constraints were fulfilled and which were not, (ii) to what extent the system believes that the recommended alternative is appropriate given the stated preferences, or (iii) which inputs were the most decisive when determining the suggestion.

Table 7 Codes related to the type of information conveyed in explanations

Label	Description
User preferences and input	
Decisive input values	Indication of the inputs that determined the resulting advice
Preference match	Provision of information about which of the user preferences and constraints are fulfilled by the suggested alternative
Feature importance analysis	Justification of the advice in terms of the relative importance of features, e.g. by showing that changing feature weights would cause the selected alternative to be different
Suitability estimate	Indication of how the system believes that the user would evaluate the suggested alternative, e.g. by showing a predicted rating
Decision inference process	
Inference trace	Provision of details of the reasoning steps that led to the suggested alternative, e.g. a chain of triggered inference rules
Inference and domain knowledge	Provision of information about the decision domain or process, e.g. about the main logic of the inference algorithm. For example: "We suggest this option because similar users liked it."
Decision method side-outcomes	Provision of algorithm-specific outcomes of the internal inference process, e.g. a calculated number that expresses the system's confidence
Self-reflective statistics	Provision of facts regarding the system's performance, e.g. by informing the user how many times the system made decision suggestions in the past that were accepted
Background and complementary information	
Knowledge about peers	Provision of information about the preferences of related users, e.g. ratings given to a suggested alternative by social friends
Knowledge about similar alternatives	Indication of similar alternatives that were an appropriate (system's or user's) decision in a similar context in the past, e.g. items that the user or related peers showed interest in
Relationship between knowledge objects	Provision of information about the relationship between features, or features and users. This can be done, for example, in the form of a directed acyclic graph representing a causal network
Background data	Provision of (external) background data specific to the current problem instance, e.g. data derived from processing posts in a social network, which were considered in the decision inference process
Knowledge about the community	Provision of information that supports the decision based on the behaviour and preferences of a community, e.g. showing the general popularity of the proposed alternative

Table 7 continued

Label	Description
Alternatives and their features	
Decisive features	Indication of the features of the alternative that are key to the decision
Pros and Cons	Indication of the key positive and negative features of the alternative
Feature-based domination	Justification of a decision in terms of the dominance relationship between two alternatives, e.g. by showing that an alternative is not selected because it is dominated by another
Irrelevant features	Indication of features that are irrelevant for the decision, typically when the values of such features in the suggested alternative are not considered good

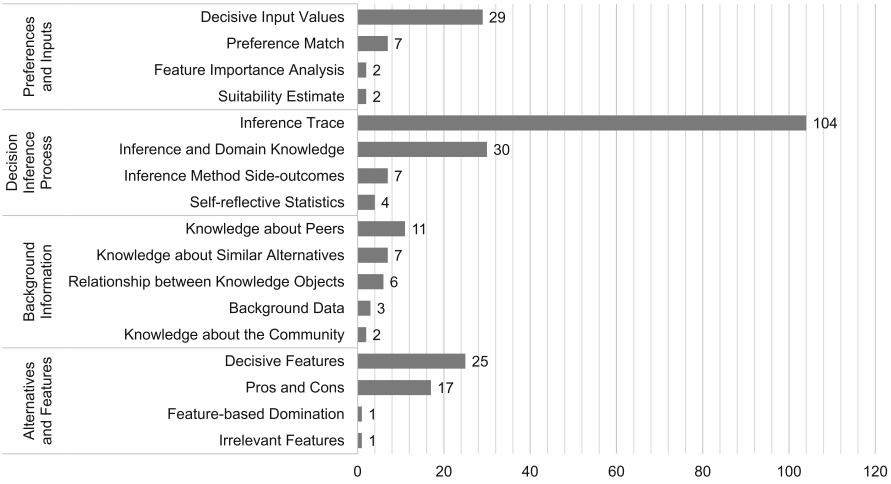

Fig. 2 Occurrence of codes related to explanation content

Decision inference process Providing information about the inference process of a specific decision problem (e.g. in the form of traces) was the most common approach in classical expert systems. Some explanations only provide the general logic of the system's internal inference process; others mention system confidence in the suggestion or the success rate in past decision making situations.

Background and complementary information A reduced amount of explanations provide additional background information that is specific to the given decision making instance. Various types of background and complementary information were identified. Explanations can, for example, provide more information about the knowledge sources that were used in the inference process or how relevant entities in the knowledge base are interrelated. They can also refer to past suggestions or user choices in similar situations, or mention which users liked the suggested alternative.

Alternatives and their features A common approach in the literature is to explain the system's suggestion by analysing the features of the alternatives. Some explanations consist of lists of features, pro and con, for each alternative; others refer to dominance relationships based on the features, but most explanations show which features were decisive in the inference process.

We made the following additional observations regarding orthogonal aspects when analysing which kind of information is conveyed to the user within the explanations.

Baselines and multiple alternatives First, in most cases the provided explanations focus solely on one single (recommended) alternative. For example, such explanations contain details about the features of the alternative, describe in which ways it is suitable for the user, or how the decision was made. However, there are some approaches that use other alternatives as a *baseline* for comparison. We observed two forms of including baselines in such a comparison. One option is to use one single alternative (e.g. the second best choice), as done in Papamichail and French (2003) and Labreuche (2011). As an example, the provided explanation can detail in which ways the best alternative is favourable over the second best option. A different approach is to contrast one alternative with a *set* of other options (Bohnenberger et al. 2005; Mejia-Lavalle 2010; Helms et al. 1990). In this case, the explanation highlights why the best alternative is generally better than the group of other alternatives, e.g. in terms of specific features. Finally, there are two cases (Lopez-Suarez and Kamel 1994; Pu and Chen 2007) in which the explanation does not refer to one single best alternative as a reference point to compare other options with, but to a *group* of alternatives (equally suitable regarding a considered aspect) or groups of rules that were used in the inference process.

In one of these studies (Pu and Chen 2007), a key goal is to educate users about the trade-off among options by means of explanation interfaces. This can be achieved by these explicitly provided explanations and also by providing interactive decision making support, such as by means of dynamic critiquing (McCarthy et al. 2005).

Context-tailored explanations Which information an explanation should provide to the user can depend on various factors, including the expertise or interests of users, or their current situational context. We identified 16 primary studies in which explanations are tailored to the current situation in different forms, e.g. by using different levels of detail. Moreover, *group decisions* can be seen as a very specific context. In our review, we found only one single approach (Lieberman et al. 1999) that focuses on explaining decision suggestions that were made for a group of users.

External sources of explanation content As discussed, some explanations provide information associated with background knowledge, but such knowledge is almost always associated with the decision inference process. Four approaches (Zhang et al. 2014; Chen and Wang 2014; Muhammad et al. 2015; Charissiadis and Karacapilidis 2015) in the e-commerce domain exploited external sources of information, namely *product reviews*.

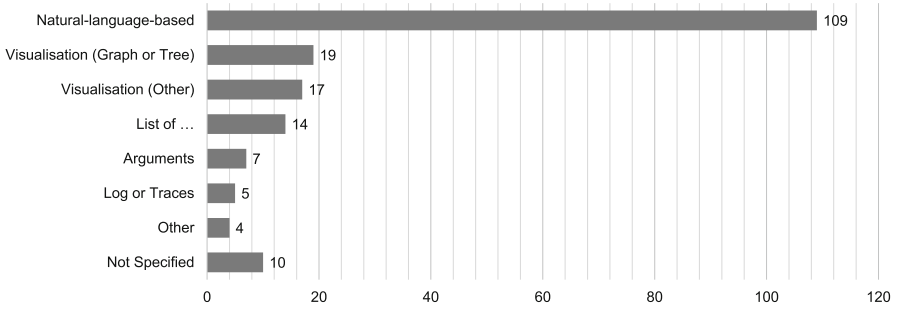

Fig. 3 Occurrence of codes related to explanation presentation

Interactive explanations In some approaches, the explanations provided by the system represent a starting point for further user interactions, e.g. asking the user for additional input. The common types of questions associated with user interaction are: (i) *what-if* (what the output would be if alternative input data were provided); (ii) *why* (why the system is asking for a particular input); and (iii) *why-not* (why the system has not provided a given output). These questions were addressed in 10, 37, and 13 studies, respectively. These three types of questions, together with the *how*-question (how the system reached a certain conclusion), are those typically addressed in expert systems. Explanations that address the *how*-question appear in Table 7, mainly as inference traces. In addition, some explanations given as answers to this question trace the path from a decision to the given user input. As result, they only report the input that actually led to the decision.

4.2.3 Explanation presentation

Now that we have discussed the content of explanations, we proceed to how they are *presented* to users. Codes were also used in this analysis. As result of the categorisation of the different ways of how explanations are presented, we identified 8 presentation codes, which are shown along with their occurrence frequencies in Fig. 3.

The most frequent code is by far *natural-language-based*, i.e. most of the approaches use a text format to display explanations to the user. Note that we also used this code to label explanations that are based on pre-defined templates, which were, for example, instantiated with lists of features before they were presented to the user. Alternatively, displaying simple lists of various things (e.g. features, users, alternatives, past cases, conclusions derived in the decision process) was chosen as a presentation form by a number of studies. Different forms of visualisations, e.g. in the form of graphs, are also quite common means to convey the explanatory information to the user. Only a smaller number of studies presented inference traces (or other forms of logs) to the user as a final explanation presentation format. Finally, some studies structure the explanation as an argument, typically in the form of the Toulmin's argument structure (Toulmin 2003).

There are four studies that used unusual forms of presentations, and we assigned them to the group called *Other*. These are the types of outputs in this group: (i)

audio (Terano et al. 1989), which uses voice as output; (ii) highlighting (Roitman et al. 2010), which presents an alternative with highlighted aspects; (iii) query results (Basu and Ahad 1992), when the explanation is the result of a database query; and (iv) OWL (Ontology Web Language) (Sherchan et al. 2008), when explanations are given using this technical language (to be further processed for presentation to the end user). Studies that do not specify how the explanations are presented to the user were assigned to the group *Not Specified*.

4.3 RQ-2: How are explanations generated?

Having discussed what is actually presented as explanations to end users, we now proceed to the investigation of *how* they are generated in the proposed approaches. Specifically, we are interested in the inner workings of processes that take the outcomes of a decision inference method (that was used to determine the suggested alternative) to produce what is finally shown as an explanation.

4.3.1 Explanation generation process

We observed that most of the studies investigated in our review do not provide many details about this process. The reason is that in most cases the explanation generation process is closely tied to the underlying decision inference method and the data that is used to determine the suggested alternative. If, for example, the underlying inference method is rule-based, the explanation presented to the user might consist of a set natural language representations of the rules that were triggered. In many of these cases, no further information is provided regarding whether any additional processing was needed to generate what is finally presented to the user in one of the different forms shown in the previous section.

Only a few studies implemented more complicated explanation generation processes. In particular, when the underlying inference method is based on multi-criteria decision theory (MAUT) (Keeney and Raiffa 1976), to derive an explanation for the user, specific algorithms are often provided to analyse user preferences with respect to the features of the alternatives (e.g. Charissiadis and Karacapilidis 2015; Klein and Shortliffe 1994; Labreuche 2011). Approaches based on artificial neural networks (e.g. Ray 1991; Mitra 1994; Mitra and Pal 1995) also often specify algorithms to extract rules from the network to generate the explanations.[8]

Although there is limited information regarding the explanation generation process in most of the approaches, we identified three factors that strongly influence the selection of what sort of explanation will be provided. In a few studies, the *application domain* plays a key role in this process. 18 (out of 101, or 17.8%) explanation generation TECHNIQUES are domain-specific, that is, their proposals are focused on, and possibly tailored to, a particular domain. Domain-specific approaches can be found

[8] We remind the reader that approaches that simply provide a rule extraction algorithm without detailing how the explanations are provided to the end user are excluded from our review as specified in our inclusion and exclusion criteria.

in the recent past mostly in the fields of *Computing & Robotics* and *Media Recommendation*. They exploit information types that are only available in a certain domain to produce richer explanations. The approach proposed by Briguez et al. (2014) is an example of such a work, in which the authors identified specific argument structures that can be used in the media (movie) recommendation domain. However, from TECHNIQUES that focus on a particular domain, many do not exploit any domain specificities. Domain specificity is not analysed in studies involving TOOLS, because they are always developed with a focus on a particular domain.

In addition to the application domain, we identified two other key drivers of the explanation generation process. The first is the *purpose* for which explanations are provided in an advice-giving system and, second, the adopted *underlying decision inference method*, as briefly discussed above. These are further discussed in the next sections.

4.3.2 Purpose of explanations

Previous surveys highlighted the importance of considering the intended purpose of explanations when designing an explanation facility (Tintarev and Masthoff 2007b; Buchnan et al. 1984). The intended purpose, which can be, for example, to persuade users to accept a suggested alternative, should determine what information should be conveyed to the user. However, most of the studies analysed in our review do not explicitly state their purpose or, more specifically, their *intended* purpose (Friedrich and Zanker 2011).

Therefore, we adopted an analysis method, in which we searched in TECHNIQUE and TOOL studies for sentences that indicate the underlying explanation purpose of the study. To categorise the different studies, we used the list of possible purposes proposed by Tintarev and Masthoff (2007b) as a basis. We then extended this list with additional categories that we found during the analysis and specifically included categories that were identified in an earlier work on explanations by Buchnan et al. (1984).

The resulting list of explanation purposes found in the studies is shown in Table 8. Some of the mentioned explanation purposes from the two reference works are in fact similar, but the authors used different labels, e.g. transparency *vs.* understanding or trust *vs.* acceptance. In these cases, we used the label that is used in the more recent reference work.

Figure 4 shows how often each intended purpose is in the focus of the investigation in our primary studies. Each study can be associated with more than one purpose. The majority of the studies (124, or 65%) are, however, concerned only with one single purpose. From the studies associated with multiple purposes, only one study (Hunt and Price 1988) explicitly states five of them (maximum). In one case (Buchanan et al. 1995), the purpose is a parameter of the explanation generation process (and classified as *parameterised* in Fig. 4). In 11% of the cases (21 studies), we could not identify a single sentence within the respective paper that indicates the purpose of the explanations.

As can be seen in Fig. 4, in which we also indicate the decade in which the respective studies were published, the most common explanation purpose is to provide *transparency*, i.e. to explain how the system ended up with the suggested alternative(s).

Table 8 Explanation purposes identified in primary studies (based on Tintarev and Masthoff 2007b; Buchnan et al. 1984)

Purpose	Source	Description
Transparency	Tintarev and Masthoff (2007b), Buchnan et al. (1984)	Explain how the system works
Effectiveness	Tintarev and Masthoff (2007b)	Help users make good decisions
Trust	Tintarev and Masthoff (2007b), Buchnan et al. (1984)	Increase users' confidence in the system
Persuasiveness	Tintarev and Masthoff (2007b), Buchnan et al. (1984)	Convince users to try or buy
Satisfaction	Tintarev and Masthoff (2007b)	Increase the ease of use or enjoyment
Education	Buchnan et al. (1984)	Allow users to learn something from the system
Scrutability	Tintarev and Masthoff (2007b)	Allow users to tell the system it is wrong
Efficiency	Tintarev and Masthoff (2007b)	Help users make decisions faster
Debugging	Buchnan et al. (1984)	Allows users to identify that there are defects in the system

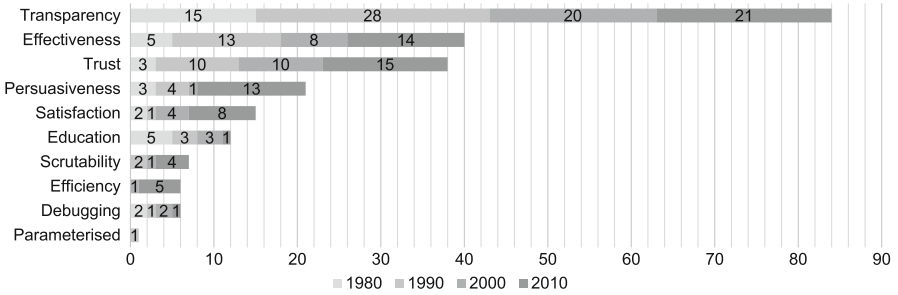

Fig. 4 Purpose analysis: number of TECHNIQUE or TOOL studies per purpose

The provided explanations in these studies focused on exposing the system's inference process in order to make the recommended decision understandable. The authors of one study (Norton 1988) argued that this aspect is particularly critical because "*[the] human user bears the ultimate responsibility for action*" and, therefore, she should be able to explain the decision. Transparency is also seen as key for users to develop *trust* toward the system (Gedikli et al. 2011; Yasdi 1989; Mocanu 2015). In some studies that focus on transparency, trust is thus not the direct purpose of the system's explanation facility, but an expected indirect effect of transparency. In a number of other works, however, trust-building is explicitly mentioned as the goal of the explanations.

The second most frequent purpose of explanations in the analysed primary studies is *effectiveness*, i.e. to help users assess if the recommended alternative is truly adequate for them.[9] *Persuasiveness*, i.e. a system's capability to nudge the user to a certain

[9] One of the examined studies (Roitman et al. 2010) focused on *safety* in the context of decisions with critical consequences. We included it in the category *effectiveness*.

Table 9 Decision inference methods: method type by decade

Category	Subcategory	1980	1990	2000	2010	Total	%
Knowledge-based		33	50	28	31	142	68.3
	Rule-based	28	33	11	7	79	55.6
	Logic-based	2	3	3	8	16	11.3
	Multi-criteria decision making	0	1	3	5	9	6.3
	Constraint-based	0	1	1	0	2	1.4
	Case-based reasoning	0	1	3	2	6	2.9
	Other	3	11	7	9	30	21.1
Machine learning		2	13	12	24	51	24.5
	Feature-based	2	13	6	11	32	62.7
	Collaborative-filtering	0	0	5	5	10	19.6
	Hybrid	0	0	1	8	9	17.6
Mathematical model		0	2	0	0	2	1.0
Human-made decision		0	1	0	1	2	1.0
Algorithm-independent		0	3	3	5	11	5.3

direction, which can be conflicting with effectiveness (Chen and Wang 2014), was also in the focus of a number of studies.

Looking at the historical developments, we can observe in Fig. 4 that other potential purposes, like user satisfaction, scrutability, and efficiency, received more attention in the recent past. Reducing the user's cognitive load and trying to increase their satisfaction with the system (thereby increasing their intention to return) are essential aspects for e-commerce applications, which have been increasingly investigated during the past few years. In this commercial context, the potential persuasive nature of explanations (Bilgic et al. 2005) also attracted more research interest recent years.

4.3.3 Underlying decision inference methods

Given that in most cases the explanation generation process is highly coupled with the underlying inference method, we analysed which methods are used in the investigated primary studies to infer the suggested alternative. Table 9 shows the outcomes of this analysis, in which one study may fall into more than one category, if it uses different methods.

Given the historical importance of explanations in the context of (rule-based) expert systems, it is not surprising that the majority of the examined studies adopted a knowledge-based approach for decision inference and, correspondingly, for generating the explanations. Again, we can see the declining role of rule-based systems over the years and an increasing adoption of approaches based on ML. We use a two-level categorisation scheme to be able to detect the developments over time in a fine-grained manner. Based on such a categorisation, we can see, for example, that

explanations for collaborative-filtering recommender systems are only investigated after the year 2000.

The subcategory *Other* covers various alternative forms of knowledge-based approaches, including those that use special heuristics or ontologies, as well as studies that state that they use a form of knowledge-based reasoning without providing further details. Besides two studies that use mathematical models to infer the suggested alternative, there are two approaches (Nunes et al. 2012b; Buchanan et al. 1995) in which a decision is actually provided by the user, and the system seeks to complement this decision with an explanation. Their goals are: (i) to record the decision rationale for future inspections; or (ii) to provide an explanation on behalf of decision makers, who have expertise on the domain, to save their time.

Finally, Table 9 shows that only 11 approaches discuss explanation generation approaches that are independent of the underlying inference method. In some sense, this is not surprising, because generating an explanation using solely the user inputs or context together with the selected alternative is not trivial. However, this deserves further investigation, because extracting explanations from today's decision inference methods is becoming increasingly complex due to the increasing complexity of widely adopted ML algorithms, which may be even confidential, as argued by both Zanker and Ninaus (2010) and Vig et al. (2009). These authors decouple explanations from the underlying inference method by proposing the so called *knowledgeable explanations* and *tag-based explanations*, respectively. This can lead to explanations that may be disconnected from the reasons of why a decision inference method suggested a particular alternative. Although this alleviates the problems discussed above, it compromises system transparency. Another direction is the use of ML algorithms that allow to produce *explainable recommendations*, proposed by Zhang et al. (2014), which use latent factor models.

4.4 RQ-3: How are explanations evaluated?

We now focus on *how* explanation generation approaches were *evaluated* and, in some cases, compared. We first discuss whether the respective proposed forms of explanations were evaluated in the paper in which they were published. We then detail the types of evaluations adopted as well as the domains chosen to perform them. Finally, as the most common way of evaluating explanations is by means of user studies, we report their characteristics, such as their independent and dependent variables as well as sample size.

4.4.1 Presence of an evaluation

Considering TECHNIQUE and TOOL studies, we investigated to what extent explanations were evaluated in the paper that they were proposed. In this analysis, we considered any form of evaluation except cases in which the authors simply described a simple scenario (i.e. a *toy example*) to illustrate the use of the proposed approach and the explanation produced, even if this toy example was referred to as case study by the authors.

Table 10 TECHNIQUE and TOOL studies with evaluation (number and percentage)

	1980		1990		2000		2010		Total	
	#	%	#	%	#	%	#	%	#	%
TECHNIQUE	0	0.0	5	21.7	11	50.0	19	46.3	35	34.7
TOOL	0	0.0	3	7.7	1	5.6	1	6.7	5	5.6
Total	0	0.0	8	12.9	12	30.0	20	35.7	40	21.1

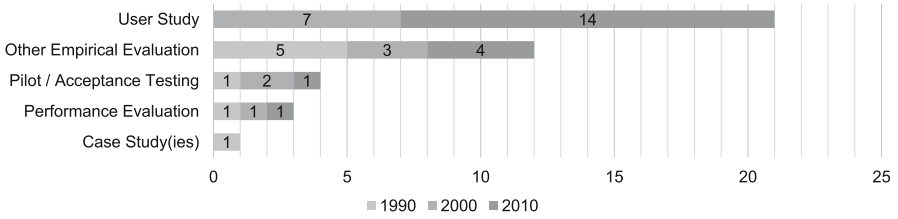

Fig. 5 Evaluation types

The results are shown in Table 10, in which we can observe that less than a quarter (21.5%) of the studies involving TECHNIQUES and TOOLS contains any form of evaluation, apart from toy examples. This can be partially explained by the fact that, when many expert systems papers were published (in the 1980s and 1990s), the methodological requirements in this field were probably lower in terms of evaluations than they are today.

Nevertheless, it is surprising that even nowadays (from 2010-present), the presence of an evaluation accompanying the proposal of a new form of explanation is still not typical, with almost two thirds of all analysed studies lacking a proper evaluation. In some studies, this can be explained by the fact that the main focus of the work was on another contribution, e.g. a recommendation algorithm or an expert system that included an explanation facility, and the evaluation is then limited to this main contribution.

4.4.2 Evaluation types and domains

Next, we analyse which *type* of evaluation researchers applied to assess or compare different explanations provided by a system. Figure 5 shows the five main types of evaluation that we found in the primary studies in which new forms of explanations are proposed. As can be seen, all identified evaluation types are empirical. We do not include the 22 EVALUATION studies in Fig. 5, because all of them describe results of user studies.

User studies are the predominant research method used to evaluate proposed TECHNIQUES and TOOLS, occurring in more than half (52.4%) of the cases. This is expected because there is no formal definition of a *correct* or *best* explanation in many application scenarios. In these cases, the only way to evaluate the provided explanations

is to capture the subjective perception of the users or to monitor the impact of the explanations in the user behaviour.

Possibly due to the time required to conduct user studies, alternative evaluation types, which do not require the availability of participants, were the choice in the remaining studies. Most of these, 12 in total, included a customised form of empirical evaluation, which involved generating explanations based on the proposed approach, and collecting measurements that are possibly specific to the explanation problem. An example of such a measurement is *explanation coverage* (Symeonidis et al. 2008), which is defined as the fraction of features that are part of the user preferences that are used in the explanation. Three studies focused specifically on the performance (computational efficiency) of their approach. Finally, the remaining evaluations either comprised a pilot study, or alternatively an acceptance test, or a set of case studies (reported in a single study Lopez-Suarez and Kamel 1994).

Having looked at the evaluation type, we now analyse which domains researchers chose to perform their evaluations. Figure 6a summarises the results of the analysis of all studies that include an evaluation. The results show that the most common application domain is *Media Recommendation*, followed by *Health*. In addition, a number of evaluations was done in the context of movie or music recommendation (which are included in the *Media Recommendation* category) in the past few years. The same applies to the domain of *(e-)Commerce*. In four studies, although there is an evaluation, the selected application domain is not reported (*No Domain* in Fig. 6a).

To understand how the interest in particular application domains changed over time, we present in Fig. 6b the domains associated with studies that describe a TOOL. Such TOOLS were more frequently proposed in the past decades, as opposed to evaluation studies, which received more attention in recent years. Considering TOOLS, the *Health* domain has been consistently in the focus of researchers over time. However, although it is the second most targeted domain in evaluations, the comparison between Fig. 6a and b shows that the *Health* domain has been largely more explored in the context of TOOLS than in evaluations. The popularity of this domain in developed TOOLS is mainly due to the many expert systems of the 1980s and 1990s that were designed to support medical decision making. Furthermore, we can observe that many domains were not explored in the past few years, but this is also a consequence of the generally declining number of studies that describe TOOLS.

4.4.3 User studies

We now investigate the user studies in more depth. We first discuss their design details in terms of the *independent and dependent variables*, followed by an analysis of their *sample size*, i.e. the number of involved participants.

Independent variables Figure 7 shows the independent variables of the study designs. In a few (four) cases, only one single treatment was used (marked as ST). In ten studies, explanations were presented in one condition but not in the other (marked as WN, standing for *With explanations* and *No explanations*). The majority of the studies, however, compared different kinds of explanations or the impact of their presence when providing a recommendation. In some of these studies, one of the alternatives

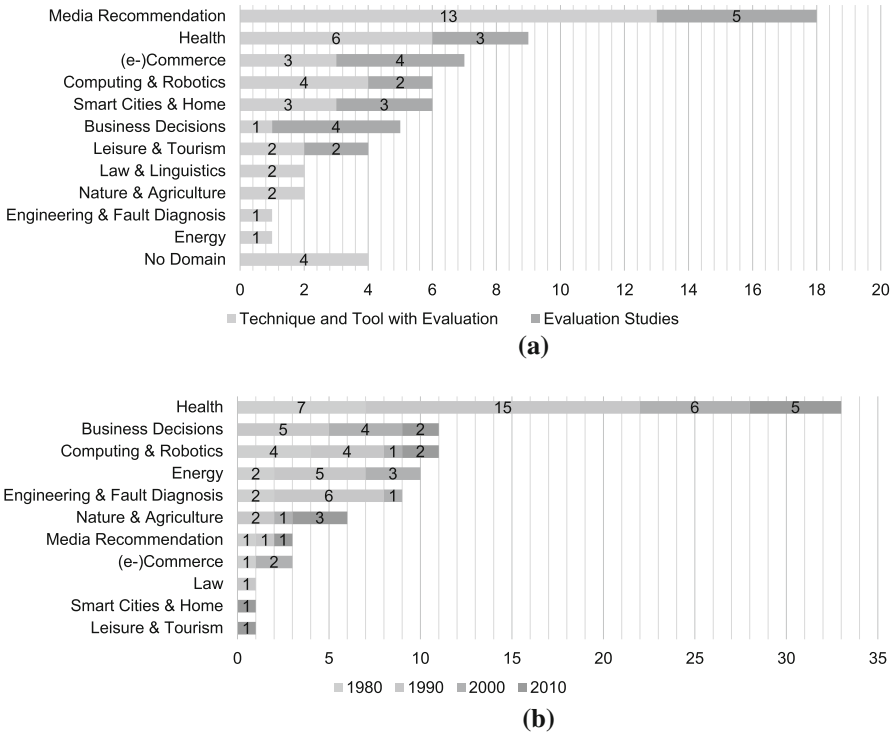

Fig. 6 Domain analysis. **a** Evaluations per domain. **b** Number of TOOL studies per domain

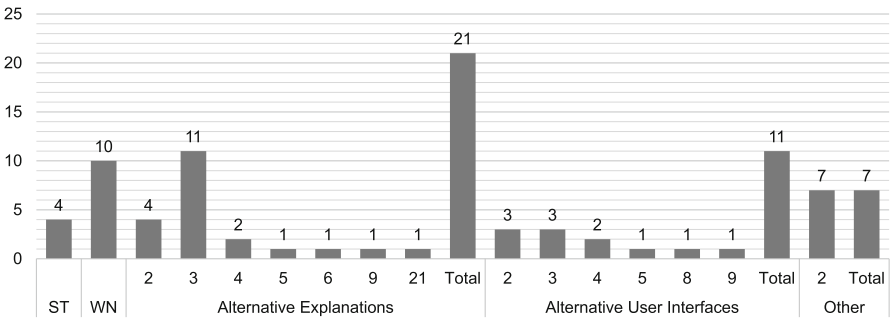

Fig. 7 Types of independent variables used in user studies

was providing participants with no explanation. The label at the bottom of the bars in Fig. 7 indicates how many alternatives were compared, while the label at the top provides the number of studies that had this number of alternatives.

We distinguished cases in which the study focused solely on different types of explanations (i.e. the only varying component of the information presented to the user is the explanation, while other components remain fixed) and cases in which the focus was on aspects of the user interface (i.e. presence or absence of alternative interface components, which influence the provided explanation). The former cases are referred

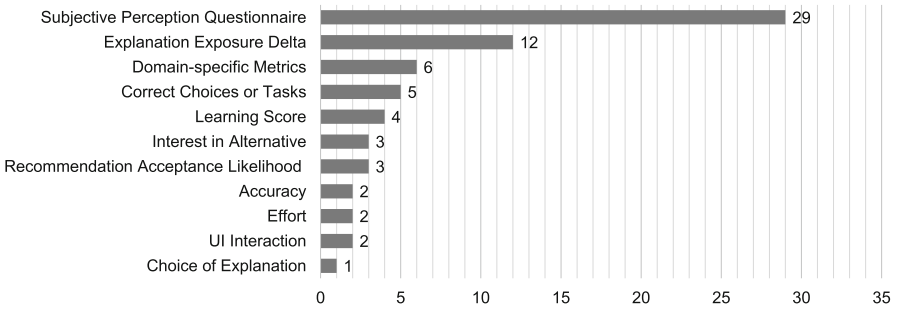

Fig. 8 Types of dependent variables used in user studies

to as *alternative explanations* in Fig. 7, while the latter are considered *alternative user interfaces*.

A smaller amount of studies (*Other*, in Fig. 7) investigated other relevant explanation-related aspects, such as the impact of the user's expertise level and different explanation properties. Examples of such properties are: (i) *source*, which indicates from where the information presented in the explanation is extracted; (ii) *length*, measured by the number of characters or number of features used in an explanation; (iii) *direction*, which can either be positive and aimed at justifying why a certain alternative fits a certain user, or negative, when it justifies the relaxation of certain input conditions; and (iv) *confidence*, which is associated with the vocabulary used in explanations, indicating how confident the system is with respect to the suggestion made to the user.

Dependent variables The different measurements that were collected in the user studies are summarised in Fig. 8. Note that several measurements were made in the majority of the studies. Various studies included the collection of the opinion of study participants on certain aspects. The second most adopted measurement, which we refer to as *explanation exposure delta*, was first adopted by Bilgic et al. (2005)[10] and uses a specific protocol to evaluate explanations. These types of dependent variables are described in Table 11, together with the many other variable types shown in Fig. 8.

Looking at the subjective perception questionnaires in more detail, we observed that participants were asked a wide variety of questions in the studies in order to investigate different aspects of explanations. We selected terms commonly used to refer to these aspects, given that there is no standardised terminology to classify the analysed aspects. Each question was classified using our selected terms. An example is asking the participant to indicate the agreement with the sentence: "The user interface is easy to use," which is associated with *usability*. Since such a classification approach leaves room for interpretation, we do not report specific occurrence numbers here, but represent the information in the form of a tag cloud (Fig. 9). The tag cloud shows that *transparency* is one of the main aspects that are evaluated in user studies. This observation matches the results from Sect. 4.3.2 (Fig. 4), in which transparency is mentioned

[10] Bilgic and Mooney's work is not included in this review because it was published in a workshop and not part of the databases searched in our review. We discuss the choice of databases and possible research limitations later in the paper.

Table 11 Types of dependent variables in user studies

Measurement type	Description
Subjective perception questionnaire	Participants are asked a set of questions in order to obtain participants' subjective view with respect to different explanation aspects. Responses are typically collected using a Likert-scale
Explanation exposure delta	Measures the difference between a score given by participants before and after the presentation of an explanation
Domain-specific metrics	Measurement of metrics that are meaningful only in particular domains, e.g. percentage of forecasts that were adjusted (forecasting domain)
Correct choice or tasks	Evaluates how many correct choices (or accomplished tasks) participants are able to make, when there is a notion of choice correctness assumed in the study (e.g. a disease diagnosed based on symptoms)
Learning score	Measures how much participants learn by interacting with the system
Interest in alternative	Measures to what extent participants are interested in the recommendation, e.g. by rating the recommended alternative
Recommendation acceptance likelihood	Measures whether participants agree with a recommendation or predicted suitability score
Accuracy	Measures the difference between the predicted suitability of suggested alternatives and how participants evaluate them considering provided explanations
Effort	Measures how much effort (in time) participants spend making a decision or evaluating an explanation
UI interaction	Measures how participants interact with the user interface, e.g. number of clicks or requests to more detailed explanations
Choice of explanation	Asks participants which from a set of alternative explanations they prefer

Fig. 9 Tag cloud of question topics in subjective perception questionnaires

as the most frequently stated intended purpose of providing explanations. A similar observation can be made for *trust*. Interestingly, *satisfaction* was often in the focus of the questionnaires, even though it is not commonly listed as an investigated explanation purpose. Note that satisfaction, according to Tintarev and Masthoff (2007b), refers to *usefulness* and *usability*, but they are often explicit targets of specific questions.

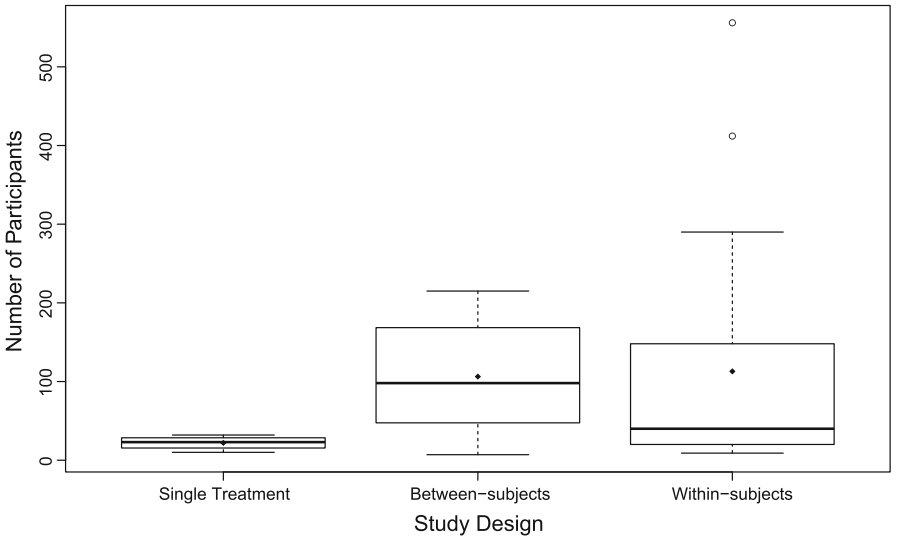

Fig. 10 Sample size in user studies

Table 12 Number of participants in user studies

Study design	Number of studies	Mean	SD	Median	Min	Max
Single treatment	4	22.0	9.2	23	10	32
Between-subjects	20	106.3	68.4	98	7	215
Within-subjects	21	112.9	147.8	40	9	556
All studies	45	101.9	112.2	54	7	556

Sample size We now consider the number of subjects that were involved in the user studies. We differentiate between three study designs: (i) single treatment, in which no comparison is made; (ii) between-subjects, in which subjects are split into groups and receive different treatments; and (iii) within-subjects, in which all subjects experience all of the different treatments. Typically, between-subjects studies should have a higher number of participants to achieve statistical significance. We present a box plot of the number of participants according to the different study designs in Fig. 10 and provide the exact numbers and further details in Table 12. Two between-subjects studies were not considered in this analysis because they do not report the number of participants. Moreover, when there is more than one user study reported in the same work, we considered each user study individually.

Even though the single largest study had a within-subjects design, between-subjects studies indeed had the largest median of participants. Moreover, not only the study design should influence the number of required participants, but also the number of factors that are investigated. However, we did not observe a relationship between these aspects. For example, the study performed by Guy et al. (2010) with 412 participants is a within-subjects study and has five treatments (different versions of recommender

systems). In contrast, Felfernig and Gula (2006) performed a study with 116 participants followed a between-subjects design and evaluated eight different versions. Finally, we observe that the number of participants of single treatment studies is typically low. A possible reason is that empirical studies are expected in an increasing number of subfields of computer science, and researchers resort to comparably simple studies in order to provide some form of evaluation. In any case, the low number of participants involved in these studies is a potential threat to their external validity.

4.5 RQ-4: What are the conclusions of evaluation or foundational studies of explanations?

After analysing how researchers evaluate different forms of explanation, we now focus on the conclusions they reached. This analysis also includes the main findings of the few FOUNDATIONAL studies that we identified.

4.5.1 Analysis method

A number of primary studies associated with TECHNIQUES and TOOLS only provided anecdotal evidence that the proposed approach is generally feasible or produces reasonable explanations. Additionally, in some cases, data was gathered to describe general characteristics of the generated explanations, such as their length. Such forms of unstructured evaluation, unfortunately, only provide limited evidence of the true benefits of the proposed approaches. We next thus only analyse conclusions that are based on user studies, including those published in EVALUATION papers.

As there is no standard design for user studies that evaluate forms of explanations, combining the results reported in the literature is not trivial. Again, to avoid researcher bias, we devised the following systematic way to analyse and contrast the obtained results. First, we extracted explicitly stated *key observations* and *conclusions* that were made by the authors of the studies. In this step, we did not infer any additional conclusions based on the provided data and also did not question the validity of the conclusions stated by the authors. Second, we classified the conclusions according to the explanation characteristics that were investigated in the studies. Often, these characteristics match possible explanation purposes and, in case they were not explicitly stated, we inferred this information from the measurements.[11] Third, we organised the conclusions as tuples of the form ⟨`direction target-explanation-style [compared-explanation-styles]`⟩. The direction can be positive, negative, or neutral, indicating that the target (proposed) explanation style increases, decreases, or has no impact in a certain measurement when compared with other explanation styles. We use the term *explanation style* to refer to the specific forms of explanation considered in the studies. When no information about other explanation styles is given, it means that the baseline is the provision of no explanations.

[11] Some authors measured the *accuracy* of the decisions or ratings of the study participants. Such a measurement is categorised as investigating the *effectiveness* of the explanations.

Table 13 User studies with positive results with respect to no explanations

Purpose	Explanation style
Effectiveness	Confidence + sensor data with low robot ability (Wang et al. 2016b)
	Decisive input values (Suermondt and Cooper 1993)
	Derived topic models and time intervals (Schaffer et al. 2015)
	Explanation interfaces (Pu and Chen 2007)
	Peer graph navigation (O'Donovan et al. 2009)
	Tag-based (Vig et al. 2009)
Transparency	Confidence + sensor data with low robot ability (Wang et al. 2016b)
	Decisive input values (Suermondt and Cooper 1993)
	Derived topic models and time intervals (Schaffer et al. 2015)
	Justification (Li and Gregor 2011)
	POMDP* translation (Wang et al. 2016a)
	Tag-based (Vig et al. 2009)
Persuasiveness	Peer information (Guy et al. 2009)
	Peer- and tag-based (Guy et al. 2010)
Satisfaction	Justification (Li and Gregor 2011)
	Music domain knowledge (Oramas et al. 2016)
	Pros and Cons (Felfernig and Gula 2006)
Trust	Confidence + sensor data with low robot ability (Wang et al. 2016b)
	How/why/trade-off (Wang et al. 2016c)
	POMDP* translation (Wang et al. 2016a)
	Pros and Cons (Felfernig and Gula 2006)
Usefulness	Decisive features (Muhammad et al. 2016)
	Match score + decisive features (Zanker and Ninaus 2010)
	Match score + decisive features (Zanker 2012)
Ease of use	Explanation interfaces (Pu and Chen 2007)
Efficiency	Decisive features (Zanker and Ninaus 2010)
Education	Decisive features (Reyes et al. 2011)

* POMDP stands for partially observable Markov decision process

4.5.2 Conclusions reached in user studies

We split the results of our analysis, i.e. the set of recorded tuples, into three parts. Studies that found positive effects of providing explanations *vs.* providing no explanations are listed in Table 13. Studies that report positive effects in a comparison with alternative explanation styles are given in Table 14. Studies with neutral and negative conclusions are shown in Table 15.

Table 14 User studies with positive results with respect to alternative explanations

Purpose	Explanation style
Effectiveness	Non-personalised decisive features (personalised, popularity): cameras (Tintarev and Masthoff 2012)
	(Non-)personalised tag-based (keyword-based) (Gedikli et al. 2011)
	(Non-)personalised tag-based (popularity, neighbours) (Gedikli et al. 2014)
	Predicted rating (popularity, single feature) (Hanshi et al. 2016)
	Trace (justification, strategy) (Tan et al. 2012)
	Visualisation (histogram) (Cleger-Tamayo et al. 2012)
Transparency	Personalised tag-based (non-personalised, confidence, neighbours) (Gedikli et al. 2014)
	Visualisation (histogram) (Cleger-Tamayo et al. 2012)
Persuasiveness	Histogram (20 explanations) (Herlocker et al. 2000)
	Justification (trace, strategy) (Ye 1995)
	Personalised features (people also viewed, no explanation) (Zhang et al. 2014)
	Social explanations (peers, personalisation) (Sharma and Cosley 2013)
	Social explanations + decisive features (authority, social proof) (Gkika and Lekakos 2014)
Satisfaction	Personalised decisive features (non-personalised, popularity): movies I (Tintarev and Masthoff 2012)
	Personalised decisive features (non-personalised): cameras (Tintarev and Masthoff 2012)
	Non-personalised decisive features (popularity): movies final (Tintarev and Masthoff 2012)
	(Non-)personalised tag-based (keyword-based) (Gedikli et al. 2011)
	Personalised tag-based (popularity, neighbours) (Gedikli et al. 2014)
	Predicted rating (popularity, single feature) (Hanshi et al. 2016)
Trust	Confidence + decisive features (confidence) (Bussone et al. 2015)
Usefulness	Contextualised deep explanations (non-contextualised) (Mao and Benbasat 2001)
	Trace (justification, strategy) (Tan et al. 2012)
Ease of use	Contextualised deep explanations (non-contextualised) (Mao and Benbasat 2001)
	Tag-based (item-based, feature-based) (Chen et al. 2013a)
Efficiency	(Non-)personalised tag-based (keyword-based) (Gedikli et al. 2011)
	(Non-)personalised tag-based (8 explanations) (Gedikli et al. 2014)
Education	

Table 15 User studies with neutral or negative results

Purpose	Explanation styles with	
	Neutral results	Negative results
Effectiveness	Confidence + sensor data with high robot ability (Wang et al. 2016b)	(Non-)personalised decisive features (popularity): movies final (Tintarev and Masthoff 2012)
	Hierarchic + deep explanations (Nakatsu and Benbasat 2003)	
	Match score (no recommendation, recommendation) (Ehrlich et al. 2011)	
	MovieLens (Herlocker et al. 2000)	
	(Non-)personalised decisive features (popularity): movies I (Tintarev and Masthoff 2012)	
	Non-personalised decisive features	
	(Personalised): movies II (Tintarev and Masthoff 2012)	
Transparency	Confidence + sensor data with high robot ability (Wang et al. 2016b)	
Persuasiveness		Tag-based (Guy et al. 2010)
Satisfaction	Personalised decisive features (non-personalised): movies II (Tintarev and Masthoff 2012)	
	Trace (justification, strategy) (Tan et al. 2012)	
Trust	Confidence + sensor data with high robot ability (Wang et al. 2016b)	
	Match score + decisive features (Zanker 2012)	
Usefulness		
Ease of use	Match score + decisive features (Zanker and Ninaus 2010)	
	Match score + decisive features (Zanker 2012)	
Efficiency		Derived topic models and time intervals (Schaffer et al. 2015)
		Pros and Cons (Felfernig and Gula 2006)
Education	Domain knowledge (Murphy and Phillips 1991)	

The largest amount of conclusions reached in the studies are related to the *effectiveness* of explanations, typically measured by the *explanation exposure delta*, which in this case the lower, the better.[12] These conclusions are also those that diverge the most. Given that the reported results concern different explanation styles, the observed divergence means that specific forms of explanations lead to more effective decisions. Moreover, this also suggests that there may be confounding variables in some stud-

[12] Higher deltas mean that users tend to overestimate or underestimate suggested alternatives based on explanations.

ies, such as the accuracy of the underlying decision inference method and the study domain, which may influence the observed outcomes. Three studies illustrate the possible existence of such confounding effects. A study in the robotics domain (Wang et al. 2016b) showed that explanations lead to higher effectiveness only *when the robot ability is low*. Ehrlich et al. (2011), who initially observed no statistical difference in their user study, based on a finer-grained analysis of their results, concluded that explanations are helpful *when the correct recommendation is provided*, which is not the case in the absence of such a recommendation. Furthermore, the four user studies reported by Tintarev and Masthoff (2012), which involved more than one domain, led to slightly different results regarding effectiveness—some results were not significant while others provided evidence that presenting popularity-based or non-personalised decisive features are more effective than presenting decisive features in a personalised way.

Contradicting results were also observed when the goal of the studies included the investigation of the *persuasiveness* of explanations. The data in Table 13 shows that persuasiveness was mainly achieved when explanations are based on social information, such as peer ratings. Negative results were obtained when tags were used as a basis for explanations. In addition, the studies that compare different explanation styles (Table 14) confirm the value of social information when designing persuasive explanations (Herlocker et al. 2000; Sharma and Cosley 2013; Gkika and Lekakos 2014). Only in one single study (Zhang et al. 2014), it turned out that a traditional explanation of the form "people also viewed" was less persuasive than *personalised features*. These are decisive features of an alternative selected based on preferences of the user receiving the recommendation.

Transparency and many of the user-centric purposes—*trust*, *satisfaction*, and *usefulness*—share similar results. Most of the studies indicate that explanations can in fact help to achieve these purposes. If not, the results do not provide evidence of negative effects. This is not the case, however, of *ease of use*. There are two studies associated with no effect in that respect, and only one reporting improvement. This improvement was achieved not only through the provision of explanatory information to users, but an enhanced user interface that categorises the suggested alternatives, possibly helping the user while analysing them. The non-existence of an effect on ease of use in the other studies is probably caused by the increased cognitive load for the users when more explanatory information is displayed. The fact that users have more information to process in such situations also explains the mostly negative effects of the provision of explanations on *efficiency*.

With respect to the purpose of *education*, there are only two studies, which reached different conclusions. Furthermore, none of the analysed user studies focused on the remaining purposes of explanations mentioned in the literature, namely *scrutability* and *debugging*.

Aside from the explanation purposes, some studies analysed the orthogonal aspect of *personalising* explanations. Gedikli et al. (2014) showed that a personalised version of an explanation approach based on tag clouds led to higher levels user-perceived transparency than the non-personalised version and had a modest positive effect on satisfaction. However, the opposite can be observed with respect to effectiveness and efficiency. Their earlier study (Gedikli et al. 2011) indicates that personalisation had

only a modest effect on efficiency, satisfaction, and effectiveness. Similarly, based on their four user studies, Tintarev and Masthoff (2012) concluded that personalised decisive features led to (sometimes modest) increased satisfaction. Nevertheless, this type of explanations caused a significant lower effectiveness in one of the studies.

A few studies are not reported in the summary tables as they focus on specific aspects of explanations. Ramberg (1996) analysed the impact of *different expertise levels* of the users and concluded that experts and novices have different preferences regarding the provided explanations. One study (Swinney 1995) investigated two explanation aspects: *direction*, which can be positive or negative, and *source*, which can be a decision support system or self-generated (by participants). The authors concluded that negative explanations are more influential than positive explanations, when they are generated by a decision support system. Finally, Gönül et al. (2006) focused on the particular explanation characteristics *confidence* and *length* and their study indicates that strongly confident and long explanations are more persuasive.

4.5.3 Conclusions reached in FOUNDATIONAL *Studies*

Our literature search returned only a few FOUNDATIONAL studies. A study performed by Tintarev and Masthoff (2007a) confirms the potential value of personalisation as discussed above. According to their findings (Tintarev and Masthoff 2007a), explanations should be customised to the user, focusing on an appropriate set of features of the suggested alternative. They also suggest that explanations should be tailored to the context. Furthermore, they proposed two additional guidelines that state that (i) from the many features of alternatives, considering only a short list from which personalised features are selected is enough to satisfy most of the users; and (ii) the source of the explanations matters (e.g. peers mentioned in the explanation). Additional guidelines and patterns that complement this study were proposed by Nunes et al. (2012a) which, for example, state that explanations should be concise and focus on the most relevant criteria.

Three additional user studies did not evaluate aspects of the explanations themselves (thus not classified as EVALUATION studies), but investigated extrinsic aspects that contribute to a better understanding of explanations. The impact of *mental models* was investigated by Rook and Donnell (1993), who concluded that it is important that users understand the expert system's reasoning process and the information provided in explanations to make good decisions. Giboney et al. (2015) found out that justifications that match user preferences (*cognitive fit*) are valuable for increasing the acceptance of recommendations of knowledge-based systems. The results show that when this match occurs, users tend to be more engaged, leading to longer interaction times with the system. Finally, Gregor (2001) investigated the usefulness of explanations in different contexts. The main outcome of the study is that explanations are more often accessed and helpful in cooperative problem solving situations. Moreover, when explanations are more often accessed, the problem-solving performance increases, particularly when system-user collaboration is required. Note that both these latter studies may indicate that more effective explanations are also those that lead to decreased efficiency.

5 Discussion

Based on the results from our systematic review, we present our insights in the field of explanations. Specifically, we (i) propose a new comprehensive taxonomy that captures the many facets that one might consider when designing an explanation facility for an advice-giving systems; and (ii) outline possible directions for future works. We also discuss limitations of our review.

5.1 Explanation taxonomy

A number of explanation taxonomies of different granularity levels has been proposed in the literature in the past (Ye and Johnson 1995; Chandrasekaran et al. 1989; Gregor and Benbasat 1999; Lacave and Díez 2002, 2004; Nakatsu 2006; Papadimitriou et al. 2012; Vig et al. 2009; Blanco et al. 2012; Friedrich and Zanker 2011; Langlotz and Shortliffe 1983). These taxonomies cover a variety of different aspects of explanation facilities, such as their purpose, the knowledge they use internally, or the information they convey to the user. Our review, however, revealed that there are a number of facets should be considered when designing a new explanation approach, which are not covered by these existing taxonomies.

The comprehensive new taxonomy that we discuss next is based both on the primary studies that were investigated in our review as well as on the existing—and sometimes not fully compatible—taxonomies from the literature. The underlying idea of our taxonomy is that explanations that are presented to the user consist of one or more *user interface components*. A component can be a justification in natural language, a histogram, or some other way of conveying information to the user. Our taxonomy, shown in Fig. 11, includes both general facets that are associated with explanations and their generation approach as well as facets related to the content and the presentation of individual explanation components (referred to as user interface components), which collectively comprise an explanation.

5.1.1 General facets of explanations

When designing an explanation approach, we should first determine *what is the objective* of providing explanations to the users. Previous research work described a number of possible *explanation purposes*, as discussed in Sect. 4.3. These possible purposes (e.g. transparency, efficiency, and trust) are often provided as a flat list, indicating that they are independent. In reality, as stated by Tintarev and Masthoff (2007b) and others, the possible purposes can, however, be related in different ways. Achieving transparency can, for example, constitute one of several factors that contribute to user trust in the system (Gedikli et al. 2011; Yasdi 1989; Mocanu 2015; Nilashi et al. 2016). Moreover, trust is usually not even the ultimate goal from the perspective of the provider of the advice-giving systems, who might be interested in increasing the user intention to continue to use the system in the future. In the spirit of Jannach and Adomavicius's work (2016), we distinguish between three levels of possible *objectives* of explanations:

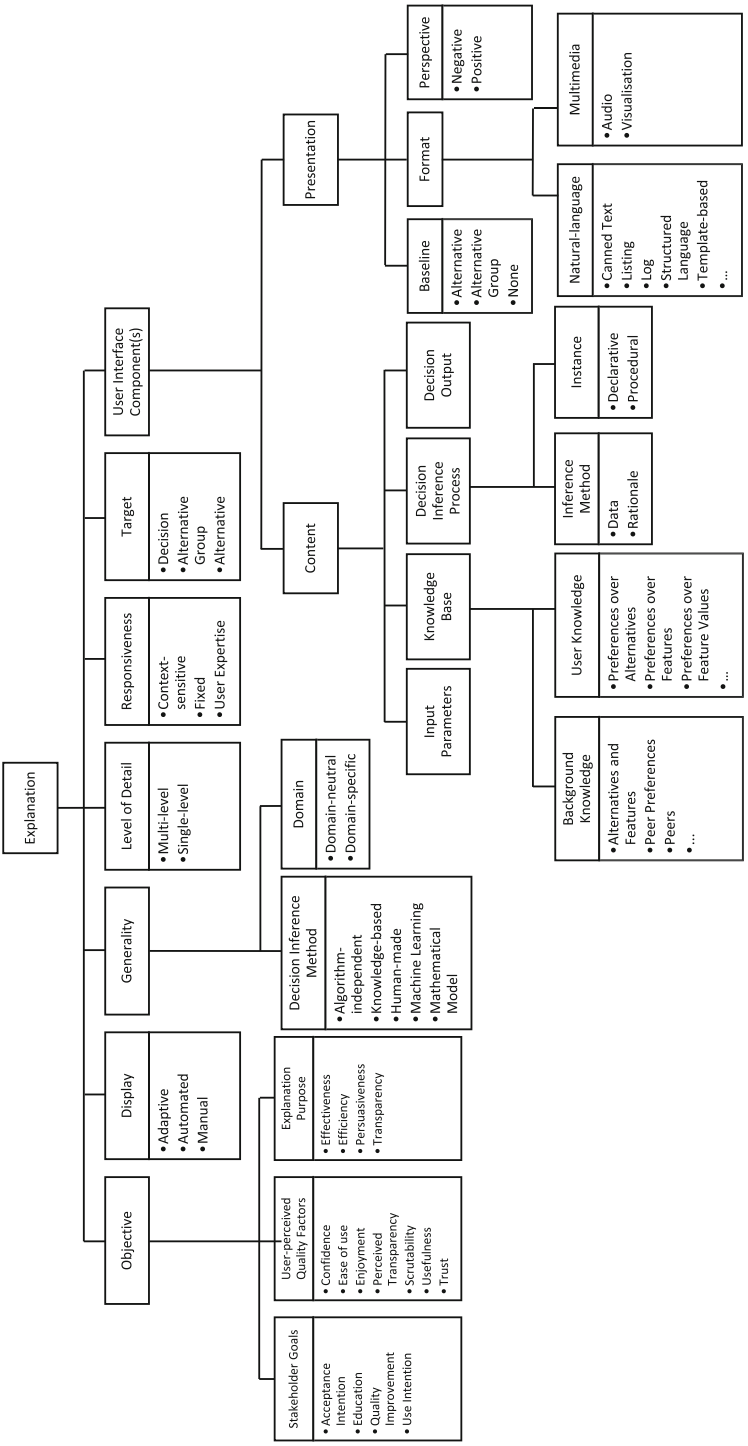

Fig. 11 Explanation taxonomy

(i) *stakeholder goals*, e.g. to increase the user's intention to reuse the system;
(ii) *user-perceived quality factors*, which are system quality factors that can contribute to the achievement of the stakeholder goals (e.g. trust); and
(iii) *explanation purposes*, which are explanation-specific objectives that can contribute to achieving the objectives at the other levels.

The key idea of having these three levels is that the choice and design of an explanation mechanism must be guided by the overarching goals that should be achieved with the explanations. Without clearly knowing the underlying motivations and in which ways the explanations are related to the quality perceptions of users, explanations might be added to a system in an arbitrary way. If, for example, the objective of the service provider is establish long-term customer relations (i.e. quality improvement), the service provided must optimise an adequate related quality factor at the user side (e.g. confidence). Based on the selected quality factor, the chosen explanation design should then support a corresponding purpose (e.g. effectiveness).

Based on previously proposed explanation purposes (Tintarev and Masthoff 2007b; Buchnan et al. 1984), the different types of objectives can be further refined as shown in Fig. 11. Additional details about the objectives at the lower levels are provided in Table 16. We highlight that, differently from Tintarev and Masthoff (2007b), we do not consider satisfaction as a single objective, but split it into ease to use, enjoyment, and usefulness. As shown in the analysis of EVALUATION studies, these are investigated and evaluated separately, in addition to being possibly conflicting. Explanations may be provided to increase the perceived usefulness of the system, although this may compromise ease of use. Moreover, note that we distinguish *perceived transparency* (a user-perceived quality factor) from *transparency* (an explanation purpose). The information provided to a user may be perceived as detailing how a system works. However, such information may not necessarily match how the system *actually* works.

Our taxonomy includes five additional general facets. With the explanation *Target* facet, we distinguish between situations in which the provided explanations refer to a single decision alternative, to a group of alternatives, or the decision output as a whole. The *Generality* facet captures whether an explanation is generated using information that comes from a particular domain and, therefore, the explanation generation process is domain-specific and not general enough to generate explanations in other domains. Furthermore, generality also refers to the question whether the approach to generate explanations is tied to a specific underlying decision inference method or if it is able to generate explanations only considering specific decision inference methods. With the *Responsiveness* facet, we distinguish between explanations that are adapted to the current user context and those that are not. In addition, the *Level of Detail* indicates that explanations with more or less details can be provided, depending on a certain criterion, such as user expertise. More detailed explanations are more informative, but can require a higher cognitive effort from the user. To deal with this trade-off, there are approaches that provide multiple levels of detail, which can be explored by users as needed or displayed according to the current context. Finally, the *Display* facet characterises what triggers an explanation to be displayed. The usual alternatives are manual (shown upon user request), automatic (always shown), or adaptive (depending on the context).

Table 16 Description of the objective facet

Category	Property	Description
Stakeholder	Acceptance intention	Increasing the probability of users accepting the suggestion (referred to as purchase intention in e-commerce systems)
Goals	Education	Providing users with knowledge to make decisions in the system domain
	Use intention	Increasing the probability of users using the system
	Quality	Improving the quality of user decisions (in terms of correctness) by detecting possible system flaws.
	Improvement	
User-perceived	Confidence	Being perceived as a system that helps users make good decisions, i.e. making users confident of the decision quality
Quality factors	Ease of use	Being perceived as easy to use
	Enjoyment	Being perceived as a system that brings enjoyment to users
	Perceived transparency	Being perceived as transparent, i.e. a system that exposes its inner workings
	Scrutability	Being able to receive and use user feedback about the decision advice
	Usefulness	Being perceived as a useful system
	Trust	Being perceived as a trustworthy system
Explanation	Effectiveness	Providing information to allow assessment of whether the suggested alternative is appropriate
Purposes	Efficiency	Providing information to help users make faster decisions
	Persuasiveness	Providing information to convince users that the suggested alternative is appropriate
	Transparency	Providing information to understand the inference logic of the advice-giving system

5.1.2 User interface components

The other facets of our taxonomy are concerned with aspects related to the *content* and the *presentation* of explanations. This part of the taxonomy is mainly orthogonal to the previously introduced facets. However, the design decisions that are made along the dimensions of these general facets can impact on these remaining choices. For example, if the explanation target is the decision as a whole, a baseline cannot be selected for a user interface component, because baselines make only sense when comparing alternatives.

Regarding the *Content* facet, we identified four key types of information that can be presented in explanations, detailed as follows.

Input parameters User interface components that refer to the inputs that were provided for a particular decision problem, e.g. the set of symptoms that were

entered in a medical decision support system or the current user mood in the case of a movie recommender system.
Knowledge base User interface components that include information that resides in the knowledge base of an advice-giving systems. The provided information can be personalised, i.e. tailored to the specific user that receives the advise, or it can be background (or world) knowledge that is selected independent of the current user. Examples of the different types of knowledge are provided in Fig. 11.
Decision inference process User interface components can also include information that is related to the system's internal process of determining the suggested alternatives. Such explanations can refer to a specific decision problem instance or provide general information about the internal inference method. In this latter case, the system can either explain the general idea behind the algorithm (e.g. recommendation of alternatives that similar users like) or the data it uses (e.g. use of users' shopping history to identify what they like). When the explanations are tied to the specific decision problem instance, the explanations can be procedural, i.e. describe the steps taken to reach a decision (e.g. rule trace), or declarative, providing information such as the confidence in the decision.
Decision output Finally, user interface components can focus on the decision reasoning outcome and, for example, describe the particular features and feature values of the recommended and non-recommended alternatives. The explanation style *Pros and Cons* presented in Table 7 illustrates an explanation component that falls into this category.

Looking at the *Presentation* facet of the taxonomy, there are three sub-facets. First, there are different ways of including none or multiple *baselines* for comparison in the explanation. The baseline (or anchor) can be a single alternative to that recommended or a group of alternatives. Second, different output *formats* can be chosen, as discussed previously in our review, such as using natural language or different types of visualisations. We list possible alternatives in their corresponding boxes. For example, canned text consists of a set of text segments that, when combined, form one sentence. Templates, in contrast, are usually almost complete sentences that must be completed with a set of arguments. Finally, the *perspective* in which an explanation is presented can either be *positive*, i.e. focusing on why an alternative is suitable for a user, or *negative*, i.e. detailing why certain negative aspects of an alternative could be acceptable.

5.2 Future directions

While a substantial amount of work has already been done in the context of explanations, a variety of open issues still need to be addressed. In this section, we give examples of such open research questions. These questions refer to several *concerns* that need further investigation, ranging from questions related to the general objectives of explanations, over questions regarding the choice of the explanation content, to open methodological issues.

Understanding the relationship among stakeholder goals, user-perceived quality factors, and explanation purposes In the previous section, we argued that the explanation

purposes that are mentioned in the literature can be different in nature and some are only indirectly related to explanations. Correspondingly, we distinguished between Stakeholder Goals, User-Perceived Quality Factors, and Explanation Purposes in our taxonomy. These objectives are interrelated, e.g. transparency can have an impact on trust. Although one can form intuitive hypotheses about the relationships among these objectives, more systematic studies of these relationships still need to be done.

Selecting the right explanation content As shown in our taxonomy and in the discussion regarding explanation content, various types of information can be presented within explanations. Most of the user studies that we examined in this work compared largely different forms of explanations and there is no common *baseline* explanation form that is used in many studies. Furthermore, the compared explanation forms often vary in many different aspects so that it is not possible to understand what content should be presented to which kind of users and when. In the early years of expert systems, explanations often focused on the decision inference process and provided inference traces. However, it soon became clear that explanations that focus on *why an alternative is adequate* are more helpful for users. Based on these considerations, we argue that explanation generation approaches should be further investigated in the future, being as independent as possible of the underlying decision inference process, thereby increasing the possibility to reuse approaches for different types of advice-giving systems. In our systematic review, only a small number (11) of primary studies consists of approaches that are detached from the underlying decision inference method. As a consequence, with the increasing adoption of complex machine learning methods, such forms of explanations will become increasingly more important in the future. The same holds for application domains in which the internal inference methods are confidential (Blanco et al. 2012). Nevertheless, algorithm-independent explanations might not be appropriate when systems autonomously make decisions regarding individuals. Regulations, such as the General Data Protection Regulation[13], have emerged to give citizens the right to go against algorithmic-based decisions that affect them. Consequently, the provision of *transparent* explanations is required in these cases.

Investigating fine-grained details of presentation aspects Various questions are also open with respect to the fine-grained details of how to present explanations to the users. Explanations can vary, for example, with respect to (i) their length; (ii) the adopted vocabulary if natural language is used; (iii) the presentation format, and so on. When explanation forms are compared in user studies that are entirely different in these respects, it is impossible to understand how these details impact the results. Therefore, more studies are required to investigate the impact of these variables. Such user studies could then provide a more solid foundation for the development of new explanation approaches. From all investigated studies, only two (Bader et al. 2012; Bohnenberger et al. 2005) of the proposed explanation forms were explicitly founded and motivated by a preliminary study. Many of the others studies might

[13] http://www.eugdpr.org/.

have used existing research results in the literature as foundation of their work. However, this could have been highlighted to justify design decisions associated with explanations.

Towards more responsive explanations Intelligent advice-giving systems are adopted in a wide range of domains. In some of them (like health), decisions are more critical than in others (like movie recommendation). Therefore, it is important to better understand which forms of explanations are appropriate for which scenarios. When designing an explanation facility one should, for example, consider how much effort users might be willing to make in order to analyse explanations. Moreover, taking into account the user who is interacting with the system, and her background knowledge, is also relevant—as discussed in some of the analysed studies. However, only 16 explanation TECHNIQUES and TOOLS, i.e. 8% of all studies of these types, aimed at providing explanations that are tailored to a given context or user expertise.

The need for adequate objective evaluation protocols and metrics The most common way of assessing different explanation aspects, e.g. transparency or persuasiveness, is to use questionnaires that the participants fill out as part of or after an experiment, which is a well-established research instrument. Such studies rely on the participants' subjective perception of certain system or explanation aspects and on their behavioural intentions. Nonetheless, this research approach has limitations, as study participants may find it difficult to express to what extent they feel persuaded by a system or consider the system's explanations transparent, for example. A further limitation is that there are no standardised study designs or list of questionnaire items in the field. Consequently, it is important that researchers develop a standardised set of evaluation protocols to measure certain aspects of explanations. Moreover, these protocols should rely on *objective* measures, in addition to the subjective quality perception statements. One example of such a protocol is what we referred to as the *explanation exposure delta*. However, this protocol can only be used to assess certain aspects, and more work towards a comprehensive evaluation framework for explanations is still required.

5.3 Research limitations

The main goal of our systematic review is to develop a comprehensive understanding of what has been done in the field of explanations based on an unbiased selection and analysis of a large amount of primary studies. However, due to the systematic process of selecting the studies from a specified set of digital libraries, our survey does not cover all existing work on explanations. We generally selected widely used digital libraries as sources, which we assumed that would contain the largest number of relevant studies. An example of a comparably often cited study that was not retrieved in our search is that of Bilgic et al. (2005), because it was published in workshop proceedings. However, we believe that the number of relevant studies that were published only in a workshop and were not continued in a conference or journal paper is low. A few other relevant studies (Nunes et al. 2014; Junker 2004; Carenini and Moore 2001) were also

not included in our review, because they are part of the the AAAI Digital Library[14], which unfortunately provides too limited search support to be usable for our survey. Nevertheless, the advantage of systematic reviews is that they can be further extended in future reviews, which can follow the specified procedure with other databases or in a future publication time range.

6 Summary

The increasing trend of using software systems as advice givers and also making them more autonomous calls for approaches that allow systems to be able to *explain* their decisions. Due to this trend, explanations increasingly receive more attention. However, a substantial amount of work has already been done in the field of explanations in advice-giving systems, mainly in the context of expert systems. Therefore, to have a comprehensive view of what has been done in this field, we presented the results of a systematic review of studies that proposed new techniques to generate explanations, described tools that include an explanation facility, or detailed the results of evaluation and foundational studies. A wide range of aspects associated with explanations were discussed in the 217 analysed studies.

We observed that most of the explanations provided in existing work consist of inference traces that were collected during the internal process of reasoning about which alternative(s) should be suggested to users. To be presented, such traces are often transformed into natural language statements. The adoption of this form of explanation was mainly due to the underlying inference method, which was some form of rule-based reasoning in most cases. The use of traces also explains the most frequent intended purpose of providing explanations, which is transparency (i.e. detail how a system reached a particular conclusion), so that users can trust the system. However, such traces are often not helpful for end users. Therefore, other forms of explanations were explored, such as those that focus on the features of the suggested alternatives and those that present visualisations rather than natural-language statements. This also led to the exploitation of explanations to persuade users or increase their satisfaction while interacting with the system. By analysing the results of studies which evaluated and compared explanations, we observed that there is a lack of standardised protocols and measurements to guide the study design, which makes a comparison of the obtained results difficult. However, a deeper analysis of results that were obtained mainly through user studies indicated divergence in terms of the reached conclusions. Consequently, there is a need for conducting further studies to identify the real pros and cons of explanations.

Based on our literature review, we proposed an explanation taxonomy, which covers a wide range of facets that can be used as a guideline by researchers when proposing and evaluating future approaches to generate explanations. Our taxonomy highlights the importance of specifying clear objectives as a key driver when designing new approaches, with objectives ranging from stakeholder goals to the specific explanation purposes. Moreover, we discussed open challenges that remain open, such as the

[14] http://www.aaai.org/Library/library.php.

understanding of the right explanation content according to different contexts and the subtleties of the explanations that may have an impact on the user, e.g. the used vocabulary. The identified challenges leave room for much research work that needs to be done in order for users to develop trust towards future advice-giving systems and autonomous systems.

Our review focused on explanations provided to end users, who need further information when receiving recommendations from advice-giving systems, with varying purposes. However, due to the increasing complexity of machine learning techniques, including those based on deep learning, providing explanations for data scientists to understand the outcomes of these techniques is becoming crucial, leading to what is generally referred to as explainable artificial intelligence. This broader topic is out of the scope of our work but should be further investigated in the future.

Acknowledgements The authors would like to thank Michael Jugovac for carefully proofreading this paper. Ingrid Nunes also would like to thank for research grants CNPq ref. 303232/2015-3, CAPES ref. 7619-15-4, and Alexander von Humboldt, ref. BRA 1184533 HFSTCAPES-P.

References

Abu-Hakima, S., Oppacher, F.: Rationale: reasoning by explaining. In: Proceedings of the Fourth International Conference on Data Engineering, pp. 258–265 (1988)

Allgayer, J., Harbusch, K., Kobsa, A., Reddig, C., Reithinger, N., Schmauks, D.: XTRA: a natural-language access system to expert systems. Int. J. Man Mach. Stud. **31**(2), 161–195 (1989)

Amer-Yahia, S., Galland, A., Stoyanovich, J., Yu, C.: From Del.Icio.Us to x.Qui.Site: recommendations in social tagging sites. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD '08, pp. 1323–1326 (2008)

Artioli, E., Avanzolini, G., Martelli, L., Ursino, M.: An expert system based on causal knowledge: validation on post-cardiosurgical patients. Int. J. Bio Med. Comput. **41**(1), 19–37 (1996)

Bader, R., Woerndl, W., Karitnig, A., Leitner, G.: Designing an Explanation Interface for Proactive Recommendations in Automotive Scenarios, pp. 92–104. Springer, Berlin (2012)

Balleda, K., Satyanvesh, D., Sampath, N.V.S.S.P., Varma, K.T.N., Baruah, P.K.: Agpest: an efficient rule-based expert system to prevent pest diseases of rice amp; wheat crops. In: 2014 IEEE 8th International Conference on Intelligent Systems and Control (ISCO), pp. 262–268 (2014)

Banavar, G.: Learning to Trust Artificial Intelligence Systems: Accountability, Compliance and Ethics in the Age of Smart Machines. White paper, IBM Global Services (2016)

Barbieri, N., Bonchi, F., Manco, G.: Who to follow and why: Link prediction with explanations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, pp. 1266–1275 (2014)

Basu, A., Dutta, A.: Computer based support of reasoning in the presence of fuzziness. Decis. Support Syst. **2**(3), 235–256 (1986)

Basu, A., Ahad, R.: Using a relational database to support explanation in a knowledge-based system. IEEE Trans. Knowl. Data Eng. **4**(6), 572–581 (1992)

Basu, A., Majumdar, A.K., Sinha, S.: An expert system approach to control system design and analysis. IEEE Trans. Syst. Man Cybern. **18**(5), 685–694 (1988)

Bau, D.Y., Brezillon, P.J.: Model-based diagnosis of power-station control systems. IEEE Expert **7**(1), 36–44 (1992)

Bavota, G., Gethers, M., Oliveto, R., Poshyvanyk, D., Lucia, Ad: Improving software modularization via automated analysis of latent topics and dependencies. ACM Trans. Softw. Eng. Methodol. **23**(1), 4:1–4:33 (2014)

Bedi, P., Sharma, R.: Trust based recommender system using ant colony for trust computation. Expert Syst. Appl. **39**(1), 1183–1190 (2012)

Bedi, P., Agarwal, S.K., Sharma, S., Joshi, H.: Saprs: situation-aware proactive recommender system with explanations. In: 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 277–283 (2014)

Beiley, J., Duban, S.: Explanation and learning in medicine. In: Kibby, M. (ed.) Computer Assisted Learning, pp. 91–97. Pergamon, Amsterdam (1990)

Belahcene, K., Labreuche, C., Maudet, N., Mousseau, V., Ouerdane, W.: Explaining robust additive utility models by sequences of preference swaps. Theory Decis. **82**(2), 151–183 (2017)

Benaroch, M.: Roles of design knowledge in knowledge-based systems. Int. J. Hum. Comput. Stud. **44**(5), 689–721 (1996)

Bielza, C., Gómez, M., Ríos-Insua, S., Fernándezdel Pozo, J.A.: Structural, elicitation and computational issues faced when solving complex decision making problems with influence diagrams. Comput. Oper. Res. **27**(78), 725–740 (2000)

Bilgic, M., Mooney, R.J.: Explaining recommendations: satisfaction vs. promotion. In: Proceedings of Beyond Personalization 2005: A Workshop on the Next Stage of Recommender Systems Research at the 2005 International Conference on Intelligent User Interfaces. San Diego, CA (2005)

Blake, J.N., Kerr, D.V., Gammack, J.G.: Streamlining patient consultations for sleep disorders with a knowledge-based CDSS. Inf. Syst. **56**, 109–119 (2016)

Blanco, R., Ceccarelli, D., Lucchese, C., Perego, R., Silvestri, F.: You should read this! let me explain you why: explaining news recommendations to users. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12, pp. 1995–1999 (2012)

Bofeng, Z., Na, W., Gengfeng, W., Sheng, L.: Research on a personalized expert system explanation method based on fuzzy user model. In: Fifth World Congress on Intelligent Control and Automation, vol. 5, pp. 3996–4000 (2004)

Bohanec, M., Zupan, B., Rajkovič, V.: Applications of qualitative multi-attribute decision models in health care. Int. J. Med. Inform. **5859**, 191–205 (2000)

Bohnenberger, T., Jacobs, O., Jameson, A., Aslan, I.: Decision-Theoretic Planning Meets User Requirements: Enhancements and Studies of an Intelligent Shopping Guide, pp. 279–296. Springer, Berlin (2005)

Borlea, l., Buta, A., Dusa, V., Lustrea, B.: DIASE—expert system fault diagnosis for Timisoara 220 kV substation. In: EUROCON 2005—The International Conference on "Computer as a Tool", vol. 1, pp. 221–224 (2005)

Bosnić, Z., Vraćar, P., Radović, M.D., Devedžić, G., Filipović, N.D., Kononenko, I.: Mining data from hemodynamic simulations for generating prediction and explanation models. IEEE Trans. Inf. Technol. Biomed. **16**(2), 248–254 (2012)

Bostandjiev, S., O'Donovan, J., Höllerer, T.: Tasteweights: a visual interactive hybrid recommender system. In: Proceedings of the Sixth ACM Conference on Recommender Systems, RecSys '12, pp. 35–42 (2012)

Briguez, C.E., Budn, M.C., Deagustini, C.A., Maguitman, A.G., Capobianco, M., Simari, G.R.: Argument-based mixed recommenders and their application to movie suggestion. Expert Syst. Appl. **41**(14), 6467–6482 (2014)

Buchanan, B.G., Shortliffe, E.H. (eds.): Explanations as a topic of AI research. In: Rule-Based Systems, pp. 331–337. Addison-Wesley, Massachusetts (1984)

Buchanan, B.G., Moore, J.D., Forsythe, D.E., Carenini, G., Ohlsson, S., Banks, G.: An intelligent interactive system for delivering individualized information to patients. Artif. Intell. Med. **7**(2), 117–154 (1995)

Burattini, E., Gregorio, M.D., Tamburrini, G.: Hybrid expert systems: An approach to combining neural computation and rule-based reasoning. In: Leondes, C.T. (ed.) Expert Systems, pp. 1315–1354. Academic Press, Burlington (2002)

Buschner, S., Schirru, R., Zieschang, H., Junker, P.: Providing recommendations for horizontal career change. In: Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business, i-KNOW '14, pp. 33:1–33:4 (2014)

Bussone, A., Stumpf, S., O'Sullivan, D.: The role of explanations on trust and reliance in clinical decision support systems. In: 2015 International Conference on Healthcare Informatics, pp. 160–169 (2015)

Cagnoni, S., Coppini, G., Livi, R., Poli, R., Scarpelli, P.T., Valli, G.: A neural network expert system for computer-assisted analysis of blood-pressure data. In: Proceedings Computers in Cardiology, pp. 473–476 (1991)

Carenini, G., Moore, J.D.: An empirical study of the influence of user tailoring on evaluative argument effectiveness. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence, IJCAI'01, pp. 1307–1312 (2001)
Carenini, G., Moore, J.D.: Generating and evaluating evaluative arguments. Artif. Intell. **170**, 925–952 (2006)
Castro, C., Bose, A., Handschin, E., Hoffmann, W.: Comparison of different screening techniques for the contingency selection function. Int. J. Electr. Power Energy Syst. **18**(7), 425–430 (1996)
Chandrasekaran, B., Mittal, S.: Deep versus compiled knowledge approaches to diagnostic problem-solving. Int. J. Hum. Comput. Stud. **51**(2), 357–368 (1999)
Chandrasekaran, B., Tanner, M.C., Josephson, J.R.: Explaining control strategies in problem solving. IEEE Expert Intell. Syst. Appl. **4**(1), 9-15–19-24 (1989)
Chang, C.C., Hsieh, S.C.: Applying web service technology to build a wireless lan problem diagnosis expert system. In: 2010 International Conference on Computational Aspects of Social Networks, pp. 217–220 (2010)
Charissiadis, A., Karacapilidis, N.: Strengthening the Rationale of Recommendations Through a Hybrid Explanations Building Framework, pp. 311–323. Springer, Berlin (2015)
Chelsom, J.J., Ellis, T.J., Carson, E.R., Cramp, D.G.: Blood gas analysis: a knowledge-based adviser for the interpretation of results. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1360–1361 (1988)
Chen, L., Wang, F.: Sentiment-enhanced explanation of product recommendations. In: Proceedings of the 23rd International Conference on World Wide Web, WWW '14 Companion, pp. 239–240 (2014)
Chen, W., Hsu, W., Lee, M.L.: Tagcloud-based explanation with feedback for recommender systems. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13, pp. 945–948 (2013a)
Chen, Y.C., Lin, Y.S., Shen, Y.C., Lin, S.D.: A modified random walk framework for handling negative ratings and generating explanations. ACM Trans. Intell. Syst. Technol. **4**(1), 12:1–12:21 (2013b)
Cheng, S.J., Chen, D.S., Peng, X.L.: An expert system for a thermal power station alarm processing. In: International Conference on Advances in Power System Control, Operation and Management, APSCOM-91, pp. 316–320 (1991)
Chiou, A., Yu, X.: Industrial decision support system (IDSS) in weed control and management strategies: expert advice using descriptive schemata and explanatory capabilities. In: IECON 2007—33rd Annual Conference of the IEEE Industrial Electronics Society, pp. 105–110 (2007)
Chouicha, M., Siller, T.: An expert system approach to liquefaction analysis part 1: development and implementation. Comput. Geotech. **16**(1), 1–35 (1994)
Cleger-Tamayo, S., Fernandez-Luna, J.M., Huete, J.F.: Explaining neighborhood-based recommendations. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12, pp. 1063–1064 (2012)
Davey-Wilson, I.: Development of a prolog-based expert system for groundwater control. Comput. Struct. **40**(1), 185–189 (1991)
David, J.M., Krivine, J.P.: Designing knowledge-based systems within functional architecture: the DIVA experiment. In: Proceedings of the Fifth Conference on Artificial Intelligence Applications, pp. 173–180 (1989)
Davis, K.: DORIS (diagnostic oriented rockwell intelligent system). IEEE Aerosp. Electron. Syst. Mag. **1**(7), 18–21 (1986)
de Braal, L., Ezquerra, N., Garcia, E., Cooke, C., Krawczynska, E.: PERFUSE: an interactive knowledge-based system for the interpretation and explanation of cardiac imagery. In: Proceedings of 18th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 3, pp. 1238–1239 (1996)
Deep, R., Czech, D.R., Dizek, S.G., Kennedy, D.K.: A bit-mapping classifier expert system in warranty selection. In: Proceedings of the IEEE 1988 National Aerospace and Electronics Conference, pp. 1222–1224 (1988)
Dhaliwal, J.S., Benbasat, I.: The use and effects of knowledge-based system explanations: theoretical foundations and a framework for empirical evaluation. Inf. Syst. Res. **7**(3), 342–362 (1996)
Diederich, J.: Explanation and artificial neural networks. Int. J. Man Mach. Stud. **37**(3), 335–355 (1992)
Du, G., Ruhe, G.: Two machine-learning techniques for mining solutions of the releaseplanner decision support system. Inf. Sci. **259**, 474–489 (2014)

Ehrlich, K., Kirk, S.E., Patterson, J., Rasmussen, J.C., Ross, S.I., Gruen, D.M.: Taking advice from intelligent systems: the double-edged sword of explanations. In: Proceedings of the 16th International Conference on Intelligent User Interfaces, IUI '11, pp. 125–134 (2011)

Ezquerra, N., de Braal, L., Garcia, E., Cooke, C., Krawczynska, E.: Interactive, knowledge-guided visualization of 3D medical imagery. Future Gener. Comput. Syst. **15**(1), 59–73 (1999)

Felfernig, A.: Koba4ms: selling complex products and services using knowledge-based recommender technologies. In: Seventh IEEE International Conference on E-Commerce Technology (CEC'05), pp. 92–100 (2005)

Felfernig, A., Gula, B.: An empirical study on consumer behavior in the interaction with knowledge-based recommender applications. In: The 8th IEEE International Conference on E-Commerce Technology and the 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEE'06), pp. 37–37 (2006)

Fong, J., Lam, H.P., Robinson, R., Indulska, J.: Defeasible preferences for intelligible pervasive applications to enhance eldercare. In: IEEE International Conference on Pervasive Computing and Communications Workshops, pp. 572–577 (2012)

Friedrich, G., Zanker, M.: A taxonomy for generating explanations in recommender systems. AI Mag. **32**(3), 90–98 (2011)

Gallagher, S., Trainor, J., Murphy, M., Curran, E.: A knowledge based system for competitive bidding. In: Proceedings of the 1995 Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems, pp. 314–317 (1995)

Garca, A.J., Chesevar, C.I., Rotstein, N.D., Simari, G.R.: Formalizing dialectical explanation support for argument-based reasoning in knowledge-based systems. Expert Syst. Appl. **40**(8), 3233–3247 (2013)

Gedikli, F., Ge, M., Jannach, D.: Understanding Recommendations by Reading the Clouds, pp. 196–208. Springer, Berlin (2011)

Gedikli, F., Jannach, D., Ge, M.: How should i explain? A comparison of different explanation types for recommender systems. Int. J. Hum. Comput. Stud. **72**(4), 367–382 (2014)

Giboney, J.S., Brown, S.A., Lowryc, P.B., Nunamaker Jr., J.F.: User acceptance of knowledge-based system recommendations: explanations, arguments, and fit. Decis. Support Syst. **72**, 1–10 (2015)

Gkika, S., Lekakos, G.: Investigating the effectiveness of persuasion strategies on recommender systems. In: Proceedings of the 9th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP '14, pp. 94–97. IEEE Computer Society, Washington, DC, USA (2014)

Glaser, B.G.: Basics of Grounded Theory Analysis: Emergence vs. Forcing. Sociology Pr, Mill Valley (1992)

Gómez-Vallejo, H.J., Uriel-Latorre, B., Sande-Meijide, M., Villamarín-Bello, B., Pavón, R., Fdez-Riverol, F., Glez-Peña, D.: A case-based reasoning system for aiding detection and classification of nosocomial infections. Decis. Support Syst. **84**, 104–116 (2016)

Gönül, M.S., Önkal, D., Lawrence, M.: The effects of structural characteristics of explanations on use of a dss. Decis. Support Syst. **42**(3), 1481–1493 (2006)

Goud, R., Hasman, A., Peek, N.: Development of a guideline-based decision support system with explanation facilities for outpatient therapy. Comput. Methods Progr. Biomed **91**(2), 145–153 (2008)

Gowri, K., Marsh, C., Bedard, C., Fazio, P.: Knowledge-based assistant for aluminum component design. Comput. Struct. **38**(1), 9–20 (1991)

Grando, M.A., Moss, L., Glasspool, D., Sleeman, D., Sim, M., Gilhooly, C., Kinsella, J.: Argumentation-Logic for Explaining Anomalous Patient Responses to Treatments, pp. 35–44. Springer, Berlin (2011)

Gregor, S.: Explanations from knowledge-based systems and cooperative problem solving. Int. J. Hum. Comput. Stud. **54**(1), 81–105 (2001)

Gregor, S., Benbasat, I.: Explanations from intelligent systems: theoretical foundations and implications for practice. MIS Q. **23**(4), 497–530 (1999)

Grierson, D.E., Cameron, G.E.: A knowledge-based expert system for computer automated structural design. Comput. Struct. **30**(3), 741–745 (1988)

Guida, G., Zanella, M.: Active operator support: a case study in steel production. In: IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century, vol. 4, pp. 3340–3345 (1995)

Guida, G., Mussio, P., Zanella, M.: User interaction in decision support systems: the role of justification. In: 1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation, vol. 4, pp. 3215–3220 (1997)

Guy, I., Zwerdling, N., Carmel, D., Ronen, I., Uziel, E., Yogev, S., Ofek-Koifman, S.: Personalized recommendation of social software items based on social relations. In: Proceedings of the Third ACM Conference on Recommender Systems, RecSys '09, pp. 53–60 (2009)

Guy, I., Zwerdling, N., Ronen, I., Carmel, D., Uziel, E.: Social media recommendation based on people and tags. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '10, pp. 194–201 (2010)

Gvenir, H., Emeksiz, N.: An expert system for the differential diagnosis of erythemato-squamous diseases. Expert Syst. Appl. **18**(1), 43–49 (2000)

Hair, D.C., Pickslay, K., Chow, S.: Explanation-based decision support in real time situations. In: Proceedings of the Fourth International Conference on Tools with Artificial Intelligence, TAI '92, pp. 22–25 (1992)

Hanshi, W., Qiujie, F., Lizhen, L., Wei, S.: A probabilistic rating prediction and explanation inference model for recommender systems. China Commun. **13**(2), 79–94 (2016)

Hasling, D.W., Clancey, W.J., Rennels, G.: Strategic explanations for a diagnostic consultation system. Int. J. Man Mach. Stud. **20**(1), 3–19 (1984)

Hatzilygeroudis, I., Prentzas, J.: Symbolic-neural rule based reasoning and explanation. Expert Syst. Appl. **42**(9), 4595–4609 (2015)

Helms, G.L., Richardson, J.W., Cochran, M.J., Rister, M.: A farm level expert simulation system to aid farmers in selecting among crop insurance strategies. Comput. Electron. Agric. **4**(3), 169–190 (1990)

Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, CSCW '00, pp. 241–250 (2000)

Hodgkinson, L., Walker, E.: An expert system for credit evaluation and explanation. J. Comput. Sci. Coll. **19**(1), 62–72 (2003)

Holman, J.G., Wolff, A.H.: An expert adviser for oliguria occurring on the intensive care unit. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1442–1443 (1988)

Horan, J., O'Sullivan, B.: Towards diverse relaxations of over-constrained models. In: 2009 21st IEEE International Conference on Tools with Artificial Intelligence, pp. 198–205 (2009)

Horn, W., Popow, C., Miksch, S., Seyfang, A.: Quicker, more accurate nutrition plans for newborn infants. IEEE Intell. Syst. Appl. **13**(1), 65–69 (1998)

Hornung, T., Ziegler, C.N., Franz, S., Przyjaciel-Zablocki, M., Schtzle, A., Lausen, G.: Evaluating hybrid music recommender systems. In: IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), vol. 1, pp. 57–64 (2013)

Hu, Y., Koren, Y., Volinsky, C.: Collaborative filtering for implicit feedback datasets. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 263–272 (2008)

Hudson, D.L., Cohen, M.E.: Human–computer interaction in a medical decision support system. In: Proceedings of the Twenty-Second Annual Hawaii International Conference on System Sciences. Volume II: Software Track, vol. 2, pp. 429–435 (1989)

Hunt, J., Price, C.: Explaining qualitative diagnosis. Eng. Appl. Artif. Intell. **1**(3), 161–169 (1988)

Hussain, S., Abidi, S.S.R.: Ontology driven CPG authoring and execution via a semantic web framework. In: 40th Annual Hawaii International Conference on System Sciences, HICSS 2007, pp. 135–135 (2007)

Hussein, T., Neuhaus, S.: Explanation of spreading activation based recommendations. In: Proceedings of the 1st International Workshop on Semantic Models for Adaptive Interactive Systems, SEMAIS '10, pp. 24–28 (2010)

Jabri, M.A.: Knowledge-based system design using prolog: the PIAF experience. Knowl. Based Syst. **2**(1), 72–79 (1989)

Jaimes, A., Gatica-Perez, D., Sebe, N., Huang, T.S.: Guest editors' introduction: human-centered computing-toward a human revolution. Computer **40**(5), 30–34 (2007)

Jamieson, P.W.: A model for diagnosing and explaining multiple disorders. Comput. Biomed. Res. **24**(4), 307–320 (1991)

Janjua, N.K., Hussain, F.K.: Defeasible reasoning based argumentative Web-IDSS for virtual teams (VTs). In: 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 330–334 (2011)

Jannach, D., Adomavicius, G.: Recommendations with a purpose. In: Proceedings of the 2016 ACM Conference on Recommender Systems, RecSys '16, pp. 7–10 (2016)

Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender Systems: An Introduction. Cambridge University Press, New York (2010)
Jannach, D., Resnick, P., Tuzhilin, A., Zanker, M.: Recommender systems—beyond matrix completion. Commun. ACM **59**(11), 94–102 (2016)
Ji, K., Shen, H.: Jointly modeling content, social network and ratings for explainable and cold-start recommendation. Neurocomputing **218**, 1–12 (2016)
Joch, J., Dudeck, J.: Decision support for infectious diseasesa working prototype. Int. J. Med. Inform. **64**(23), 331–340 (2001)
Jugovac, M., Jannach, D.: Interacting with recommenders—overview and research directions. ACM Trans. Interact. Intell. Syst. **7**(3), 46 (2017)
Jung, D., Burns, J.R.: Connectionist approaches to inexact reasoning and learning systems for executive and decision support. Decis. Support Syst. **10**(1), 37–66 (1993)
Junker, U.: Quickxplain: preferred explanations and relaxations for over-constrained problems. In: AAAI'04, pp. 167–172. USA (2004)
Kadhim, M.A., Alam, M.A., Kaur, H.: Design and implementation of intelligent agent and diagnosis domain tool for rule-based expert system. In: 2013 International Conference on Machine Intelligence and Research Advancement, pp. 619–622 (2013)
Kagal, L., Pato, J.: Preserving privacy based on semantic policy tools. IEEE Secur. Priv. **8**(4), 25–30 (2010)
Karwowski, W., Mulholland, N.O., Ward, T.L., Jagannathan, V.: A fuzzy knowledge base of an expert system for analysis of manual lifting tasks. Fuzzy Sets Syst. **21**(3), 363–374 (1987)
Katarya, R., Jain, I., Hasija, H.: An interactive interface for instilling trust and providing diverse recommendations. In: International Conference on Computer and Communication Technology (ICCCT), pp. 17–22 (2014)
Keeney, R.L., Raiffa, H.: Decisions with Multiple Objectives: Preferences and Value Tradeoffs. Wiley Series in Probability and Mathematical Statistics. Wiley, Hoboken (1976)
Kim, B.O., Lee, S.M.: A bond rating expert system for industrial companies. Expert Syst. Appl. **9**(1), 63–70 (1995)
Kim, S.K., Park, J.I.: A structural equation modeling approach to generate explanations for induced rules. Expert Syst. Appl. **10**(3), 403–416 (1996)
Kitchenham, B., Charters, S.: Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE-2007-01, School of Computer Science and Mathematics, Keele University (2007)
Kitchenham, B., Brereton, P.: A systematic review of systematic review process research in software engineering. Inf. Softw. Technol. **55**(12), 2049–2075 (2013)
Klein, D.A., Shortliffe, E.H.: A framework for explaining decision-theoretic advice. Artif. Intell. **67**(2), 201–243 (1994)
Koussev, T., Weiss, M.P., Reiss, K.: A graphic explanation environment for expert systems. In: Second International Conference on Software Engineering for Real Time Systems, pp. 11–15 (1989)
Labreuche, C.: A general framework for explaining the results of a multi-attribute preference model. Artif. Intell. **175**(7), 1410–1448 (2011)
Lacave, C., Díez, F.J.: A review of explanation methods for bayesian networks. Knowl. Eng. Rev. **17**(2), 107–127 (2002)
Lacave, C., Díez, F.J.: A review of explanation methods for heuristic expert systems. Knowl. Eng. Rev. **19**(2), 133–146 (2004)
Lacave, C., Oniśko, A., Díez, F.J.: Use of Elvira's explanation facility for debugging probabilistic expert systems. Knowl. Based Syst. **19**(8), 730–738 (2006)
Lambert, S.C., Ringland, G.A.: Knowledge representations and interfaces in financial expert systems. In: UK IT 1990 Conference, pp. 434–441 (1990)
Langlotz, C.P., Shortliffe, E.H.: Adapting a consultation system to critique user plans. Int. J. Man Mach. Stud. **19**(5), 479–496 (1983)
Lee, H.M., Hsu, C.C.: Building expert systems by training with automatic neural network generating ability. In: Proceedings Eighth Conference on Artificial Intelligence for Applications, pp. 197–203 (1992)
Levy, M., Ferrand, P., Chirat, V.: SESAM-DIABETE, an expert system for insulin-requiring diabetic patient education. Comput. Biomed. Res. **22**(5), 442–453 (1989)
Li, M., Gregor, S.: Outcomes of effective explanations: empowering citizens through online advice. Decis. Support Syst. **52**(1), 119–132 (2011)

Libório, A., Furtado, E., Rocha, I., Furtado, V.: Interface design through knowledge-based systems: an approach centered on explanations from problem-solving models. In: Proceedings of the 4th International Workshop on Task Models and Diagrams, TAMODIA '05, pp. 127–134 (2005)
Lieberman, H., van Dyke, N., Vivacqua, A.: Let's browse: a collaborative browsing agent. Knowl. Based Syst. **12**(8), 427–431 (1999)
Liu, K.F.R., Lee, J., Chiang, W., Yang, S.J.: Fpnes: fuzzy Petri net based expert system for bridges damage assessment. In: Proceedings Tenth IEEE International Conference on Tools with Artificial Intelligence, pp. 302–309 (1998)
Lopez-Suarez, A., Kamel, M.: Dykor: a method for generating the content of explanations in knowledge systems. Knowl. Based Syst. **7**(3), 177–188 (1994)
Machado, R.J., da Rocha, A.F.: Inference, inquiry and explanation in expert systems by means of fuzzy neural networks. In: Proceedings of the Second IEEE International Conference on Fuzzy Systems, vol. 1, pp. 351–356 (1993)
Mahmoud, M., Algadi, N., Ali, A.: Expert system for banking credit decision. In: 2008 International Conference on Computer Science and Information Technology, pp. 813–819 (2008)
Malheiro, N., Vale, Z.A., Ramos, C., Santos, J., Marques, A.: Enabling Client-Server Explanation Facilities in a Real-Time Expert System, pp. 333–342. Springer, Berlin (1999)
Mao, J.Y., Benbasat, I.: The effects of contextualized access to knowledge on judgement. Int. J. Hum. Comput. Stud. **55**(5), 787–814 (2001)
Martincic, C.J.: QUE: an expert system explanation facility that answers "why not" types of questions. J. Comput. Sci. Coll. **19**(1), 336–348 (2003)
Marx, P., Hennig-Thurau, T., Marchand, A.: Increasing consumers' understanding of recommender results: a preference-based hybrid algorithm with strong explanatory power. In: Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys '10, pp. 297–300 (2010)
Matelli, J.A., Bazzo, E., da Silva, J.C.: An expert system prototype for designing natural gas cogeneration plants. Expert Syst. Appl. **36**(4), 8375–8384 (2009)
Matsatsinis, N., Doumpos, M., Zopounidis, C.: Knowledge acquisition and representation for expert systems in the field of financial analysis. Expert Syst. Appl. **12**(2), 247–262 (1997)
Maybury, M.T.: Enhancing explanation coherence with rhetorical strategies. In: Proceedings of the Fourth Conference on European Chapter of the Association for Computational Linguistics, EACL '89, pp. 168–173 (1989)
McCarthy, K., Reilly, J., McGinty, L., Smyth, B.: Experiments in dynamic critiquing. In: Proceedings of the 10th International Conference on Intelligent User Interfaces, IUI '05, pp. 175–182. ACM (2005)
Mcsherry, D.: Explanation in recommender systems. Artif. Intell. Rev. **24**(2), 179–197 (2005)
Mejia-Lavalle, M.: Outlier detection with innovative explanation facility over a very large financial database. In: 2010 IEEE Electronics, Robotics and Automotive Mechanics Conference, pp. 23–27 (2010)
Mendes, D., Rodrigues, I.P., Baeta, C.: Ontology based clinical practice justification in natural language. Procedia Technol. **9**, 1288–1293 (2013)
Metzler, D.P., Martincic, C.J.: QUE: explanation through exploration. Expert Syst. Appl. **15**(34), 253–263 (1998)
Mitra, S.: Fuzzy mlp based expert system for medical diagnosis. Fuzzy Sets Syst. **65**(2), 285–296 (1994)
Mitra, S., Pal, S.K.: Fuzzy multi-layer perceptron, inferencing and rule generation. IEEE Trans. Neural Netw. **6**(1), 51–63 (1995)
Mller-Kolck, U.: Expert system support for the therapeutic management of cerebrovascular disease. Artif. Intell. Med. **2**(1), 35–42 (1990)
Mocanu, A.: Envisioning a collaborative smart home solution based on argumentative dialogues. In: Proceedings of the 7th Balkan Conference on Informatics Conference, BCI '15, pp. 23:1–23:6 (2015)
Moulin, B., Irandoust, H., Bélanger, M., Desbordes, G.: Explanation and argumentation capabilities: towards the creation of more persuasive agents. Artif. Intell. Rev. **17**(3), 169–222 (2002)
Muhammad, K., Lawlor, A., Rafter, R., Smyth, B.: Great Explanations: Opinionated Explanations for Recommendations, pp. 244–258. Springer, Berlin (2015)
Muhammad, K.I., Lawlor, A., Smyth, B.: A live-user study of opinionated explanations for recommender systems. In: Proceedings of the 21st International Conference on Intelligent User Interfaces, IUI '16, pp. 256–260 (2016)
Murphy, D.S., Phillips, M.E.: The effects of expert system use on entry-level accounting expertise: an experiment. Expert Syst. Appl. **3**(1), 129–134 (1991)
Nakatsu, R.T.: Explanatory Power of Intelligent Systems, pp. 123–143. Springer, London (2006)

Nakatsu, R.T., Benbasat, I.: Improving the explanatory power of knowledge-based systems: an investigation of content and interface-based enhancements. Trans. Syst. Man Cybern. Part A **33**(3), 344–357 (2003)
Narayanan, T., McGuinness, D.L.: Towards leveraging inference web to support intuitive explanations in recommender systems for automated career counseling. In: First International Conference on Advances in Computer–Human Interaction, pp. 164–169 (2008)
Nart, D.D., Tasso, C.: A personalized concept-driven recommender system for scientific libraries. Procedia Comput. Sci. **38**, 84–91 (2014)
Ng, G., Ong, K.: Using a qualitative probabilistic network to explain diagnostic reasoning in an expert system for chest pain diagnosis. Comput. Cardiol. **2000**(27), 569–572 (2000)
Nilashi, M., Jannach, D., bin Ibrahim, O., Esfahani, M.D., Ahmadi, H.: Recommendation quality, transparency, and website quality for trust-building in recommendation agents. Electron. Commer. Res. Appl. **19**, 70–84 (2016)
Norton, S.W.: An explanation mechanism for bayesian inferencing systems. In: Lemmer, J.F., Kanal, L.N. (eds.) Uncertainty in Artificial Intelligence, Machine Intelligence and Pattern Recognition, vol. 5, pp. 165–173. North-Holland, Amsterdam (1988)
Nunes, I., Miles, S., Luck, M., de Lucena, C.J.P.: Investigating explanations to justify choice. In: Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization, UMAP'12, pp. 212–224 (2012a)
Nunes, I., Chen, Y., Miles, S., Luck, M., Lucena, C.: Transparent Provenance Derivation for User Decisions, pp. 111–125. Springer, Berlin (2012b)
Nunes, I., Miles, S., Luck, M., Barbosa, S., Lucena, C.: Pattern-based explanation for automated decisions. In: Proceedings of the Twenty-first European Conference on Artificial Intelligence, ECAI'14, pp. 669–674 (2014)
Nuthall, P., Bishop-Hurley, G.: Expert systems for animal feeding management part i: presentation aspects. Comput. Electron. Agric. **14**(1), 9–22 (1996)
O'Donovan, J., Gretarsson, B., Bostandjiev, S., Hollerer, T., Smyth, B.: A visual interface for social information filtering. In: 2009 International Conference on Computational Science and Engineering, vol. 4, pp. 74–81 (2009)
Omran, A.M., Khorshid, M.: Intelligent environmental scanning approach (a case study: the Egyptian wheat crop production). IERI Procedia **7**, 28–34 (2014a)
Omran, A.M., Khorshid, M.: An intelligent recommender system for long view of Egypt's livestock production. AASRI Procedia **6**, 103–110 (2014b)
Oramas, S., Espinosa-Anke, L., Sordo, M., Saggion, H., Serra, X.: Information extraction for knowledge base construction in the music domain. Data Knowl. Eng. **106**, 70–83 (2016)
Overby, M.A.: Psyxpert: an expert system prototype for aiding psychiatrists in the diagnosis of psychotic disorders. Comput. Biol. Med. **17**(6), 383–393 (1987)
Pal, K.: An approach to legal reasoning based on a hybrid decision-support system. Expert Syst. Appl. **17**(1), 1–12 (1999)
Pal, K., Palmer, O.: A decision-support system for business acquisitions. Decis. Support Syst. **27**(4), 411–429 (2000)
Papamichail, K., French, S.: Explaining and justifying the advice of a decision support system: a natural language generation approach. Expert Syst. Appl. **24**(1), 35–48 (2003)
Papadimitriou, A., Symeonidis, P., Manolopoulos, Y.: A generalized taxonomy of explanations styles for traditional and social recommender systems. Data Min. Knowl. Discov. **24**(3), 555–583 (2012)
Pazzani, M., Iyer, R., See, D., Schroeder, E., Tilles, J.: CTSHIV: a knowledge-based system for the management of HIV-infected patients. In: Intelligent Information Systems, 1997. IIS '97, pp. 7–13 (1997)
Perlin, M., Kanal, E., John, A.: A user interface for visualizing concepts in magnetic resonance imaging. In: Proceedings of the First Conference on Visualization in Biomedical Computing, pp. 260–267 (1990)
Popchev, I.P., Zlatareva, N.P., Sinapova, L.J.: EDDY: an expert system in dysmorphology based on truth-maintenance. In: Images of the Twenty-First Century. Proceedings of the Annual International Engineering in Medicine and Biology Society, pp. 1877–1878 (1989)
Pu, P., Chen, L.: Trust-inspiring explanation interfaces for recommender systems. Knowl. Based Syst. **20**(6), 542–556 (2007)
Rahwan, I., Simari, G.R.: Argumentation in Artificial Intelligence, 1st edn. Springer, Berlin (2009)
Ramberg, R.: Construing and testing explanations in a complex domain. Comput. Hum. Behav. **12**(1), 29–48 (1996)
Ray, A.K.: Equipment fault diagnosisa neural network approach. Comput. Ind. **16**(2), 169–177 (1991)

Reggia, J.A., Perricone, B.T., Nau, D.S., Peng, Y.: Answer justification in diagnostic expert systems—part I: abductive inference and its justification. IEEE Trans. Biomed. Eng. **BME–32**(4), 263–267 (1985)
Reilly, J., McCarthy, K., McGinty, L., Smyth, B.: Explaining compound critiques. Artif. Intell. Rev. **24**(2), 199–220 (2005)
Reyes, A., Ibarguengoytia, P.H., Elizalde, F., Snchez, L., Nava, A.: ASISTO: an integrated intelligent assistant system for power plant operation and training. In: 16th International Conference on Intelligent System Applications to Power Systems, pp. 1–6 (2011)
Richards, D.: The reuse of knowledge: a user-centred approach. Int. J. Hum. Comput. Stud. **52**(3), 553–579 (2000)
Ringer, M.J., Quinn, T.M., Merolla, A.: Autonomous power system: intelligent diagnosis and control. Telemat. Inform. **8**(4), 365–383 (1991)
Riordan, D., Carden, K.J.: Explanation in ecological systems. In: Proceedings of the 1990 ACM SIGSMALL/PC Symposium on Small Systems, SIGSMALL '90, pp. 249–254 (1990)
Roitman, H., Messika, Y., Tsimerman, Y., Maman, Y.: Increasing patient safety using explanation-driven personalized content recommendation. In: Proceedings of the 1st ACM International Health Informatics Symposium, IHI '10, pp. 430–434 (2010)
Rook, F.W., Donnell, M.L.: Human cognition and the expert system interface: mental models and inference explanations. IEEE Trans. Syst. Man Cybern. **23**(6), 1649–1661 (1993)
Samarasinghe, S.: Neural Networks for Applied Sciences and Engineering. Auerbach Publications, Boston (2006)
Santoso, N.I., Darken, C., Povh, G., Erdmann, J.: Nuclear plant fault diagnosis using probabilistic reasoning. In: Proceedings of the 1999 IEEE Power Engineering Society Summer Meeting, vol. 2, pp. 714–719 (1999)
Sarkar, A., Bandyopadhyay, S., Jullien, G.A.: Bit-level designer's assistant-a knowledge based approach to systolic processor design. In: Proceedings of the 33rd Midwest Symposium on Circuits and Systems, pp. 1001–1004 (1990)
Saunders, V.M., Dobbs, V.S.: Explanation generation in expert systems. In: IEEE Conference on Aerospace and Electronics, pp. 1101–1106 (1990)
Schaffer, J., Giridhar, P., Jones, D., Höllerer, T., Abdelzaher, T., O'Donovan, J.: Getting the message? A study of explanation interfaces for microblog data analysis. In: Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI '15, pp. 345–356 (2015)
Scheel, C., Castellanos, A., Lee, T., De Luca, E.W.: The Reason Why: A Survey of Explanations for Recommender Systems, pp. 67–84. Springer, Berlin (2014)
Schröder, O., Möbus, C., Folckers, J., Thole, H.J.: Supporting the construction of explanation models and diagnostic reasoning in probabilistic domains. In: Proceedings of the 1996 International Conference on Learning Sciences, ICLS '96, pp. 60–67 (1996)
Shaalan, K., Rafea, M., Rafea, A.: KROL: a knowledge representation object language on top of Prolog. Expert Syst. Appl. **15**(1), 33–46 (1998)
Sharma, A., Cosley, D.: Do social explanations work? Studying and modeling the effects of social explanations in recommender systems. In: Proceedings of the 22nd International Conference on World Wide Web, WWW '13, pp. 1133–1144 (2013)
Sherchan, W., Loke, S.W., Krishnaswamy, S.: Explanation-aware service selection: rationale and reputation. Serv. Oriented Comput. Appl. **2**(4), 203–218 (2008)
Shoval, P.: Principles, procedures and rules in an expert system for information retrieval. Inf. Process. Manag. **21**(6), 475–487 (1985)
Slagle, J.R.: Applications of a generalized network-based expert system shell-artificial intelligence mini-tutorial. In: Proceedings of the Symposium on the Engineering of Computer-Based Medical, pp. 33–42 (1988)
Slotnick, S.A., Moore, J.D.: Explaining quantitative systems to uninitiated users. Expert Syst. Appl. **8**(4), 475–490 (1995)
Song, W., Shi, H., Li, Q.: Study of an explanation mechanism in expert system based on fault tree for safety risk assessment. In: 2nd International Conference on Future Computer and Communication, vol. 2, pp. V2–479–V2–483 (2010)
Sørmo, F., Cassens, J., Aamodt, A.: Explanation in case-based reasoning-perspectives and goals. Artif. Intell. Rev. **24**(2), 109–143 (2005)

Srivastava, R.P.: Automating judgmental decisions using neural networks: a model for processing business loan applications. In: Proceedings of the 1992 ACM Annual Conference on Communications, CSC '92, pp. 351–357 (1992)
Strachan, S.M., McArthur, S.D.J., Judd, M.D., McDonald, J.R.: Incremental knowledge-based partial discharge diagnosis in oil-filled power transformers. In: Proceedings of the 13th International Conference on Intelligent Systems Application to Power Systems (2005)
Strat, T.M., Lowrance, J.D.: Explaining evidential analyses. Int. J. Approx. Reason. **3**(4), 299–353 (1989)
Štrumbelj, E., Kononenko, I., Šikonja, M.R.: Explaining instance classifications with interactions of subsets of feature values. Data Knowl. Eng. **68**(10), 886–904 (2009)
Suermondt, H.J., Cooper, G.F.: An evaluation of explanations of probabilistic inference. Comput. Biomed. Res. **26**(3), 242–254 (1993)
Swartout, W.R., Moore, J.D.: Explanation in Second Generation Expert Systems, pp. 543–585. Springer, Berlin (1993)
Swinney, L.: The explanation facility and the explanation effect. Expert Syst. Appl. **9**(4), 557–567 (1995)
Symeonidis, P., Nanopoulos, A., Manolopoulos, Y.: Providing justifications in recommender systems. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **38**(6), 1262–1272 (2008)
Tan, W.K., Tan, C.H., Teo, H.H.: Consumer-based decision aid that explains which to buy: decision confirmation or overconfidence bias? Decis. Support Syst. **53**(1), 127–141 (2012)
Tanner, M.C., Keuneke, A.M.: Explanations in knowledge systems: the roles of the task structure and domain functional models. IEEE Expert **6**(3), 50–57 (1991)
Terano, T., Suzuki, M., Onoda, T., Uenishi, K., Matsuura, T.: CSES: an approach to integrating graphic, music and voice information into a user-friendly interface. In: International Workshop on Industrial Applications of Machine Intelligence and Vision, pp. 349–354 (1989)
Thirumuruganathan, S., Huber, M.: Building bayesian network based expert systems from rules. In: 2011 IEEE International Conference on Systems, Man, and Cybernetics, pp. 3002–3008 (2011)
Tintarev, N., Masthoff, J.: Effective explanations of recommendations: user-centered design. In: Proceedings of the 2007 ACM Conference on Recommender Systems, RecSys '07, pp. 153–156 (2007a)
Tintarev, N., Masthoff, J.: A survey of explanations in recommender systems. In: Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop, pp. 801–810 (2007b)
Tintarev, N., Masthoff, J.: Designing and evaluating explanations for recommender systems. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Recommender Systems Handbook, pp. 479–510. Springer, Berlin (2011)
Tintarev, N., Masthoff, J.: Evaluating the effectiveness of explanations for recommender systems. User Model. User Adapt. Interact. **22**(4–5), 399–439 (2012)
Tjahjadi, T., Bowen, D., Bevan, J.R.: 3M: a user modelling interface of an expert system for x-ray topographic image interpretation. Interact. Comput. **2**(3), 259–278 (1990)
Tong, L.C.: An explanation facility for a grammar writing system. In: Proceedings of the 13th Conference on Computational Linguistics, COLING '90, pp. 359–364 (1990)
Tong, X., Ang, J.: Explaining control strategies in second generation expert systems. IEEE Trans. Syst. Man Cybern. **25**(11), 1483–1490 (1995)
Toulmin, S.E.: The Uses of Argument. Cambridge University Press, Cambridge (2003)
Tzafestas, S., Konstantinidis, N.: ENGEXP—an integrated environment for the development and application of expert systems in equipment and engine fault diagnosis and repair. Adv. Eng. Softw. **14**(1), 3–14 (1992)
van Aarle, E., van den Bercken, J.: The development of a knowledge-based system supporting the diagnosis of reading and spelling problems. Comput. Hum. Behav. **8**(23), 183–201 (1992)
Vashisth, P., Chandoliya, D., Yadav, B.K., Bedi, P.: Trust enabled argumentation based recommender system. In: 12th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 137–142 (2012)
Vig, J., Sen, S., Riedl, J.: Tagsplanations: explaining recommendations using tags. In: Proceedings of the 14th International Conference on Intelligent User Interfaces, IUI '09, pp. 47–56 (2009)
Vogiatzis, D., Karkaletsis, V.: A cognitive framework for robot guides in art collections. Univers. Access Inf. Soc. **10**(2), 179–193 (2011)
Wall, R., Cunningham, P., Walsh, P., Byrne, S.: Explaining the output of ensembles in medical decision support on a case by case basis. Artif. Intell. Med. **28**(2), 191–206 (2003)
Wang, L., Libert, G., Liu, B.: An expert system for forecasting model selection. In: Proceedings of the First IEEE Conference on Control Applications, pp. 704–709 (1992)

Wang, N., Pynadath, D.V., Hill, S.G.: The impact of pomdp-generated explanations on trust and performance in human-robot teams. In: Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, AAMAS '16, pp. 997–1005 (2016a)
Wang, N., Pynadath, D.V., Hill, S.G.: Trust calibration within a human-robot team: comparing automatically generated explanations. In: Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 109–116 (2016b)
Wang, W., Qiu, L., Kim, D., Benbasat, I.: Effects of rational and social appeals of online recommendation agents on cognition- and affect-based trust. Decis. Support Syst. **86**(C), 48–60 (2016c)
Washington, E.S., Ali, M.: PISCES: an expert system for coal fired power plant monitoring and diagnostics. In: Proceedings of the 1st International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, IEA/AIE '88, pp. 87–93 (1988)
Wick, M.R., Slagle, J.R.: An explanation facility for today's expert systems. IEEE Expert **4**(1), 26–36 (1989a)
Wick, M.R., Slagle, J.R.: The partitioned support network for expert system justification. IEEE Trans. Syst. Man Cybern. **19**(3), 528–535 (1989b)
Widyantoro, D.H., Baizal, Z.K.A.: A framework of conversational recommender system based on user functional requirements. In: 2nd International Conference on Information and Communication Technology (ICoICT), pp. 160–165 (2014)
Wong, K.P., Cheung, H.N.: Expert system for protection current transformer design specification preparation. IEE Proc. C Gener. Transm. Distrib. **136**(6), 391–400 (1989)
Yasdi, R.: Design of the exis's explanation component. Comput. Ind. **13**(1), 15–21 (1989)
Ye, L.R.: The value of explanation in expert systems for auditing: an experimental investigation. Expert Syst. Appl. **9**(4), 543–556 (1995)
Ye, L.R., Johnson, P.E.: The impact of explanation facilities on user acceptance of expert systems advice. MIS Q. **19**, 157–172 (1995)
Yen, J.: Gertis: a dempster-shafer approach to diagnosing hierarchical hypotheses. Commun. ACM **32**(5), 573–585 (1989)
Yoon, Y., Guimaraes, T., Swales, G.: Integrating artificial neural networks with rule-based expert systems. Decis. Support Syst. **11**(5), 497–507 (1994)
Yu, C., Lakshmanan, L., Amer-Yahia, S.: It takes variety to make a world: diversification in recommender systems. In: Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, EDBT '09, pp. 368–378 (2009)
Zain, M.F.M., Islam, M.N., Basri, I.H.: An expert system for mix design of high performance concrete. Adv. Eng. Softw. **36**(5), 325–337 (2005)
Zanker, M.: The influence of knowledgeable explanations on users' perception of a recommender system. In: Proceedings of the Sixth ACM Conference on Recommender Systems, RecSys '12, pp. 269–272 (2012)
Zanker, M., Ninaus, D.: Knowledgeable explanations for recommender systems. In: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 657–660 (2010)
Zeleznikow, J., Stranieri, A., Gawler, M.: Project report: split-up—a legal expert system which determines property division upon divorce. Artif. Intell. Law **3**(4), 267–275 (1995)
Zhang, Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y., Ma, S.: Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '14, pp. 83–92 (2014)

Dr. Ingrid Nunes is a Senior Lecturer at the Institute of Informatics, Universidade Federal do Rio Grande do Sul (UFRGS), Brazil, currently in a sabbatical year at TU Dortmund in Germany. She obtained her Ph.D. in Informatics at the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil. Her Ph.D. was in cooperation with King's College London (UK) and University of Waterloo (Canada). She is the head of the Prosoft research group, and her main research areas are decision making and multi-agent systems.

Dr. Dietmar Jannach is a Professor of Computer Science at TU Dortmund, Germany and head of the department's e-services research group. Dr. Jannach has worked on different areas of artificial intelligence, including recommender systems, model-based diagnosis, and knowledge-based systems. He is the leading author of a textbook on recommender systems and has authored more than hundred technical papers, focusing on the application of artificial intelligence technology to practical problems.