IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 16, 2020, accepted January 3, 2021, date of publication January 13, 2021, date of current version January 22, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3051315

A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence

ILIA STEPIN[1], JOSE M. ALONSO[1], (Member, IEEE), ALEJANDRO CATALA[1], AND MARTÍN PEREIRA-FARIÑA[2]

[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain
[2]Departamento de Filosofía e Antropoloxía, Universidade de Santiago de Compostela, 15705 Santiago de Compostela, Spain

Corresponding author: Ilia Stepin (ilia.stepin@usc.es)

This work was supported in part by the Spanish Ministry of Science, Innovation and Universities under Grant RTI2018-099646-B-I00 and Grant RED2018-102641-T, in part by the Galician Ministry of Education, University and Professional Training under Grant ED431F 2018/02, Grant ED431C 2018/29, Grant ED431G/08, and Grant ED431G2019/04; and in part by the European Regional Development Fund (ERDF/FEDER Program).

ABSTRACT A number of algorithms in the field of artificial intelligence offer poorly interpretable decisions. To disclose the reasoning behind such algorithms, their output can be explained by means of so-called evidence-based (or factual) explanations. Alternatively, contrastive and counterfactual explanations justify why the output of the algorithms is not any different and how it could be changed, respectively. It is of crucial importance to bridge the gap between theoretical approaches to contrastive and counterfactual explanation and the corresponding computational frameworks. In this work we conduct a systematic literature review which provides readers with a thorough and reproducible analysis of the interdisciplinary research field under study. We first examine theoretical foundations of contrastive and counterfactual accounts of explanation. Then, we report the state-of-the-art computational frameworks for contrastive and counterfactual explanation generation. In addition, we analyze how grounded such frameworks are on the insights from the inspected theoretical approaches. As a result, we highlight a variety of properties of the approaches under study and reveal a number of shortcomings thereof. Moreover, we define a taxonomy regarding both theoretical and practical approaches to contrastive and counterfactual explanation.

INDEX TERMS Computational intelligence, contrastive explanations, counterfactuals, explainable artificial intelligence, systematic literature review.

I. INTRODUCTION

In the last few decades, the field of Artificial Intelligence (AI) has witnessed major changes. As available computational resources have grown significantly, AI algorithms are attracting a significant amount of attention in industry and research [1]. While a great number of such algorithms present strikingly accurate decisions, their decision-making apparatus is frequently left unclear to users of such applications. In particular, a number of Machine Learning (ML)-based algorithms are often perceived as ''black-box'' algorithms because they are overloaded with millions of hardly interpretable parameters to be optimized at the training stage. This fact makes the algorithm's output hard to explain. A lack of the ability to explain such automatic decisions undermines users' trust and hence decreases usability of such systems [2]. Furthermore, it prevents users from a responsible exploitation of their decisions [3]. In addition, many of the existing eXplainable AI (XAI[1]) methods provide summaries of automatically made predictions rather than true explanations [4]. As a result, the need to motivate automatic decisions with a clear explanation of why the algorithm outputs a particular decision has made the XAI research field grow quickly [5].

Since the number of high-stakes AI applications found in daily life increases, the requirements to their explana-

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

[1]XAI stands for eXplainable Artificial Intelligence. This acronym was made popular by the USA Defense Advanced Research Projects Agency when launching to the research community the challenge of designing self-explanatory AI systems (https://www.darpa.mil/program/explainable-artificial-intelligence).

tory capacity increase accordingly. This also provokes the introduction of regulations and laws concerned with explanation requirements for AI-based applications. For instance, the need for explaining reasoning mechanisms behind such applications is now legally regulated in the European Union by means of the General Data Protection Regulation.[2] According to these legal provisions, the data subject must be provided with ''meaningful information about the logic involved'' in the automatic decision making process, which is commonly referred to as the ''right to explanation'' [6]. Thus, an AI application is expected not only to provide accurate decisions but also to justify them in a comprehensive manner to end-users.

The goal of approaching human-centric AI has led towards a deeper research on the nature of explanation. However, no agreement about a definition of explanation has been reached despite the fact that explanation has called a significant amount of attention in, e.g., philosophy of science [7], [8]. In its most general form, explanation is normally treated as ''an answer to the question of why something is the case'' [9]. In the context of AI, it often bases on judgments about why a certain outcome is predicted by an AI algorithm and hypotheses about causes with respect to given effects [10].

The need of generating more human-like explanations has attracted AI researchers' attention to particular properties of explanation as well as its sub-types [11]. Thus, it appears particularly challenging to explain a given algorithm's output in terms of reasonable yet non-occurring alternatives given a possibly infinite set of such options. Furthermore, this can be enhanced with the ability of suggesting relevant changes in the input so that the algorithm outputs a different decision.

Given a rising interest towards these types of explanation (referred to as contrastive and counterfactual, respectively) within the XAI community, it is of crucial importance to review the existing theoretical accounts of contrastive and counterfactual explanation as well as state-of-the-art computational frameworks for automatic generation thereof. Thus, the aim of this study is to fulfill the next three objectives: (1) to scrutinize theoretical works on the contrastive and counterfactual accounts of explanation; (2) to summarize state-of-the-art methods in the field of automatic explanation generation thereof; and (3) to discuss a degree of synergy between the revised theories and their related up-to-date implementations.

The rest of the manuscript is organized as follows. Section II introduces the notions of contrastive and counterfactual explanation as well as their main application areas. Section III presents the terminology used throughout the review, poses the research questions, and describes the methodology employed to address the given questions. Section IV presents the main findings collected within the present survey and the emerging taxonomy thereof. Section V discusses peculiarities of the existing theoretical and computational frameworks of contrastive and counterfactual explanation. Finally, we conclude in Section VI.

[2]https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02016R0679-20160504

II. BACKGROUND

A. CONTRASTIVE EXPLANATION

Findings on explanation accumulated in humanities and social sciences show that it is intrinsically contrastive [11]. The property of contrastiveness presupposes that an explanation answers the given why-question regarding the cause of the event in question (''Why did P happen?'') in terms of hypothesized non-occurring alternatives (''Why did P happen rather than Q?'') [12]. Thus, supporters of the pragmatic approach to explanation argue that it is exactly the ability to distinguish the answer to an explanatory question from a set of contrastive hypothesized alternatives that provides the explainee with sufficiently comprehensive information on the reasoning behind the question [13]. This approach is also claimed to set a minimum criterion that an explanation must fulfill: it must favor the probability of the observed event P to all the hypothetical alternatives $(Q_1, Q_2, \ldots, Q_n)$ [14].

Contrastive explanation is among influential topics in cognitive science [15]–[17]. Thus, contrastive explanations are claimed to be inherent to human cognition [16]. Indeed, we are used to question those decisions that we once made, especially if such decisions or coinciding circumstances resulted in tragic events [18].

In addition, contrastive reasoning forms the basis of abductive inference [19], i.e., the process of inferring certain facts that render some observation plausible [20]. In other words, a given observation can be explained on the basis of the most likely among a pool of competing hypotheses [21].

B. COUNTERFACTUAL EXPLANATION

Given the property of contrastiveness, it is possible to imagine explanatory alternatives to how things would stand if a different decision had been made at some point. They can serve to explain potential consequences of such contrastive non-taken alternative decisions. In this case, the mind is assumed to construct and compare mental representations of an actually happened event and that of some event alternative to it [22]. Cognitive scientists refer to such mental representations of alternatives to past events as counterfactuals (''contrary-to-fact'') [15]. The process of ''thinking about past possibilities and past or present impossibilities'' is therefore called counterfactual thinking [23]. Alternatively, the combination of imagining an alternative scenario in relation to the one that actually happened and the exploration of its consequences is referred to as counterfactual reasoning [24]. In addition, counterfactual reasoning is claimed to be a key mechanism for explaining adaptive behavior in a changing environment [25], [26].

Counterfactuals describe events or states of the world that did not occur and implicitly or explicitly contradict factual world knowledge [27]. Formulated in natural language, counterfactuals are usually presented in the form of conditional

statements. Broadly speaking, they contain: (1) an antecedent describing an outcome alternative to an actual event; (2) a consequent describing (a set of) consequences, had the antecedent been the case; and (3) a binary counterfactual dependency relation between them. Thus, Grahne defines a counterfactual to be a conditional statement where the antecedent ''can contradict the current state of affairs, or our current knowledge thereof'' [28]. However, despite a general agreement on structural properties of counterfactuals, existing interpretations of counterfactual conditionals still compete. As such, further constraints imposed on their structure differ depending on the approach adopted. According to Ginsberg [29], a counterfactual is a conditional statement of the form ''If P, then Q'' where P is ''expected to be false''. Aumann limits a counterfactual to be a conditional with a false antecedent only [30]. In contrast, Spohn argues that both the antecedent and the consequent of a counterfactual must be false [31]. All in all, counterfactual conditional statements are claimed to enable people to produce utterances that are factually false yet truthful irrespective of the interpretation adopted [32].

A line of research devoted to modeling human counterfactual reasoning has been thoroughly investigated in computer science. Thus, counterfactual reasoning in computer science is defined as the process of evaluating conditional claims about alternative possibilities and their consequences [33]. It is argued to be valid arising from antecedents that are true in a hypothetical model but false in reality [34]. In this setting, the truth of a counterfactually inferred statement is resolved by: (1) modeling a situation where the smallest possible change in features of the actual world (as set in the antecedent) leads to a different (possibly, desired) state of things (the so-called ''closest'' or ''nearest'' possible world); and (2) estimating what is true in that setting [35].

Moreover, counterfactuality is among the most fundamental concepts in theories of causation [36], [37]. Indeed, counterfactuals are argued to represent a causal relation between the event happened in reality and its imaginary counterpart. A counterfactual definition of a cause of an arbitrary event traces back to Hume [38]. According to him, a cause is an object (antecedent) that justifies the existence of another object (consequent) which it is followed by: ''If the first object had not been, then the second never had existed''. Therefore, once a causal connection between the antecedent and the consequent is established, a counterfactual conditional can be generalized to be a conditional claim about an alternate possibility and its consequences of the form ''If X were to occur, then Y would (or might) occur'' [33]. Similarly, Kment applies a similarity-based approach between possible worlds to formulate a general account of counterfactuals [39] driven by a non-epistemic interpretation of explanation (i.e., factors that serve as reasons for some fact to obtain are responsible for that fact).

The conditional structure of counterfactual statements gave rise to a probabilistic account of such statements. Thus, Pearl extended the definition of the causal counterfactual to estimate the probability of the truth of the consequent caused by the antecedent (''a probability statement about the truth of y, had x been true, when it is known that y had been false when x was false'') [37]. This approach to counterfactuals motivated a number of experiments on the existence of the relation between counterfactuals and conditional probability. In support of this assumption, Over *et al.* [40] showed the existence of connection between counterfactuals and conditional probability, as they experimented with probability judgments about counterfactuals. Thus, they proposed that the subjective probability of the counterfactual at the present time is the same as the conditional probability $P(y|x)$ at some earlier time. Twenty-six subjects were asked to estimate the probability of truth of thirty-two counterfactual conditionals with both affirmative and negative antecedents and consequents. Their findings point to a strong correlation between the probability of the counterfactual conditional and causal strength judgments. On a similar note, Edgington regarded counterfactual judgments as uncertain conditional statements and therefore evaluated them by estimating their conditional probability given some endorsing event [41].

C. DISTINCTION BETWEEN CONTRASTIVE AND COUNTERFACTUAL EXPLANATION

It is important to note that some researchers tend to either collapse or intentionally distinguish contrastive reasoning from counterfactual reasoning despite their conceptual similarity. For instance, Lombrozo treated counterfactual and contrastive explanations as equivalent assuming hypothesized events non-occurred in reality to be ''counterfactual cases'' where a subset of these cases forms a contrastive explanation [10]. In contrast, McGill and Klein distinguished contrastive reasoning from its counterfactual counterpart [42]. According to them, contrastive reasoning is concerned with situations where different target situations are analyzed (''What made the difference between the employee who failed and the employees who did not fail?''). On the other hand, counterfactual reasoning is claimed to deal with cases where the antecedent is altered to account for changes in the outcome (''Would the employee have failed had she not been a woman?''). Alternatively, Fang *et al.* [43] referred to contrastive reasoning as a procedure operating on ''but-statements'', as in ''all cars are polluting, but hybrid cars are not polluting'', which serves a principally different explanation generation task in comparison with the other aforementioned approaches.

D. CONTRASTIVE AND COUNTERFACTUAL EXPLANATION IN THE CONTEXT OF XAI

The stochastic nature of predictions made by various AI algorithms is claimed to be among the main obstacles in reaching a true explanation [44]. Research on automatic contrastive and counterfactual explanation generation shows a number of considerable observations that help overcome this issue. Thus, empirical studies prove that incorporating contrastiveness improves the quality of explanations offered

to the end-user [45]. Furthermore, contrastive explanations can be used to personalize human-machine interaction when a user is engaged in an explanatory dialogue with an AI application. Thus, they can be employed with the aim of adjusting the contents of the explanation for the algorithm's output in accordance with the user's preferences [46]. Finally, the ability to explain a decision contrastively is claimed to lead to responsible decision-making [47].

It is important to note that contrastive explanations point to the difference between the actual and a hypothetical decision. On the other hand, counterfactual explanations specify necessary minimal changes in the input so that a contrastive output is obtained. However, these terms are sometimes used interchangeably in the context of XAI [48], [49].

Various families of techniques have been proposed to generate contrastive and counterfactual explanations of AI algorithm output. In the context of XAI, an explanation for an automatic decision or prediction, treated as an observation, can be obtained abductively by attempting the search problem over the set of the known information concerning that observation [50]. Alternatively, counterfactual explanation is widely addressed in the paradigm of case-based reasoning, i.e., a family of problem solving methods based on appeals to precedent solutions. In this setting, generating the most suitable counterfactual may be viewed as a search problem where the most similar precedent is looked for among those making part of the case database [14]. Furthermore, Keane *et al.* argue that applying case-based reasoning techniques for generating counterfactuals increases their explanatory competence [51].

Counterfactual explanations are normally considered contrastive by nature and therefore present a source of valuable complementary information to a given automatic prediction [52]. For instance, a counterfactual explanation of an ML-based algorithm prediction may describe "the smallest change to the feature values that changes the prediction to a predefined output" [53]. An important advantage of counterfactual explanations over their non-counterfactual analogs is that they are devoid of any prerequisites to the data or model. Indeed, counterfactual explanations are data-agnostic as they can be based on the features of the neighbouring data examples extracted from the same training set and/or on the data generated synthetically around the data instance in question. In addition, counterfactual explanations are, in principle, model-agnostic, as they are suitable to explain the output of any black-box algorithm in a post-hoc manner.

Whereas counterfactual explanation generation is concerned with a number of technical challenges, it also requires to take into account several ethical aspects. For instance, their use is expected to be safe (revealing model's internals through counterfactuals may lead to model stealing) [54], fair (discriminatory explanations should be avoided) [55], actionable (suggested changes in the input should be feasible) [56], and accountable (ensuring responsibility for the explanations provided) [57].

III. METHODOLOGY

The present survey has been undertaken as a systematic literature review following the guidelines by Kitchenham and Charters [58], Kitchenham *et al.* [59], and Wohlin [60]. The background notation necessary to follow the findings of the review is specified in Section III-A.

In short, the study comprises three phases as established in the research method by Kitchenham and Charters [58]: (1) planning the review procedure; (2) conducting the review; and (3) reporting the results. During the first phase, three research questions (RQ_1, RQ_2, and RQ_3) were specified (see Section III-B). Subsequently, we determined a search strategy to retrieve primary studies, i.e., we collected all the relevant publications investigating the research questions (see Section III-C). Then, we developed inclusion and exclusion criteria (see Section III-D) in order to select the studies relevant for this article. When the same publication was retrieved from multiple sources, all-but-one instances of the publication (duplicates) were discarded. In addition, we identified and added manually other relevant publications extracted from the bibliography lists of the previously selected manuscripts to ensure a maximum coverage of the related subject areas. It is worth noting that this additional procedure is informally known as snowballing [60]. Finally, we extracted and synthesized the data necessary to address the research questions (see Section III-E).

A. PRELIMINARY TERMINOLOGY

As has been shown in Section II, contrastive and counterfactual explanations presuppose a diverse nature across various application domains. Hence, let us now define the general terms used henceforth in this manuscript. As we are primarily concerned with explainability of AI algorithms, we define explanation in terms of the observed output of such an algorithm. Thus, we regard an *explanation* as a non-empty set of pieces of information justifying the given algorithm's output for an input data instance. The explanation for the given output on the basis of the features of the input data instance is deemed as *factual*. An explanation opposing the actual outcome to one of possible other outcomes is considered to be *contrastive* (e.g., "The data instance is of class A and not B because ..."). An explanation containing instructions on how the output could have been changed constitutes a *counterfactual* explanation (e.g., "The data instance would be of class B if ..."). Explanations exhibiting patterns of both contrastive and counterfactual explanation are deemed to be *contrastive-counterfactual* explanations (e.g., "The data instance is of class A and not B because .... However, it would be of class B if ...").

We distinguish between contrastive and counterfactual explanation throughout the rest of the manuscript if and only if only one of these two terms is used in the given primary study. In contrast, we unify the notions of counterfactual and contrastive explanation introducing the term *"contfactual explanation"* or *"contfactual"* to identify potential similarities and differences of both types of explanation

within a broader scope of literature. This term is used hereafter wherever both terms for contrastive and counterfactual explanation can be used interchangeably. The terms ''contrastive explanation'' and ''counterfactual explanation'' are only used when they are found in the corresponding study and cannot be used interchangeably in the given context. Notice that the term ''contfactual explanation'' is not equivalent to ''contrastive-counterfactual explanation'' but covers both independently used types of explanation as well as their fusion.

A theoretical framework providing justification and a reasoning mechanism for obtaining a contfactual explanation is regarded as a *theory of contfactual explanation*. Altogether, we use the term *contfactual explanation generation* to refer to the process of automatic composition of contfactual explanations for a given output of an AI algorithm in the form of a complementary piece of information associated to a factual explanation.

B. RESEARCH QUESTIONS

In order to reach the three objectives of the study as formulated in Section I, the following three research questions were specified:

- RQ_1: How are contfactual explanations defined in the literature?
- RQ_2: What are the state-of-the-art methods of contfactual explanation generation?
- RQ_3: How grounded are the state-of-the-art contfactual explanation generation methods on the theoretical approaches to contfactual explanation?

C. SEARCH STRATEGY

We selected the digital libraries Scopus and Web of Science (WoS) to retrieve relevant publications from. These libraries do not only include research publications in computing but also index studies across all scientific fields, which allows for an objective analysis of the interdisciplinary literature relevant to the research questions posed.

Subsequently, we performed six queries over the title, abstract, and author keywords in the aforementioned libraries (see the overall structure of the query pipeline in Fig. 1). It is worth noting that the proximity operator *NEAR* is used following the WoS notation whereas the equivalent proximity operator *W* is used for the same queries in Scopus. The following search strings were used for querying the digital libraries:

q_1 = counterfactual* W/3 expla*
q_2 = contrastive* W/3 expla*
$q_3 = q_1$ OR q_2
$q_4 = q_3$ AND (defin* OR theor* OR infer* OR implic*)
$q_5 = q_3$ AND (generat* OR implement* OR framework* OR develop* OR software* OR model* OR artificial intelligence OR AI) AND SUBJAREA(Computer Science OR Mathematics OR Engineering)
$q_6 = q_4$ AND q_5

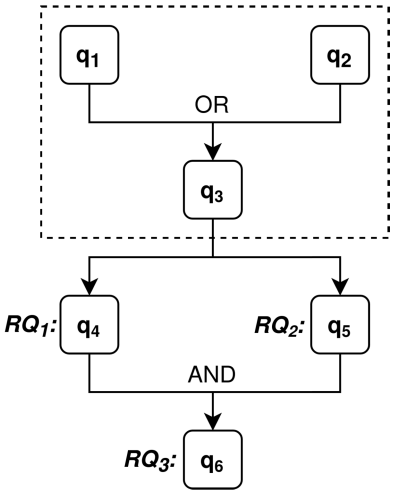

FIGURE 1. A pipeline of the queries executed. The queries found in the dashed area are considered preparatory to those directly addressing the research questions.

The search was performed on October 2nd, 2020. The search web tools of the selected digital libraries allow researchers to reproduce the original study. Furthermore, their use guarantees performing equivalent queries across both libraries. In order to capture all relevant publications, we only used the corresponding word-stems to allow for maximal diversity of the retrieved papers. For instance, the search item ''expla*'' was used to cover all publications containing such word-forms as ''explanation'', ''explaining'', ''explanatory'', and so on and so forth.

Queries q_1 and q_2 embrace all the up-to-date publications containing mentions of counterfactual and contrastive explanation, respectively, found across all subject areas. In addition, we used a window span of three words (i.e., ''NEAR/3'') to ensure that the attributes ''counterfactual'' and ''contrastive'' relate to explanation. The resulting sets of publications were then unified (q_3).

Subsequently, the preprocessed collection of publications was split into two overlapping subsets aiming to distinguish the publications covering theoretical accounts of contfactual explanation with the aim of extracting the related definitions, theories (or their inferences or implications) (q_4) and existing computational frameworks for contfactual explanation generation (q_5). The terms ''definition'', ''theory'', ''inference'', and ''implication'' as well as their corresponding word-forms (q_4) were expected to appropriately limit the pool of the unified set of publications with the aim of retrieving definitions as required for addressing RQ_1. Similarly, we used the terms ''generation'', ''implementation'', ''framework'', ''development'', ''software'', ''model'', and their corresponding word-forms (q_5) to retrieve publications concerning contfactual explanation generation frameworks. In addition, the terms ''artificial intelligence'' and ''AI'' were used to ensure retrieving relevant AI-related publications. Since RQ_2 addresses purely technical issues of

state-of-the-art implementations of such tools, we further imposed an additional restriction on q_5 so that it would return only publications from such subject areas as computer science, mathematics, and engineering. Last but not least, the findings from q_4 and q_5 were merged to examine the connection between the existing theories of contfactual explanation and frameworks for automatic contfactual explanation generation in the context of XAI (q_6).

It is important to note that publications retrieved as a result of q_4 form an exhaustive set of papers addressing RQ_1. Similarly, publications obtained as a result of q_5 address RQ_2. Finally, the papers that q_6 returned address RQ_3.

D. INCLUSION AND EXCLUSION CRITERIA

The publications retrieved during the initial search were subsequently inspected on the basis of the following inclusion and exclusion criteria. To address the epistemology of contfactual explanation, we filtered the retrieved publications to include in the collection of primary studies only those satisfying the following criteria: (1) a publication proposes a contrastive or counterfactual or contrastive-counterfactual approach to explanation or (2) it contains a clearly formulated definition of counterfactual or contrastive explanation referring to other publications in the corresponding field. In order to capture existing computational frameworks for contfactual explanation generation, we included publications that: (1) present a novel approach, method, or framework for contfactual explanation generation whose output can serve to explain the reasoning of an AI algorithm and (2) are found in such subject areas as computer science, mathematics, engineering as well as in their sub-fields.

In contrast, we excluded duplicate reports of the same studies appeared in both Scopus and WoS. As for the publications related to RQ_1, we also removed: (1) the studies whose contents did not introduce any contfactual theory of explanation or (2) those containing no formal or informal definition of contrastive or counterfactual or contrastive-counterfactual explanation. As for the publications related to RQ_2, we discarded: (1) the publications which were not related to AI algorithms or applications as well as (2) those where the proposed framework did not provide any human-comprehensible contfactual explanations as output.

E. DATA EXTRACTION AND SYNTHESIS

Table 1 shows the number of publications retrieved after each independent query, duplicates found among them in Scopus and WoS, as well as Candidate Primary Studies (CPS). Note that the numbers of duplicates indicated in Table 1 refer only to within-query duplicates, i.e., the same publications retrieved from Scopus and WoS for the given single query. Recall that q_4 and q_5 exhaustively cover all the three research questions. Hence, the numbers of CPS are calculated as a sum of the publications retrieved after q_4 and q_5. Furthermore, CPS are reduced by the number of publications addressing RQ_3 because they are found in both sets of publications collected for RQ_1 and RQ_2 and are therefore duplicates.

TABLE 1. Numbers of publications retrieved after each single query as well as those forming the pool of candidate primary studies. The numbers of publications making part of the primary studies are highlighted in bold.

	q_1	q_2	q_3	q_4	q_5	q_6	CPS
Scopus	169	125	288	**140**	**67**	22	185
WoS	122	94	212	**109**	**22**	7	124
In total (including duplicates)	291	219	500	**249**	**89**	29	309
Duplicates	108	80	184	**92**	**21**	6	107
In total (excluding duplicates)	183	139	316	**157**	**68**	23	202

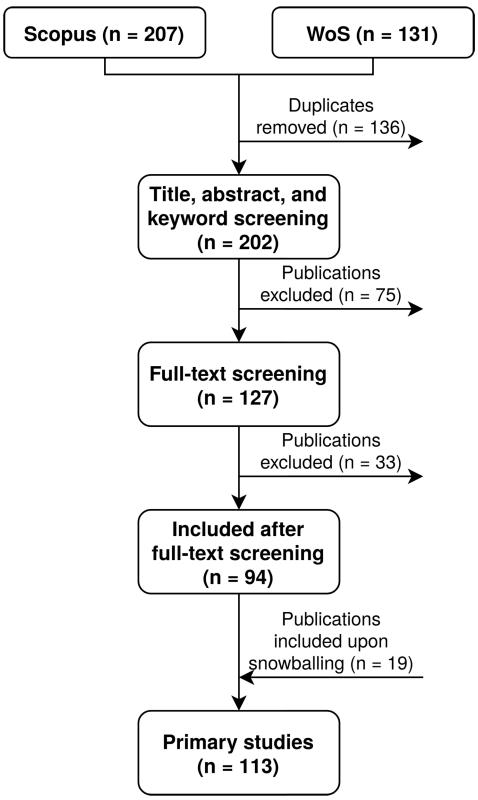

FIGURE 2. A flow diagram of the primary study selection on the basis of queries q_4 and q_5 (n is the number of publications at each stage).

Fig. 2 displays the flow diagram of the primary study selection. A sum of 338 publications (207 from Scopus and 131 from WoS) made up the collection of CPS addressing the research questions. 107 within-query duplicates were identified and removed from further analysis. In addition, 29 more duplicates were excluded when merging the sets of publications retrieved after q_4 and q_5. All in all, 136 duplicates were removed.

The title, abstract, and author keywords of each candidate primary study were screened to discard the studies irrelevant to the research questions posed. As shown in Fig. 2, 75 publications were deemed irrelevant and filtered out at this stage. A deeper analysis of the remaining 127 publications

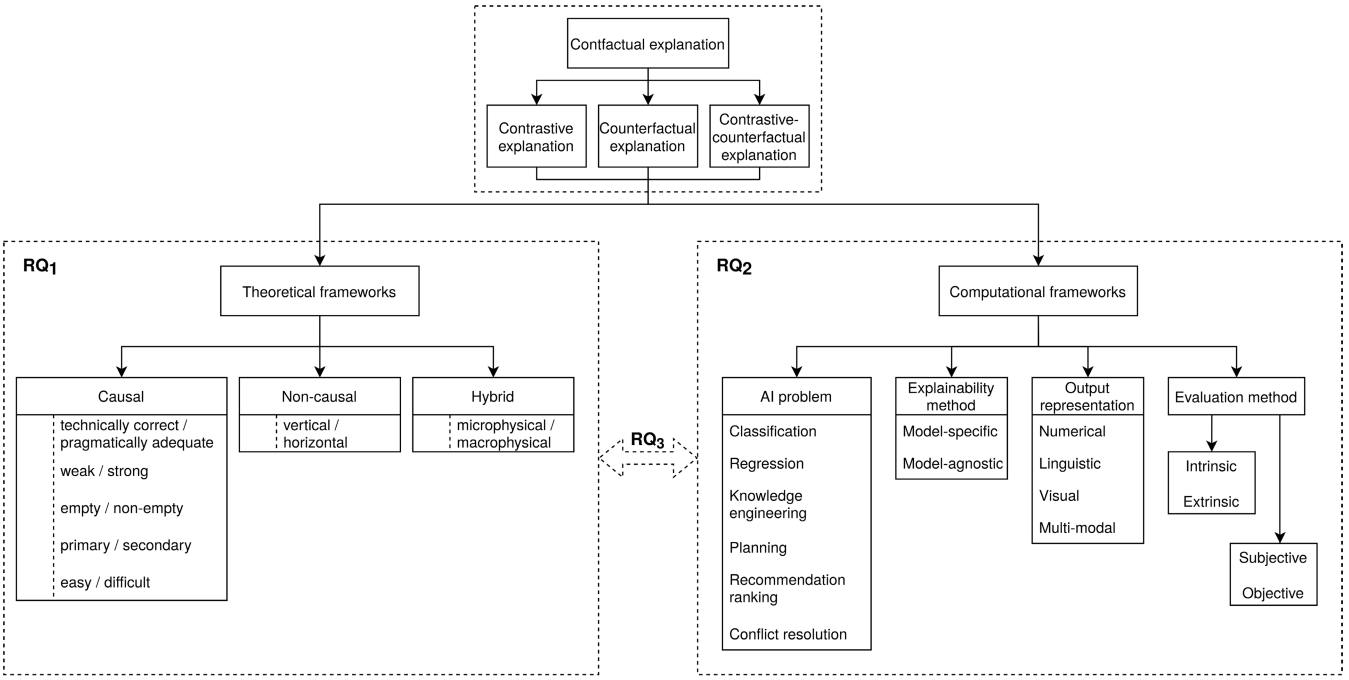

FIGURE 3. A taxonomy of contfactual explanation emerging from our systematic literature review.

TABLE 2. The exhaustive list of all the primary studies in relation to each research question.

Research question	Primary studies
RQ_1	[13], [17], [21], [22], [36], [37], [49], [61]–[127]
RQ_2	[6], [46], [48], [49], [55], [56], [61], [67], [80], [81], [84], [86], [91], [93], [94], [100], [105], [122], [128]–[161]
RQ_3	[49], [61], [67], [80], [81], [84], [86], [91], [93], [94], [100], [105], [122]

enforced us to discard 33 studies which did not satisfy the inclusion criteria. Finally, 19 papers were added to the review upon inspecting the bibliography of the primary studies. As a result, 113 unique publications formed the exhaustive pool of primary studies.

Table 2 presents the list of primary studies selected for the review. Thus, a collection of 74 out of 113 (65.49%) original primary studies were found to formulate definitions for contfactual explanation and/or address theoretical accounts thereof (RQ_1). In addition, 52 out of 113 (46.02%) publications describe frameworks (or extensions of other frameworks) for contfactual explanation generation (RQ_2). Note that 13 out of 113 (11.50%) primary studies were found to address both RQ_1 and RQ_2 and therefore answer RQ_3.

The following data were extracted from each primary study: title, authors, year of publication, author keywords. In addition, all publications related to RQ_1 were read to analyze contfactual theories of explanation and, subsequently, extract the sought-for definitions of contfactual explanation. As RQ_2 concerned a broader number of technical characteristics of contfactual explanation frameworks, we additionally extracted the following information: (1) the problem that the retrieved framework aims to solve; (2) the method proposed for contfactual explanation generation; (3) the form of output explanation (for instance, textual or visual); and (4) the corresponding evaluation methods. Based on the data extracted from the primary studies, the publications were grouped and classified in accordance with the aforementioned criteria.

IV. RESULTS

Prior to answering the research questions, we carried out a bibliometric analysis over the results of the general independent queries on counterfactual and contrastive explanation (q_1 and q_2, respectively) as well as their union (q_3). We report the results of the bibliometric analysis in Section IV-A. The findings related to the theoretical accounts of contfactual explanation (RQ_1) are presented in Section IV-B. The analysis of the computational frameworks for contfactual explanation generation (RQ_2) can be found in Section IV-C. Finally, the publications describing theoretically grounded computational frameworks (RQ_3) are reported in Section IV-D.

An emerging taxonomy of contfactual explanation frameworks is depicted in Fig. 3 and forms the core of the results discussed in the rest of the manuscript.

A. BIBLIOMETRIC ANALYSIS

The bibliometric analysis over the queries q_1, q_2, and q_3 allows us to obtain a big picture of the research area of contfactual explanation generation and spot its key characteristics. To illustrate the state of affairs within the field, we report annual scientific production and maps of author keywords revealing the main problem-specific notions. The reference

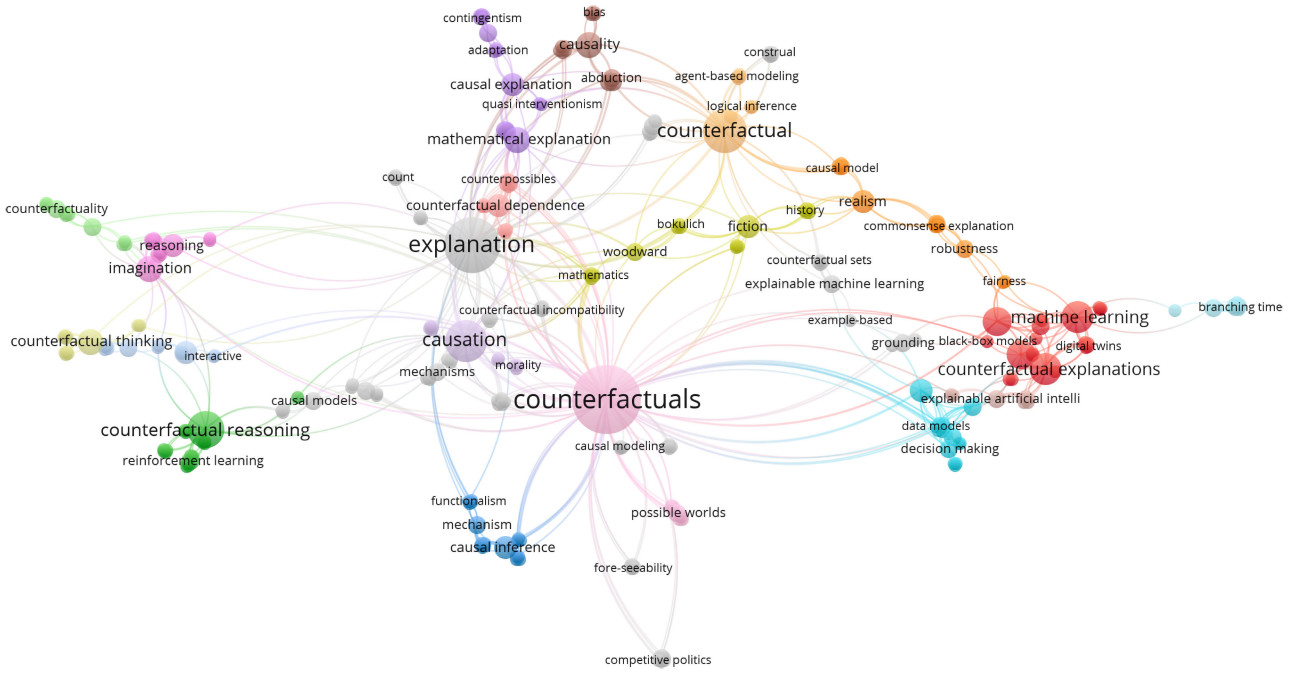

FIGURE 4. The map of author keywords for the q_1 publications.

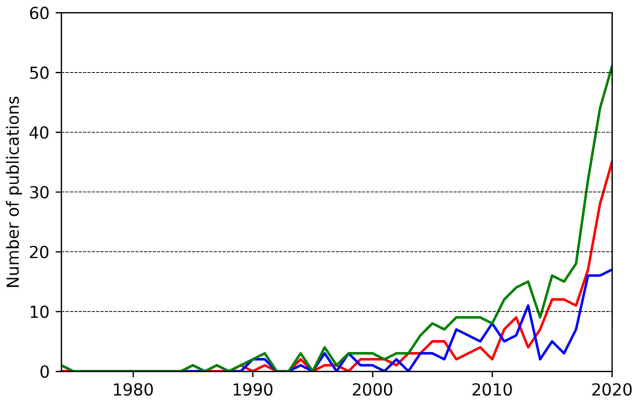

FIGURE 5. Annual scientific production for the publications retrieved after q_1 (the red line), q_2 (the blue line), and q_3 (the green line).

manager *Mendeley* was used to filter out duplicate publications. In addition, we utilized the tool *VOSViewer* [162] to generate the author keyword maps.

It can be seen that contfactual explanation appears to attract an increasing attention across all subject areas in the past two decades. Furthermore, Fig. 5 shows a rapid rise in the number of publications in the past three years. It is worth noting that the number of publications in 2020 is limited to the search date.

Author keyword maps allow us to present an overview of the terms most relevant to those specified in the preparatory queries (q_1, q_2, and q_3). For illustrative purposes, non-linked keywords were deemed to be outliers and filtered out from the analysis. Table 3 shows the overall number of keywords as well as that of linked keywords for each preparatory query.

TABLE 3. Numbers of linked and non-linked keywords in the preparatory query results (q_1, q_2 and q_3).

Query	All keywords	Linked keywords
q_1	422	323
q_2	341	228
q_3	702	530

Fig. 4 shows a graph containing the most popular author keywords for counterfactual explanation. It can be concluded that counterfactual explanation is often investigated in the context of causation (pay attention to such keywords as ''causation'', ''causal inference'' or ''causal models'') as well as cognitive science (as reflected by the keywords ''imagination'', ''reasoning'', etc.) and AI (''machine learning'', ''data models'', ''black-box models''). Similar notions are observed to be essential for contrastive explanation (see Fig. 6). However, a distinction between different clusters in the latter case is visible more clearly. This is hypothesized to be due to a more diverse usage of the term ''contrastive explanation'' across various scientific areas.

A stronger impact of counterfactual explanation in the results of the joint query q_3 appears to affect significantly the overall allocation of the related keywords in the corresponding keyword map (see Fig. 7). The keywords identified in the studies related to q_3 testify that the issue of contfactual explanation is highly interdisciplinary and finds application in both humanities and natural sciences.

B. CONTFACTUAL EXPLANATION AS DEFINED IN RELATED THEORIES (ANSWER TO RQ$_1$)

As presented in Section II, the surface form of contfactual explanation is found to preserve the same syntactic structure

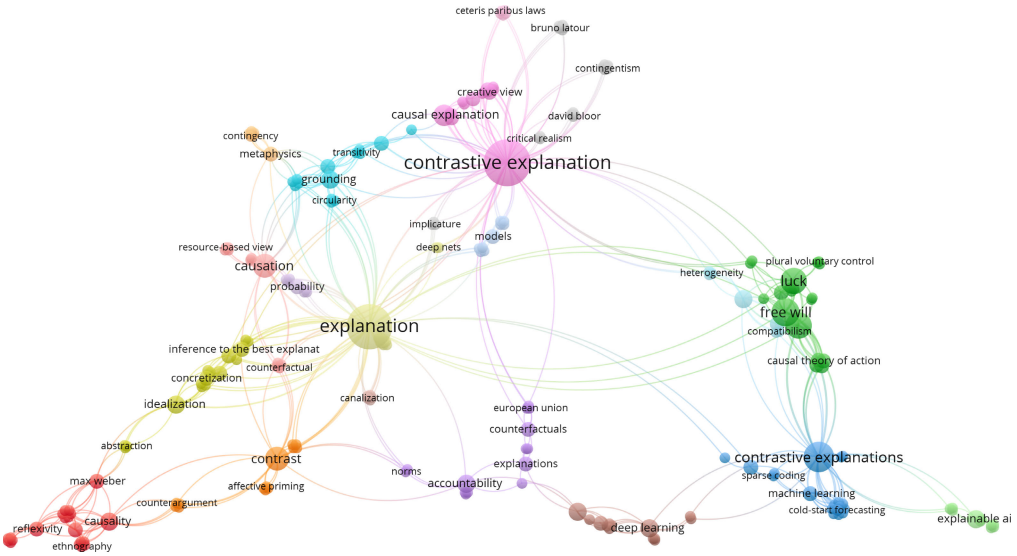

FIGURE 6. The map of author keywords for the q_2 publications.

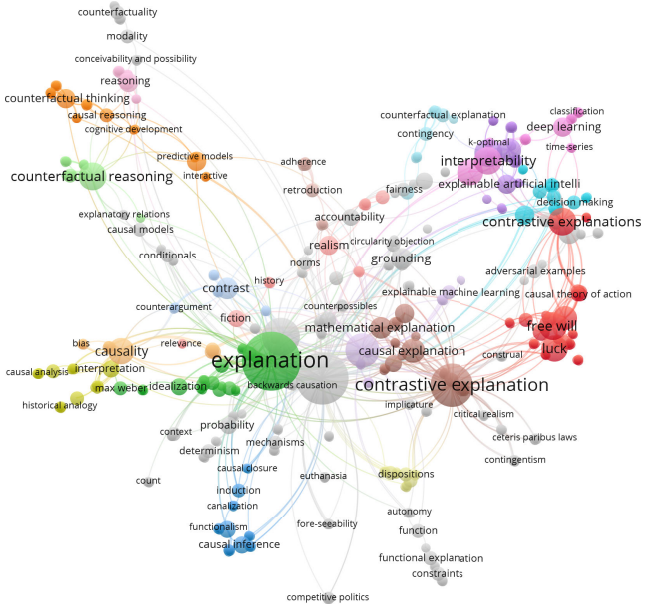

FIGURE 7. The map of author keywords for the q_3 publications.

in general. However, a major factor discerning theoretical approaches to contfactual explanation is found to consist in their relation to causation. As will be shown below, several accounts of contfactual explanation presuppose a causal nature and establish a causal contfactual dependency between the phenomenon to be explained and the explanation itself. In contrast, other theoretical frameworks seek purely non-causal dependencies in explanation. In addition, several researchers attempt to unify causal and non-causal contfactual explanation under the same paradigm. Hence, we distinguish three main groups of contfactual explanation that encompass all the retrieved primary studies: causal, non-causal, and hybrid. The exhaustive list of the retrieved publications in accordance with the suggested taxonomy is presented in Table 4 and discussed further in the following sections.

TABLE 4. A classification of approaches defining contfactual explanation.

Approach	Publications
Causal	[13], [21], [36], [37], [61]–[63], [66]–[68], [70]–[73], [75], [77]–[79], [82]–[84], [88]–[90], [92]–[94], [97], [99], [101]–[103], [106], [107], [109], [113]–[121], [123]–[127]
Non-causal	[49], [64], [65], [69], [74], [80], [81], [86], [87], [91], [96], [100], [110], [111], [122]
Hybrid	[17], [22], [76], [85], [95], [98], [104], [105], [108], [112]

Remind that contfactual explanation embraces contrastive, counterfactual, and contrastive-counterfactual explanation. Each type of contfactual explanation is present in the findings, causal counterfactual making up a majority of the considered theoretical frameworks (see Fig. 8). Hence, we analyze each contfactual explanation type independently in terms of causality in this section to draw a comparison between different approaches. In addition, we consider (1) the issue of quantitative evaluation of causality for causal contfactuals as reflected in specific primary studies and (2) different subcategorizations of causal, non-causal, and hybrid contfactual explanation.

1) CAUSAL CONTFACTUAL EXPLANATION

- **Causal contrastive explanation** is frequently found to be designed as an answer to a why-question of the following canonical form: "Why P rather than Q?" where P is an explanandum (i.e., the fact to be explained), Q being a foil (i.e., one of alternative non-occurring options) [21]. Lipton introduces the notion of

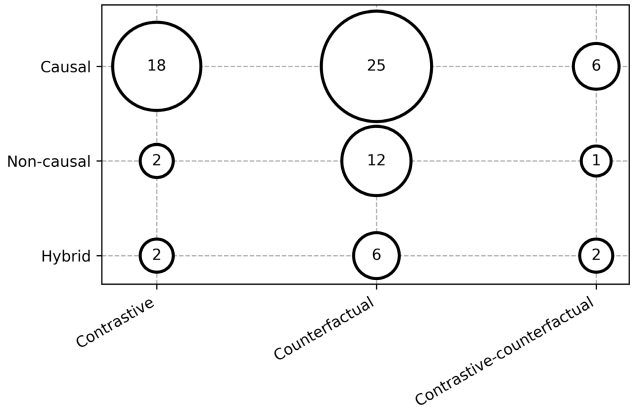

FIGURE 8. Numbers of identified theoretical contfactual explanation frameworks with respect to causality.

''difference condition'': the contrast between the fact and the foil is explained by identifying the cause of the fact and proving the absence of the corresponding cause of foil [103]. Following Lipton [103], Kean redefines a contrastive explanation to be the difference between the causal explanations for the question and the contrast [93]. Barnes further requires that ''P and Q be culminating events of a single type of natural causal process'' [63]. Similarly, Day and Botterill introduce the concept of differential inference, i.e., a form of inference based on contrastive explanation that ''can be used in order to generate causal hypotheses'' [77].

Van Fraassen formalizes a contrastive question to be the triple $\langle P, X, R\rangle$ where P is a topic (explanandum), X is a contrast space or contrast-class (i.e., a set of alternative answers to the given question), and R being the corresponding relevance criteria [13]. Then, the answer to a why-question must differentiate the topic from the contrast space. Contrarily, Chien excludes the contrast class when taking contrastive explanation as a model for scalar implicature [75].

Hitchcock generalizes the notion of contrastive explanation over all explanations bearing contrastive stress irrespective of their syntactic structure [88]. Conversely, Aguilar-Palacios *et al.* oppose the alternative explanation-seeking question ''Why P rather than Q?'' to the congruent question ''Why P but Q?'' [61]. This formulation of the question generalizes the explanation to justify why some fact P occurs in the current situation whereas some foil Q occurred in different circumstances with the aim of establishing a cause and effect relation between the fact and the foil. Similarly, Tsang and Ellsaesser claim that a contrastive explanation should point to the importance of identifying the most relevant factors differing the causal histories of the fact and the foil where both fact and foil must be true [124].

Contrastive explanations may not only concern the explanandum but also the answers to contrastive questions (often referred to as explanans). Thus, Sober stresses that the question of whether some hypothesis H explains why a non-contrastive proposition E is true is ''incomplete until H is contrasted with an alternative hypothesis'' [119]. In addition, the canonical forms of the contrastive question and the corresponding explanatory answer (i.e., the core of contrastive explanation) have raised a number of epistemological concerns in philosophy of science. For instance, Dickenson reformulates the contrastive explanation-seeking question to be: ''What explains how it is possible that an agent can act on R_1 other than R_2, given that R_2 is present?'' [78] (where reason R_1 is the cause of some action and R_2 is not). The notion of contrastive explanation is further developed in the agent-causal theory of free will. Thus, contrastive explanation is applied to agent's decision-making (i.e., why the agent makes a choice refraining from an alternative choice) [82], [101].

Campbell redefines a contrastive explanation in terms of the so-called ''structuring causes'', i.e., the traits of the structure of the causal system that trigger actual causes of some event to happen [73]. A cause of this kind is responsible for the connection between the types C and M in a system S. A contrastive explanation thus explains why a system S is claimed to be ''wired'' in such a way that an internal state of type C regularly causes a movement of type M. Similarly, Kim *et al.* regard contrastive explanation as a constraint for a system to be satisfied by a specific set of plan traces [94].

Finally, Boulter illustrates the use of contrastive explanations to distinguish between actual and non-actual biological forms [70]. Claiming all explanations in biology to be causal, the researcher introduces the following template for a causal relation in contrastive explanation: ''c_1 rather than c_2 or c_3 or c_n causes e_1 rather than e_2, or e_3 or e_n'' leaving contrasting causes implicit.

- **Causal counterfactual explanation.** Most of the considered studies on causal counterfactual explanation relate to either of the four theoretical milestones: Lewis-Stalnaker's theory of closest possible worlds [36], [120], Pearl's Structural Causal Models (SCM) [37], Woodward's Counterfactual Theory of Explanation (CTE) [125], or the Neyman-Rubin Causal Model (NRCM) [107], [115] (see Table 5).

  The Lewis-Stalnaker approach codifies a counterfactual conditional as a logical proposition where the antecedent

TABLE 5. Counterfactual explanation theories reflected in the primary studies.

Theory	Author(s)	Primary studies concerned
"Closest possible worlds"	Lewis [36], Stalnaker [120]	[66], [72], [99], [109] [114], [117], [123]
SCM	Pearl [37]	[71], [113]
CTE	Woodward [125]	[62], [67], [68], [79], [83], [116]
NRCM	Neyman [107], Rubin [115]	[90]

and the consequent are connected by means of the ''might''- or ''would''-conditional operator. Exploiting the mechanism of possible world semantics, the truthfulness of a counterfactual is assessed by assigning to it a binary truth-value in accordance with its proximity to the world in question.

Following this approach, Kutach defines counterfactuals in natural language to be ''propositions obeying a logic whose semantics is given in terms of a comparative similarity relation among possible worlds'' [99]. Similarly, Strohminger and Yli-Vakkuri assume that the counterfactual modus ponens preserves truth-functional possibility (''If it is possible that p and p counterfactually implies q, then it is possible that q'') [123] where p is a logical proposition and q is a conjunction of such propositions.

Briggs extends the Lewis-Stalnaker model by applying causal modeling language to comprise logically complex antecedents [72]. Schweder considers a counterfactual to be an implicit claim within the explanatory answer to an explanation-seeking question [117]. In addition, Pruss and Rasmussen take into account antecedents that are not necessarily ''contrary-to-fact'' and define a counterfactual to be a contingent proposition establishing a causal connection between a specific description of the circumstances of a choice and a report of an action in such circumstances [109].

Pearl's SCM operates on a predefined causal model $M = \langle U, V, F \rangle$ consisting of sets of background variables determined by factors outside the model (U) and within it (V) and a set of functions $F = \{f_i \mid 1 \leq i \leq n\}$ mapping from $U \cup (V \setminus V_i)$ to V_i, that associates each variable V_i with all the variables from U and V. Given a set of variables $X \in V$ and a causal submodel $M_x = \langle U, V, F_x \rangle$ so that $F_x = \{f_i : V_i \notin X\} \cup \{X = x\}$ and by defining a minimal change in M required to make a selected variable $X = x$ ($X \in V$) hold true under any $u \in U$, a causal counterfactual is formally defined as the solution for some subset $Y \in V$ on the set of equations F_x [37]. Counterfactuals are thus pruned by interventions on the antecedent component [113], which leads to interpreting counterfactuals as non-observable hypothetical contrasts [71].

Similarly to Pearl's SCM, Woodward's CTE establishes the counterfactual dependence between the two variables by means of the intervention mechanism. Thus, for two variables X and Y taking on some values x and y, respectively, to explain the value of y counterfactually is to show that Y would have taken on some value y' if X had taken some counterfactual value x' [125]. In other words, some small enough change in the value of X from x to x' would cause a change in Y from y to y' in the absence of changes in values of other variables.

Following Woodward's theory, Schneider and Rohlfing define a counterfactual as ''a theoretically relevant manipulation of the observed case in order to ascertain whether this manipulation would make a difference to the outcome'' [116]. Further, Bertossi defines a causal counterfactual explanation to be a set of the original feature values in the given data instance that are affected by a minimal counterfactual intervention [67] (where minimality is assumed to be based on a partial order relation on counterfactual interventions).

Conversely, Andreas and Casini reconsider explanatory counterfactuals to be ''hypothetical assumptions about the values of quantities or the values of propositions''. They argue that Woodward's interventionist account of explanation cannot handle the cases where interventions are physically impossible (e.g., due to violations of laws of nature) [62]. Applied to theorem proving, Gijsbers leaves out the mechanism of intervention from Woodward's CTE. He states that a mathematical proof has explanatory power only when the explanandum is complemented with a contrasting claim that shows how the mathematical object in question varies in the process of theorem proving. Also, Fang infers counterfactual dependencies in the form of counterfactual claims: ''in the model M, had the variable X taken such-and-such a value x_i, then the variable Y would have taken such-and-such a value y_j'' [79].

Last but not least, Holland argues that causal counterfactuals are highly relevant to research in social sciences. Thus, he follows NRCM interpreting counterfactuals in terms of potential outcomes of a dependent causal variable given some intervention with respect to that variable [90].

- **Causal contrastive-counterfactual explanation** is sometimes considered to include ''all kinds of subjunctive conditionals, regardless of whether the antecedent is true in the actual world or not'' [126]. Thus, Kuorikoski and Ylikoski elaborate on a contrastive counterfactual theory of explanation claiming that the property of contrastiveness helps to resolve linguistic ambiguity inherent in explanation [97]. In this setting, interventions (or manipulations) specify the truth conditions of such explanations: ''c [$c*$] causes e [$e*$] if we can bring about $e*$ [e] by bringing about $c*$ [c]'' [127] (where c and $c*$ are causes, e and $e*$ being the corresponding effects).

  Following Kuorikoski and Ylikoski [97], Northcott examines explanatory relevance of counterfactuals placed in a contrastive framework [106]. Similarly, Hohwy regards causal counterfactuals as an integrative part of causal contrastive explanations. Thus, he claims counterfactuals supported by laws are able to ''go into contrastive explanations even though unfavourable conditions ensure that the forces they describe are not actually occurring in the way described by any law taken alone'' [89]. On a similar note, Steglich-Petersen [121] proposes two-level semantics of contrastive causal statements requiring specific semantically complete counterfactual justifications.

- **Degree of causality in causal contfactuals.** Causal relations between the variables in explanation are not always considered binary (i.e., in the presence/absence of a cause). There have been several attempts to measure a degree of causality in contfactual explanation. Since causal contrastive explanations describe a certain aspect of the explanandum, Rips and Edwards claim them to be partial by nature [113]. In other words, the explanatory power of such explanations can be quantified and compared with that of others. In light of this assumption, Northcott defines the degree of causation (i.e., causal strength of a cause variable) to be the difference between the values that the effect variables take on in the actual and counterfactual cases [106]. In this regard, he determines a counterfactual to be the value of the effect variable. A counterfactual can thus be measured quantitatively as the distance between the target levels of the causal effect variables. Similarly, Ylikoski and Kuorikoski distinguish five dimensions of explanatory power of contrastive explanations: (1) non-sensitivity (i.e., how sensitive the explanation is to background conditions); (2) precision (i.e., how precisely the explanation characterizes the explanandum); (3) factual accuracy (i.e., a proportion of true facts captured by the given explanation in comparison with another); (4) degree of integration (i.e., unification to a larger theoretical framework); and (5) cognitive salience (i.e., ''the ease with which the reasoning behind the explanation can be followed'') [126].
- **Subtypes of causal contfactuals.** It is worth noting that several subcategorizations of causal contfactuals have been suggested within some of the aforementioned theoretical frameworks.

  As for contrastive explanation, Franklin follows Hitchcock [88] differentiating ''technically correct contrastive explanation'' (the explanation citing explanatory relevant information) and ''pragmatically adequate/defective contrastive explanations'' (the explanation providing more information than explanatory relevant) [82]. Levy distinguishes between weak contrastive explanation (if the agent is not able to explain how the agent-causal power was exercised for reasons) and strong contrastive explanations (otherwise) [101].

  As for counterfactual explanation, Holland points to a deceptive use of ''empty'' counterfactuals, i.e., counterfactuals whose antecedent ''could never occur in any real sense'' [90]. Steglich-Petersen distinguishes between primary counterfactuals (i.e., those that relate two events A and B as the cause and the effect) and secondary ones (i.e., those that establish the fact that it is event A that causes B to happen) [121]. Finally, Schneider and Rohlfing claim counterfactuals to be either easy or difficult [116]. From this perspective, easy counterfactuals are ''the assumptions about the outcome of logical remainders'' that simplify theoretical expectations. In contrast, the assumptions that simplify the solution ''but run counter to our theoretical expectations about whether single conditions involved in a remainder should or should not contribute to the outcome'' are assumed to be difficult.

2) NON-CAUSAL CONTFACTUAL EXPLANATION

- **Non-causal contrastive explanation.** Notably, non-causal contrastive explanations can address the physical nature of a modeled system. Hence, they can be used to explain the properties and relations inherent to such systems. Thus, Chakravartty extends the concept of contrastive explanation to answering non-causal what-questions, e.g.: ''What dispositions of p are relevant to circumstances x as opposed to y?'', where p is the object whose traits require an explanation and x and y are the circumstances determined by the question-dependent context [74].

  In contrast to Dickenson [78] (see Sect. IV-B1), Botterill appeals to a non-causal nature of contrastive explanations [69]. Thus, the researcher argues that ''the fact that, in the absence of R_2 but with R_1 still present the agent would perform an action of some kind does not show that when both R_1 and R_2 are present an agent does not act in that way because of both those reasons'' (where reason R_1 is the cause of some action and R_2 is not).
- **Non-causal counterfactual explanation.** Reutlinger develops a non-causal counterfactual theory of explanation to apply it to Euler's explanation[3] and the renormalization group theory[4] [110]. This counterfactual theoretical framework is subsequently extended to capture non-causal explanations in metaphysics [111].

  Driven by the assumption that physical facts and mathematical models share certain features, Baron *et al.* apply a structural equation modelling framework to model counterfactuals that could explain physical facts in terms of non-causal mathematical explanations [64]. Further, Baron introduces the concept of the so-called ''counterfactual scheme'' applied to mathematical explanation [65]. A counterfactual scheme is thus defined as a triple containing (1) a counterfactual statement with non-logical expressions substituted with variables, (2) instructions stating which parts of the statement can be substituted to produce a counterfactual, and (3) a classification for evaluating the given counterfactual. A counterfactual is then claimed explanatory if all the instances of a counterfactual scheme are true and at least two counterfactual schemes are distinct so that the corresponding physical laws relevant for evaluation of the given counterfactuals are different. Also, Hird uses

[3]Reutlinger refers to the phenomenon found in the city of Königsberg where no-one succeeded to cross the seven bridges located in four different parts of the city exactly once. Euler provided a non-causal explanation for this phenomenon in terms of graph theory.

[4]According to Reutlinger, renormalization group explanations are intended ''to provide understanding of why microscopically different physical systems display the same macrobehavior when undergoing phase transitions'' [110].

the term ''counterfactual'' to define projects that have been funded in the absence of congressional committee influence [87].

In addition, a number of definitions for non-causal counterfactual explanation come from AI. In the context of automatic decision-making, counterfactuals are found to be most generally defined as counterarguments for an alternative prediction [86]. Fernández *et al.* refer to a counterfactual as an effective type of explainable ML technique that explains predictions by describing the changes needed in a sample to flip the outcome of the prediction [81]. More precisely, Fernández *et al.* define a counterfactual for classification tasks as a ''hypothetical instance similar to an example whose explanation is of interest but with different predicted class'' [80]. Kanehira *et al.* attempt to explain counterfactually video classification output framing a (visual-linguistic) counterfactual explanation in the form of the conditional statement ''X would be classified as B and not A if C and D are not in X'' [91] (where X is the data example requiring an explanation, A is the class predicted for X, B is the contrast-class in question, C and D are specific visual patterns present or absent in the given video frame X). On a similar note, Laugel *et al.* treat a counterfactual explanation as a specific data instance, close to the observation whose prediction is explained, but predicted to belong to a different class [100]. Kostic defines a counterfactual to be a statement describing a hypothetically different situation to the actual state of affairs [96]. He distinguishes between vertical and horizontal counterfactuals. Thus, a counterfactual is considered vertical if ''a global topological property determines certain general properties of the real-world system''. In contrast, a counterfactual is deemed horizontal if ''a local topological property determines certain local dynamical properties of the real-world system''. Finally, Stepin *et al.* point that a counterfactual explanation should refer to a set of features ''minimally different from those inherent to the original data point'' [122].

- **Non-causal contrastive-counterfactual explanation.** Poyiadzi *et al.* do not distinguish between counterfactual and contrastive explanations assuming counterfactuals to be the new state of the considered object [49].

3) HYBRID CONTFACTUAL EXPLANATION

- **Hybrid contrastive explanation.** Chin-Parker and Bradner [17] as well as Chin-Parker and Cantelon [76] provide a unified theoretical framework for causal and non-causal contrastive explanation for category learning. Emphasizing the crucial importance of context for an explanation, they consider a contrast class to be a set of non-occurring alternates that delimits the set of potentially relevant information irrespective of the inherent causal relations.
- **Hybrid counterfactual explanation.** Explanatory pluralism is as well recognized in the research on counterfactual explanation. Thus, Byrne states that ''not all counterfactuals are about causes, and counterfactuals that imply a causal relation differ in systematic ways from counterfactuals that identify other sorts of relations, such as intentions'' [22]. Indeed, a large body of research on both causal and non-causal counterfactual explanation testifies that counterfactuals have a diverse nature with respect to causality [104]. Thus, Lowe claims counterfactuals to be causal ''when the modality involved is evidently natural or causal necessity''. Contrarily, other explanation cases such as those arising in mathematics ''clearly do not involve this sort of necessity, but instead something like logical necessity'' [104]. Further, Knowles and Saatsi discuss the notion of explanatory generality presuming both causal and non-causal nature of counterfactuals arguing that ''explanatory counterfactuals are appropriately directed and change-relating, capturing objective, mind-independent modal connections that show how the value of the explanandum variable depends on the value of the relevant explanans variables'' [95].

  In light of this, there have been several attempts to unify causal and non-causal counterfactuals within one framework. Hence, a hybrid approach, originating from monism,[5] has been adopted to unify causal and non-causal counterfactual explanation. Following this approach, Reutlinger introduces a unified explanation framework consisting of the following elements: a statement about the explanandum E, a set of generalizations (or explanans) $G_1,\ldots,G_m$, and a set of auxiliary statements setting initial conditions for the explanatory system [112]. A relation between an explanandum and a set of explanans is claimed to be explanatory if and only if at least one of the explanans supports the counterfactual statement ''had $S_1,\ldots,S_n$ been different than they actually are (in at least one way deemed possible in the light of the generalizations), then E or the conditional probability of E would have been different as well''. At the same time, the generalizations and auxiliary statements must logically entail the explanatory statement in question. As such, both causal and non-causal explanations are argued to be captured because they ''reveal counterfactual dependencies between the explanandum and the explanans''. Following Reutlinger's account of explanation, Held argues that the notion of counterfactuals can hardly be supported only by generalizations [85]. Furthermore, true generalizations (e.g., ''all ravens are black'') might not allow for counterfactual situations at all. Instead, he weakens the counterfactual dependency to shift from generalizations to plain counterfactuals.

  Mothilal *et al.* suggest a feature-based counterfactual explanation generation framework where importance of independent features is evaluated [105]. Nevertheless,

[5]Monism is a philosophical account of explanation that captures both causal and non-causal explanations reducing them to a single entity [112].

TABLE 6. A classification of the contfactual explanation generators by AI problem.

AI problem	Publications
Classification	[6], [46], [48], [49], [55], [67], [80], [81], [86], [91], [100], [105], [122], [128]–[131], [133]–[148], [150]–[155], [158]–[160]
Regression	[56], [61], [129]
Knowledge engineering	[93]
Planning	[94], [132], [156], [157], [161]
Recommendation	[84]
Conflict resolution	[149]

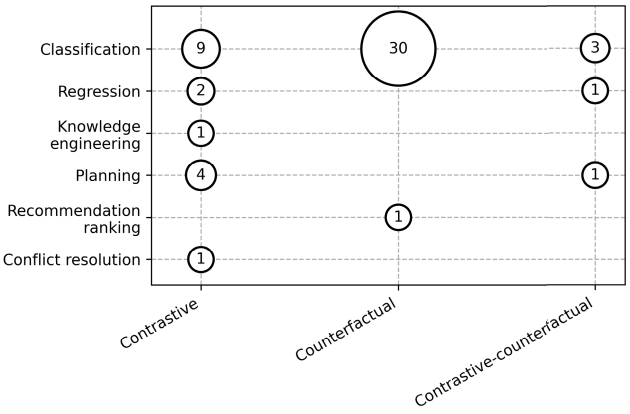

FIGURE 9. Numbers of frameworks grouped by AI problem with respect to the type of contfactual explanation generated.

they emphasize the need for causal attribution, as ignoring causal relations may lead to generating unfeasible counterfactuals. Therefore, they suggest a hybrid framework for counterfactual explanation generation.

- **Hybrid contrastive-counterfactual explanation.** Kuorikoski and Ylikoski point to the multifaceted nature of contrastive-counterfactual explanation. They argue "there exist constitutive and possibly formal counterfactual dependencies as well as combinations of these" [98]. Similarly, Pexton suggests a two-level hierarchy of explanation [108]: microphysical explanations are non-causal and form the lower-level of the hierarchy whereas manipulable causal explanations are placed at the higher-level.

C. CONTFACTUAL EXPLANATIONS AS DEFINED IN AUTOMATIC GENERATION FRAMEWORKS (ANSWER TO RQ_2)

The analysis of the primary studies related to RQ_2 allows us to categorize the state-of-the-art contfactual explanation generation frameworks in accordance with the following criteria: (1) the problem the solution for which is to be explained (i.e., the AI problem); (2) the method employed to generate such an explanation (i.e., the explainability method); (3) the output representation of the explanation; and (4) the evaluation method thereof.

1) AI PROBLEM

Contfactual explanations are used to justify automatic decisions obtained for a variety of AI-related problems. Table 6 provides the reader with a taxonomy of the state-of-the-art frameworks from the primary studies. It is derived from the considered publications in terms of the domain tasks that these frameworks are used for. As depicted in Fig. 9, most contfactual explanation generation frameworks deal with counterfactual explanation (31 out of 52 frameworks; 59.62%). In contrast, 17 out of 52 (32.69%) generate contrastive explanations. Only four studies (7.69%) fuse contrastive and counterfactual explanations. One of these studies [129] deals with both classification and regression.

- **Contfactuals for classification.** A vast majority of state-of-the-art AI applications that generate contfactuals (42 out of 52; 80.77%) are used to explain the outcome of ML-based classifiers, i.e., algorithms that learn a mapping function $f : X \longrightarrow Y$ from a training dataset of n labeled examples $X = \{x_i \mid 1 \leq i \leq n\}$ to a discrete output variable (class) $Y = \{y_j \mid 1 \leq j \leq m\}$ where m is the number of classes. Indeed, contfactuals are particularly suitable for informing the end-user why a given data example is assigned a particular class label. Thus, the outlined classification-oriented frameworks are evaluated on classifiers based on logistic regression [55], [136], [153], [158], decision trees [46], [80], [122], [140], [150], [155], [159], gradient boosted decision trees [147], support vector machines [131], [138], [146], random forests [81], [86], [142]–[144], neural networks [6], [48], [49], [91], [129], [130], [133], [135], [139], [141], [145], [148], [151], or combinations of these [100], [105], [134], [152], [154], [160]. In three studies [67], [128], [137], the classifiers used in the experiments are not specified.
- **Contfactuals for regression.** One of the classification-oriented frameworks [129] is extended to also handle the regression problem, i.e., learning a mapping function f from a training dataset X to a continuous output variable Y. However, the continuous output is, in this case, subsequently converted to a lower-scale discrete value mapped to a textual description similar to that typical of a classification problem. The other frameworks addressing the regression problem aim to leverage gradient-boosted decision trees [61] and indicate how large errors in regression tasks could be overcome [56].
- **Contfactuals for knowledge engineering.** The first of the considered frameworks (in chronological order) [93] offers explanations by reasoning abductively over the information extracted from a given knowledge base to answer a specific contrastive question. In this setting, an explanation is considered to be a consistent set of disjunctive literals for the explanation-seeking question. It is worth noting that the framework is not designed to provide explanations for ML algorithms.

TABLE 7. A classification of the contfactual explanation generators by explainability method.

Explainability method	Publications
Model-specific	[46], [48], [56], [61], [80], [84], [86], [91], [93], [94], [122], [128], [132], [138], [139], [145], [147], [149], [153], [155]–[158], [161]
Model-agnostic	[6], [49], [55], [67], [81], [100], [105], [129]–[131], [133]–[137], [140]–[144], [146], [148], [150]–[152], [154], [159], [160]

- **Contfactuals for planning.** Contfactual explanation generation appears highly relevant to sequential tasks in robotics such as automatic planning [94], [157], [161]. Moreover, some of the robotics-related frameworks found in reinforcement learning settings provide explanations for policies that a robot selects at a given time step [132], [156].
- **Contfactuals for recommendation.** Ghazimatin *et al.* propose a graph-based recommendation system in a counterfactual setup [84]. They obtain a counterfactual explanation by removing a minimal set of user actions so that the output recommendation changes.
- **Contfactuals for conflict resolution.** Mosca *et al.* introduce an argumentation-based framework for social network management [149]. They use contrastive explanations to answer critical questions about agent actions in the context of multi-user privacy conflict.

2) EXPLAINABILITY METHOD

All the frameworks generating contfactual explanations can be classified by their explainability method as either model-specific or model-agnostic. The former type of implementations is meant for explaining decisions of particular AI algorithms. The frameworks of the latter type generate explanations irrespective of the nature of the underlying algorithm. Table 7 presents the publications under study grouped in terms of the explainability method that they apply. The distribution of model-specific and model-agnostic explainability methods for generation of different types of contfactual explanation is shown in Fig. 10. Most frameworks deal with counterfactual model-agnostic methods.

- **Model-specific contfactual explanation generators.** Several model-specific frameworks generate counterfactuals to explain the output of decision trees [46], [80], [122], [155]. For instance, Fernández *et al.* [80] present a recursive algorithm which extracts counterfactuals in the form of contrast-class decision tree nodes. The relevance of the generated counterfactuals is then measured by calculating a variant of the Gower distance. The proposed metric penalizes the number of feature changes when traversing the tree so that sparsity is promoted. Alternatively, Sokol and Flash rely on the Manhattan distance measuring leaf-to-leaf distance in the tree to retrieve the most relevant counterfactuals [46], [155]. Designed specifically for decision trees, their ''Glass-box'' framework is argued to be easily extendable to capture the output of other logical (rule-based) models. Aguilar-Palacios *et al.* generate contrastive explanations using gradient boosted decision trees to forecast promotional sales [61]. The researchers make use of the weighted Euclidean distance to present the forecast as a contrast to the neighbouring vectorized promotions. Stepin *et al.* retrieve counterfactuals from a rule matrix where each rule is encoded in terms of all possible feature values [122]. Subsequently, the generated counterfactuals are ranked using a XOR-based distance to find the most relevant counterfactual pertinent to the given contrast class. This method is further extended to generating counterfactuals for fuzzy decision trees.

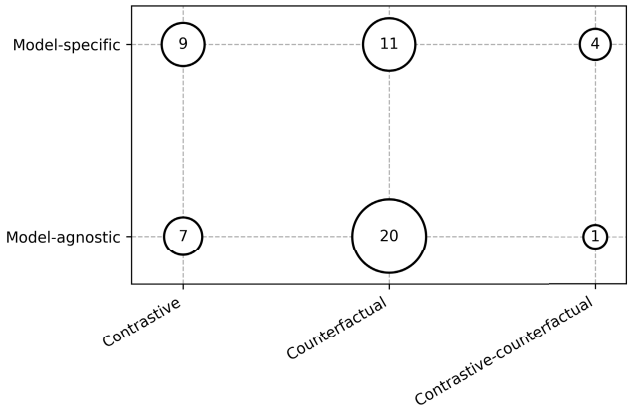

FIGURE 10. Numbers of frameworks grouped by explainability method with respect to the type of contfactual explanation generated.

A number of frameworks address specific properties of counterfactuals. Thus, Ustun *et al.* tackle the problem of actionability, i.e., constraining the generated counterfactuals in such a manner that the imposed changes ''do not alter immutable features'' and that they ''do not alter mutable features in an infeasible way'' [158]. To approach this problem, a mixed integer programming method is employed. Russell *et al.* adopt a similar approach to encompass continuous and discrete variables as well as the combination of the two [153]. The main focus of the work is however placed on assessing coherence and diversity of generated counterfactuals. In order to guarantee the coherence of the counterfactual data example used for explanation, an integer programming-based method is proposed. In addition, the generated counterfactual explanations are claimed to be diverse, as diversity constraints are applied iteratively to a set of candidate counterfactuals. However, this framework is limited to: (1) explaining predictions of only linear classifiers and (2) a simple structure of the textual explanation template.

A large number of frameworks are limited to explaining the output of particular models due to task-specific constraints. For instance, several explanation generators address computer vision tasks. Hendricks *et al.* bind

an input visual image with a paired textual counterfactual explanation generated by a recurrent neural network [86]. In their framework, a number of candidate explanations (both image-relevant and non-relevant) are generated, paired, and ranked. The best counterfactual explanation is then selected to be the most class-specific to the counterfactual image while being the most relevant to the input image. Goyal *et al.* argue that their model is more faithful by design, as it generates visual explanations directly from ''the target model based on the receptive field of the model's neurons'' [139].

Two model-specific frameworks are found in the context of video processing. For instance, Akula *et al.* [128] present an empirical study where input video frames are paired with the corresponding AND-OR graphs, i.e., compositional recursively defined graph-based knowledge representations capturing contextual information. The explanations based on such graphs are passed on to human subjects to evaluate the contrastive answers to the predefined questions. Alternatively, Kanehira *et al.* train a post-hoc explanatory model to justify a video classifier's output [91]. A counterfactual explanation is, in this case, dependent on how likely a selected region in the given frame is classified positive and not negative, hence all such regions are scored and normalized.

In accordance with the findings in the previous Section IV-C1, contfactual explanations have a great potential for automatic planning-related tasks. Most explanation generators meant for planning-based tasks are model-specific due to the problem- and approach-specific restrictions preventing them from being used for other AI challenges. For instance, Kim *et al.* employ a Bayesian probabilistic model for generating contrastive explanations [94]. Thus, the framework operates on a pair of plan traces defined in terms of linear temporal logic templates. The problem of obtaining contrastive explanations is designed as a Bayesian inference problem, with the posterior distribution to be maximized defined as the probability of a contrastive explanation given a set of positive and negative plan traces. Conversely, Sreedharan *et al.* consider the task of automatic analysis of counterfactual explanations in their ''Hierarchical Expertise-Level Modeling'' framework [156]. A robot provides a user with a plan for the next action to take. Then, the robot expects the user to respond with a set of foils. The robot's task is then to convincingly refute the foils by offering a minimal explanation for why the foils are not acceptable under the given circumstances. In addition, Chakraborti *et al.* formulate the multi-model planning problem as a tuple consisting of the planner's model of the problem and the corresponding human approximation thereof [132]. As plan explicability is reformulated in terms of its comprehensibility by an end-user. The robot's model is adapted to the updates of human's model of the problem.

The problem of contrastive explanation generation for planning is also found to be framed in the reinforcement learning setting. For instance, Sukkerd *et al.* formulate the planning problem as the shortest stochastic path problem and develop the corresponding problem solver to obtain a contrastive explanation [157]. Hence, their objective is to find an optimal policy ''that minimizes the expected cumulative cost of reaching a goal state over all closed policies''. The explanation is believed to justify the rejection of the policies alternative to the optimal one. In addition, Zhao and Sukkerd explain an autonomous system's behaviour modeling it as a Markov decision process [161]. Thus, a contrastive explanation is presented as a product of the analysis of the optimal policy at the next time step and an opposing policy on the basis of the objective values.

- **Model-agnostic contfactual explanation generators.** A large number of model-agnostic frameworks treat contfactual explanation generation as an optimization problem in a post-hoc manner. Wachter *et al.* design a generic counterfactual explanation framework to find the closest point to the test data example [6]. Fixing the optimal set of weights of a trained classifier, the objective function minimizes the distance between the nearest data points of opposing classes. Note that counterfactual data points can be synthesized artificially. The researchers suggest the use of the Manhattan distance weighted by the inverse median absolute deviation to calculate the proximity of a counterfactual to the input data example. Another case of counterfactual explanation generation regarded as an optimization problem is the ''Constrained Adversarial Examples'' framework [148]. Adversarial examples that could serve as the basis for the counterfactual explanation of the output of deep learning models are searched for with the aim of minimizing the loss with respect to the attributes (features) between the original and counterfactual data examples. The researchers attempt to find the best counterfactual explanation by minimizing the number of attributes changed. Furthermore, the gradient direction is constrained to ensure the ethical adequacy of the explanation generated. Dandl *et al.* [134] formulate counterfactual search as a multi-objective optimization problem using a distance metric for mixed feature spaces aiming to obtain sparse and most plausible counterfactuals. Labaien *et al.* generate contrastive explanations for time-series data [141]. The explanation generation is considered a two-fold optimization problem of finding pertinent positives and negatives. Pawelczyk *et al.* make use of an autoencoder architecture for a pretrained classifier performing counterfactual search in the nearest neighbor style [151].

  Model-agnostic frameworks are largely found to use decision trees as part of the reasoning mechanism instead of explaining their output. In contrast to the model-specific frameworks operating on decision tree output, Guidotti *et al.* employ decision trees as part

of reconstructing the reasoning behind any arbitrary classifier in a post-hoc fashion [140]. In their ''Local Rule-based Explanation'' framework, they generate a local neighbourhood for the given pre-classified data example using a genetic algorithm and subsequently train a decision tree on that newly obtained dataset to select a minimally distant foil within that local neighbourhood. Similarly, van der Waa *et al.* randomly sample or generate a data set in the neighbourhood local to the data point in question [159]. A decision tree is then trained to select the foil based on the minimum number of nodes between the original data point and the candidate foils. Furthermore, their ''Foil-Trees'' framework provides the methodological basis for perceptual-level contrastive explanation generation within the ''Perceptual-Cognitive Explanation'' framework [150]. Subsequently, the generated contrastive explanations are attributed to a specific group of users by means of ontology engineering at the cognitive level of the framework to make the explanations adaptive. In contrast, Martens and Provost argue that decision trees are an inadequate tool for representing, e.g., large documents [146]. Hence, they suggest the model-agnostic ''Search for Explanations for Document Classification'' algorithm for retrieving counterfactual explanations. However, it is only directly applicable to binary linear classifiers, whereas heuristics are proposed for non-linear models.

Several model-agnostic frameworks aim at measuring specific properties of contfactuals. Anjomshoae *et al.* [129] focus on contrastive explanations that maximize contextual importance and contextual utility. On the one hand, contextual importance measures the extent to which the input feature values affect the black-box algorithm's output. On the other hand, contextual utility testifies how favorable the values of the selected features are for a given decision. Thus, the context-based values are calculated for each feature used by a black-box model observing the changes in the output as the input varies across the range of all possible input values. Being based on model-agnostic and problem-independent concepts, this framework is shown to be universally applicable to various classification and regression algorithms. However, the scalability of such an algorithm is limited to the use-cases operating on a small number of features. A similar limitation is observed due to possibly high variability of the input. Laugel *et al.* raise the issue of justification for counterfactual explanation [144]. They argue that a synthesized counterfactual data point must be connected to the training data. Counterfactuals are selected from a local neighbourhood circling around the test example with the radius of the distance to the closest correctly predicted data point of a contrast-class. The candidate counterfactuals are then clustered, as the initial local neighbourhood is updated to become a more extensive hyperspherical layer, until it can no longer be extended. Laugel *et al.* [100] enhance the work on justified counterfactual explanations. They argue that the distance from the test instance to a counterfactual does not sufficiently measure counterfactual's relevance, as the counterfactual in question may appear disconnected from the ground-truth data. Thus, a counterfactual is deemed justified if it can be connected to an associated ground-truth data instance without crossing the decision boundary. Fernández *et al.* introduce the notion of counterfactual sets to enhance counterfactual diversity [81]. They explain random forest predictions by fusing different tree predictors so that the resulting counterfactual set contains the most relevant counterfactual. The other neighboring counterfactuals serve to diversify the output explanation. Mothilal *et al.* are also concerned with counterfactual diversity [105]. They design a loss function with a diversity metric over the generated counterfactuals to provide end-users with multiple relevant counterfactual explanations. Kusner *et al.* propose a causal model to assess the so-called ''counterfactual fairness'' [55]. It is worth noting that counterfactuals are presented in the form of conditional distributions and not structural equations despite the fact that the causal model employed follows Pearl's formalism [37].

Similarly to the model-specific frameworks, numerous model-agnostic explanation generators are found to be task-specific. In computer vision-related classification tasks, Chang *et al.* find the smallest region in the image whose substitution would change the classifier's prediction [133]. They employ a generative model to construct a saliency map while masking the other regions of the input image. Similarly, Dhurandhar *et al.* address an optimization problem over a perturbation variable to produce a contrastive explanation for the image classification task [135]. However, the proximity of the selected counterexample to the test point is, in this case, guaranteed by using an autoencoder.

3) OUTPUT REPRESENTATION

The considered frameworks output contfactual explanations in several ways. Depending on the problem considered, contfactuals are presented in the form of: (1) intervals or specific values of the appropriate feature values whose alteration would have changed the output (i.e., numerical or feature-based output); (2) single- or multiple-sentence coherent text (i.e., linguistic output); (3) specific regions in the input image (i.e., visual output); or (4) a multi-modal combination of (some of) the above (see Table 8). As depicted in Fig. 11, most frameworks focus on numerical counterfactual output.

- **Numerical (feature-based) contfactual explanation.** Numerical values (or intervals of values) associated to the most relevant features usually explain the behavior of AI algorithms. They can be represented as logical formulas [67], [93], [94] or in tabular form reflecting necessary changes to affect the decision [55], [56], [61],

TABLE 8. A classification of the contfactual explanation frameworks by output representation.

Output representation	Publications
Numerical (feature-based)	[55], [56], [61], [67], [80], [81], [93], [94], [100], [105], [132], [136], [140], [147], [148], [151], [152], [154], [158], [160]
Linguistic	[6], [48], [84], [122], [128], [131], [146], [149], [153], [156], [157], [159], [161]
Visual	[49], [130], [133]–[135], [139], [141], [142], [144]
Multi-modal	[46], [86], [91], [129], [137], [138], [143], [145], [150], [155]

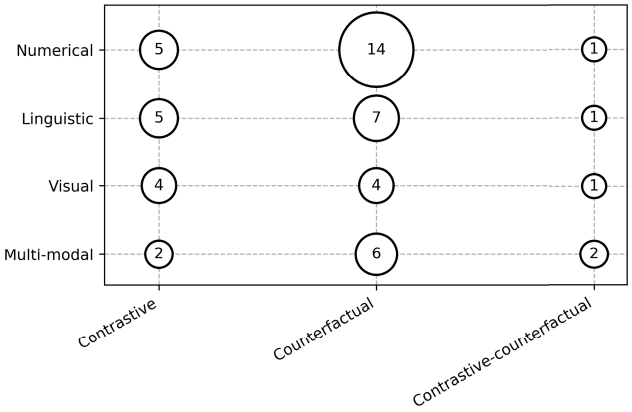

FIGURE 11. Numbers of frameworks grouped by output representation with respect to the type of contfactual explanation generated.

[81], [105], [136], [147], [148], [151], [154], [158], [160]. They can be extracted from interpretable feature-value pairs as a result of pruning in the search space [132]. In addition, they can replicate the internals of the classifier's structure, e.g., in the form of decision tree nodes or rules [80], [140], [152].

- **Linguistic contfactual explanation** is a piece of grammatical single- or multiple-sentence text in natural language. Single-sentence textual explanations combine a textual description with explicitly stated numerical feature values [6]. Such explanations suggest feature-value based instructions [48], [122], [131] or alternative actions for a possible output change [84], [149]. They also answer end-user's inquiries with respect to the automatic decision in question [128], [159]. In contrast, multiple-sentence explanations provide end-users with specific details for the given decision [153], [156], [157], [161] or explain multiple decisions at once [146].
- **Visual contfactual explanation.** On the one hand, visual explanations for non-visual input data (i.e., datasets containing continuous or categorical feature-value pairs) plot feature-value pair dependencies [134], [141], [142]. On the other hand, visual input data (i.e., images) are associated with saliency maps [133] or explained by contrastive patterns between the given data example and that of an opposing class in one iteration [130], [144] or a series thereof [49]; by depicting critical regions absent in the input data example that determine what lacks in the image to be classified differently [135]; or by visualizing spatial regions associated to data examples of opposed classes [139].
- **Multi-modal contfactual explanation** is a combination of numerical and/or linguistic and/or visual explanations. Multi-modal explanations are claimed to enhance human-robot interaction [46]. They are often selected to be the most appropriate where the problem addressed is concerned with pairing a computer vision problem with a natural language processing task such as object detection and language grounding [86]. In addition, visual-linguistic explanations identify counterfactuality in videos [91] and allow for dialogic interaction [150]. Such hybrid counterfactuals (in terms of their output representation) may as well complement each other while addressing the same task. For instance, Gomez *et al.* visualize the generated explanations in the form of bar plots combining them with explicitly stated numerical values [138]. Alternatively, Liu *et al.* combine feature importance bar plots with visual input and output [145]. While a textual explanation summarizes the degree of importance of the selected features, a visual explanation may present contextual in-method metrics that justify the classifier's reasoning [129]. Contfactual explanations, as a mixture of tabular and visual output representations, appear also in an augmented reality framework [137]. Nevertheless, explanations of different modalities are not necessarily merged. To ensure the universality of the proposed approaches, specific feature values are presented for tasks with datasets containing only continuous features, i.e., where the same method is used to output images for a handwritten digit classification problem [143]. Finally, interaction with users can be enhanced by means of voice-based explanations combined with textual explanations [46], [155].

4) EVALUATION METHOD

Evaluation of generated contfactual explanations is an issue of main concern. Unfortunately, despite an increasingly expanding use of contfactual explanations, no uniform set of evaluation methods has been adopted so far. Hence, it is worth taking a look at evaluation methods from other generation-oriented sub-areas of AI. For instance, it is common to distinguish between intrinsic and extrinsic evaluation methods in natural language generation [163]. Intrinsic evaluation implies assessing the performance of a natural language generation system (or its modules) as an isolated unit. In contrast, extrinsic (task-based) methods are designed to estimate how successfully the system performs with respect to an external task. In addition, Gatt and Krahmer make a distinction between "objective" (automatic, corpus-based) and "subjective" (human judgements) metrics [164]. Objective metrics include (but are not limited to) precision- and/or recall-oriented scores, number of insertions/deletions/substitutions,

TABLE 9. A classification of the contfactual explanation generators by evaluation method.

Evaluation method	Publications
None	[6], [46], [49], [67], [93], [129]–[131], [136], [145], [149], [150], [153], [155], [157], [158], [161]
Subjective	[56], [128], [137]
Objective	[55], [61], [80], [81], [86], [91], [94], [100], [105], [122], [132]–[135], [138]–[144], [146]–[148], [151], [152], [154], [156], [159], [160]
Hybrid	[48], [84]

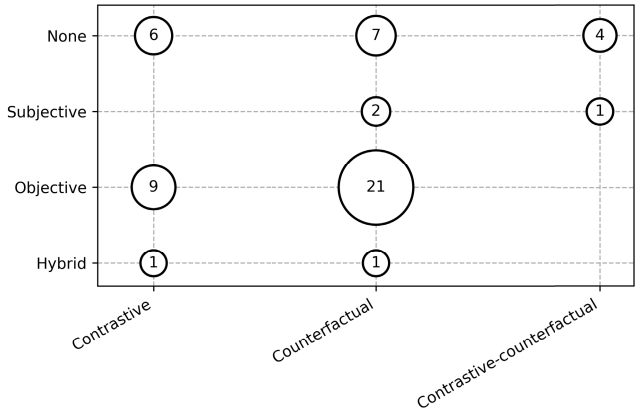

FIGURE 12. Numbers of frameworks grouped by evaluation method with respect to the type of contfactual explanation generated.

etc. In turn, subjective metrics measure readability, accuracy, relevance of the generated text, as perceived by humans. Thanks to their methodological universality, they can be extrapolated to other (non-linguistic) modalities of generated explanations (e.g., numerical, visual, or multi-modal) and are therefore used to form the basis of the evaluation method classification in this review. It is worth noting that all the considered frameworks are evaluated by means of intrinsic (either subjective or objective) metrics. Hence, we only make a clear distinction between subjective and objective evaluation methods in this study (see Table 9 for details). A distinction between the use of different types of evaluation metrics can be seen in Fig. 12. It is easy to appreciate how most frameworks deal with objective evaluation of counterfactuals. Let us give further details below, regarding the four groups of publications in Table 9.

- **No evaluation details provided.** 17 out of 52 (32.69%) of the considered publications do not evaluate their frameworks, i.e., neither automatic metrics for contfactual explanation generation are suggested nor a human evaluation survey is presented in such publications. However, whereas certain publications do not provide any specific evaluation method, some do stress that human evaluation should be encouraged to estimate the quality and effectiveness of the generated counterfactuals [46], [150], [153].
- **Subjective evaluation.** The subjective methods include human preferences for certain types of contfactual explanation over others. For instance, Akula *et al.* show that that contrastive explanation-seeking questions are in general better answered by means of contfactual explanations [128]. They classify contrastive questions in the following 10 categories suggesting the template questions for arbitrary objects x, x_1, and x_2 (all being of some class X) and y, y_1, and y_2 (all being of some other class Y):
  - WH-X: "Why x rather than not x?";
  - WH-X-NOT-Y: "Why x rather than y?";
  - WH-X1-NOT-X2: "Why x_1 rather than x_2?";
  - WH-NOT-Y: "Why not y?";
  - NOT-X: "Is it x rather than not x?";
  - NOT-X1-BUT-X2: "Is it x_1 rather than x_2?";
  - NOT-X-BUT-Y: "Is it x rather than y?";
  - DO-X-NOT-Y: "What if it is x rather than y?";
  - DO-NOT-X: "What if it is not x?".
  - DO-X1-NOT-X2: "What if it is x_1 and not x_2?"

  It is worth noting that 6 out of 10 question types (WH-NOT-Y, NOT-X, NOT-X1-BUT-X2, NOT-X-BUT-Y, DO-NOT-X, and DO-X1-NOT-X2) matched with automatically generated contfactuals are shown to be highly preferred to factual explanations.
  In addition, Ferrario *et al.* propose an augmented reality-based setting to favor interactivity and facilitate explaining ML algorithm output to non-experts [137].
  However, the two aforementioned studies [128], [137] lack an evaluation of the quality of the generated contfactual explanations themselves.
  Lucic *et al.* asked 75 subjects to judge interpretability, actionability, and trustworthiness of the generated contfactual explanations [56]. They concluded contfactual explanations are highly interpretable and actionable. In addition, they help users understand why the model makes large errors while solving a regression problem but do not support users' trust in the model's output.
  In addition, Hendricks *et al.* provide results of human evaluation for the generated explanations [86]. However, these only include evaluations for factual explanations and are therefore excluded from the taxonomy group being discussed.
- **Objective evaluation.** A majority of the researchers propose objective (automatic) methods for evaluating automatically generated contfactuals. A number of the frameworks are evaluated by means of accuracy-based metrics [61], [91], [94], [159]. Kanehira *et al.* propose one accuracy-based evaluation metric for visual and linguistic explanations each: negative class accuracy and concept accuracy, respectively [91]. Negative class accuracy estimates the quality of the visual explanation as the ratio of the probability of the contrast class after the image region in question is masked out. In turn, concept accuracy estimates how compatible the output linguistic explanation is to its visual counterpart. It is calculated as the intersection over union between a given region

and all bounding boxes in the image. Kim *et al.* define the domain-specific accuracy for the automatic planning problem of unique contrastive explanations as a sum of the number of traces in a positive set of traces satisfying a contraint for a contrastive explanation and those of the negative set where the constraint is unsatisfied over all plan traces [94]. The consistency of output explanations is otherwise shown by measuring their accuracy on the basis of mean error, mean absolute error, and mean absolute percentage error [61].

An extensive number of evaluation methods are found to be strictly task- or approach-specific. Hendricks *et al.* measure word detection (i.e., which words are not image relevant by holding out one word at a time from the sentence to determine the least relevant word in the explanation) and word correction (i.e., a number of replacements of the foiled word with words from a set of target words) [86]. Similarly, Martens and Provost estimate explanation complexity by calculating the average number of words in the shortest explanation and problem complexity according to the overall number of generated explanations [146]. Fernández *et al.* evaluate the relevance of the generated counterfactuals measuring a Gower distance-based metric in comparison with the number of feature changes and minimum distance (in terms of decision tree nodes) between the leaves in the given decision tree classifier [80]. Van der Waa *et al.* evaluate the generated explanations by means of such model-specific metrics as the average length of the explanation in terms of decision tree nodes and the F_1-score of the foil-tree on the test set compared to the model's output [159]. Kusner *et al.* estimate counterfactual fairness on the basis of the density of the predicted data for their causal models [55]. Laugel *et al.* claim that understandability of the generated explanations can be estimated by means of their sparsity defined as the number of non-zero coordinates of the explanation vector [143]. Moore *et al.* measure the number of solutions, the distances to the nearest training set data points, and the transferability of the generated counterfactuals to other datasets and classifiers [148]. Sreedharan *et al.* calculate the number of predicates that are used to generate the model lattice [156]. Similarly, Chakraborti *et al.* calculate the number of nodes in the search space remaining after pruning [132]. Goyal *et al.* report how often the discriminative regions lie inside the test data example segmentations as well as relevant specific key regions [139]. Labaien *et al.* calculate the number of changes to switch from the original to the selected contrastive sample following the dataset constraints [141]. To estimate faithfulness of the generated counterfactuals, Pawelczyk *et al.* suggest calculating the so-called degree of difficulty of a counterfactual suggestion to measure how costly it is to achieve the state of the given suggestion [151]. Aiming to provide realistic counterfactuals, Sharma *et al.* introduce the counterfactual explanation robustness-based score defined as the expected distance between the input instances and their corresponding counterfactuals [154]. In addition, the generated counterfactuals are inspected in terms of fairness which is calculated as the expected distance between the input and a counterfactual over distinct values for a specified feature set. Merrick and Taly evaluate output explanations in terms of mean feature attributions to show the importance of relevant references [147]. Gomez *et al.* evaluate counterfactuals in terms of data distribution, feature importance, as well as possible and actionable changes to the input [138]. Dandl *et al.* use the hypervolume indicator metric to estimate the quality of the estimated Pareto front during counterfactual search [134]. In addition, Chang *et al.* measure the weakly supervised localization error for an image detection task – the intersection-over-union ratio over 0.5 with any of the ground truth bounding boxes and the saliency metric, i.e., ''the log ratio between the bounding box area and the in-class classifier probability after upscaling'' [133].

Several metrics can be extended to be applied to other approaches. Lash *et al.* estimate how much the probability of a given prediction reduces given a feature perturbation as determined by a contrastive explanation [142]. Dhurandhar *et al.* employ the concept of pertinent positives (i.e., ''factors whose presence is minimally sufficient in justifying the final classification'' [135]) and pertinent negatives (i.e., ''factors whose absence is necessary in asserting the final classification'') to evaluate factuals and counterfactuals, respectively, for a given classification task. Both types of evaluation methods highlight the features supporting evidence as formulated in the contrastive explanation on the basis of the values that a perturbation variable takes on. Fernández *et al.* evaluate counterfactuals in terms of the average of the pairwise distances based on the feature type and the percentage of valid counterfactuals [81].

Mothilal *et al.* stress that counterfactuals should be evaluated in terms of validity (i.e., whether a generated counterfactual really leads to a different outcome), proximity (i.e., feature-wise distance between the original and counterfactual samples), sparsity (i.e., number of features differing in the original and counterfactual samples), and diversity (i.e., feature-wise distance between each pair of counterfactuals) [105]. Similarly, Stepin *et al.* calculate factual and counterfactual explanation length to estimate conciseness of the generated explanations [122]. They also compute the number of counterfactuals and their best minimal distance to the factual explanation to assess the relevance of counterfactuals. Rajapaksha *et al.* consider coverage (as an indicator of representativeness of a rule for a given dataset), confidence (i.e., the percentage of instances in the dataset which contain the consequent and antecedent together over the number of instances which only contain the antecedent), lift (i.e., an association between antecedent

and consequent), leverage (i.e., the observed frequency between the antecedent and consequent), and the number of features in explanation for evaluating their framework against other rule-based methods [152]. Also, White and Garcez reintroduce fidelity to the underlying classifier on the basis of distance to the decision boundary [160].

In addition, some of the model-agnostic frameworks [140], [159] allow for measuring how well the output of black boxes (i.e., actual output to be explained) and grey boxes (i.e. interpretable intermediate predictors) mimic the local neighbourhood (i.e., fidelity) and the data example to be explained (i.e., hit). Laugel *et al.* measure how justified counterfactuals are by averaging a binary score (one if the explanation is justified following the proposed definition, zero otherwise) over all the generated explanations [100], [144].

It is worth noting that the run-time of explanation generation algorithms is reported in addition to the evaluation metrics for several frameworks [132], [139], [146], [152], [156], [159].

- **Hybrid evaluation.** Two frameworks are evaluated in terms of both automatic metrics and human judgments. Ghazimatin *et al.* calculate explanation length to discuss comprehensiveness of explanations as well as estimate their usefulness and credibility by surveying 500 subjects [84]. In addition, Le *et al.* compute fidelity, conciseness, information gain, and influence [48]. Automatic metrics are complemented with a user study on intuitiveness, friendliness, comprehensibility, and understandability of generated explanations.

D. LINKS BETWEEN THEORETICAL AND PRACTICAL CONTRIBUTIONS TO CONTFACTUAL EXPLANATION GENERATION (ANSWER TO RQ_3)

We find that only few of the existing computational frameworks are grounded on theories of contfactual explanation. Indeed, only 13 out of 113 studies (11.50%) were present in both of the pools of primary studies related to RQ_1 and RQ_2. Table 10 summarizes the characteristics of such theoretically grounded contfactual explanation generation frameworks.

Moreover, only 3 out of the 13 (23.08%) studies interpret the insights from the theoretical foundations to propose their own contfactual explanation definition for problem-oriented purposes. Kean states that ''explanation in artificial intelligence is based on the inference of deduction'' [93]. He complements a deductive evidence-based explanation with a redefined abductive contrastive explanation drawing parallels to the ''inference to the best explanation'' [103]. He models Lipton's theoretical framework distinguishing two types of contrastive explanation: non-preclusive (i.e., non-restrictive) and preclusive. The key aspect distinguishing the two types of contrastive explanation is in regard to how a model explains the contrast given an explanation-seeking question. Thus, a non-preclusive contrastive explanation is ''irrelevant to the model of explaining the contrast'' being ''necessary in the model of explaining the question''. On the contrary, a preclusive contrastive explanation is assumed to be restricted by a negated model of the contrast.

Aguilar-Palacios *et al.* [61] refer to Lipton's definition of contrastive explanation [12]. Referring to Pearl [37], Bertossi redefines causal explanation in the context of XAI to be ''a set of feature values for the entity under classification that is most responsible for the outcome'' [67].

The rest of works redefine contfactual explanation on the basis of the problem-specific constraints without explicitly referring to the theoretical foundations described in Section IV-B. Driven by the task of automatic planning, Kim *et al.* define a contrastive explanation to be a constraint satisfied by a specific set plan traces [94]. Fernández *et al.* define a counterfactual to be a set of feature changes that turn the given data example to be classified differently [80]. Whereas this definition is applicable to the classification problem in general, the applicability of the framework is restricted to decision trees only. Similarly, Hendricks *et al.* explain visual concepts for the image classification task on the basis of the so-called counterfactual evidence (i.e., an attribute discriminative enough for another class of objects in the image absent in the given image) [86]. Ghazimatin *et al.* [84] define a counterfactual on the basis of their model's internal structure: an explanation is deemed counterfactual if after removing the edges from the recommendation graph, the user receives a different top-ranked recommendation. In addition, Kanehira *et al.* only specify the linguistic form of a counterfactual explanation without defining it explicitly [91].

Finally, there is a number of marginal interpretations of contfactuals among the RQ_3-related studies. Laugel *et al.* denote a counterfactual as a specific data instance that changes the algorithm's prediction [100]. Poyiadzi *et al.* denote a counterfactual to be ''the new state of the object'' [49]. Nevertheless, the most commonly acceptable definition of a contfactual in the observed RQ_3-related studies states that a contfactual explanation is a set of minimal feature modifications that makes the model change the prediction [81], [105], [122].

V. DISCUSSION

The findings show that a large body of research has been elaborated on theoretical accounts of contrastive, counterfactual, and contrastive-counterfactual explanation. In addition, the topic has recently attracted attention from researchers in XAI (see Fig. 13). Thus, 50 out of the 52 considered state-of-the-art contfactual explanation generation frameworks (96.15%; RQ2) have been developed from 2017 to 2020.

The results of the study, in relation to RQ_1, show that a majority of the considered theoretical accounts of contfactual explanation (49 out of 74; 66.22%) speculate on the causal nature of explanation. However, whereas most researchers in philosophy of science have mainly used the concept of counterfactuality to explain causal relations between entities in question, causal inference is poorly addressed in the

TABLE 10. A summary of characteristics of theoretically grounded computational frameworks for contfactual explanation generation. CT stands for contrastive explanation, CF means counterfactual explanation, and CT-CF is contrastive-counterfactual explanation.

Reference	Type of contfactual explanation			Relatedness to causality			Computational framework properties			
	CT	CF	CT-CF	Causal	Non-causal	Hybrid	AI problem	Explainability method	Output representation	Evaluation method
Aguilar-Palacios et al. (2020) [61]	✓			✓			regression	model-specific	numerical	objective
Bertossi (2020) [67]		✓		✓			classification	model-agnostic	numerical	none
Fernández et al. (2019) [80]		✓			✓		classification	model-specific	numerical	objective
Fernández et al. (2020) [81]		✓			✓		classification	model-agnostic	numerical	objective
Ghazimatin et al. (2020) [84]		✓		✓			recommendation	model-specific	linguistic	hybrid
Hendricks et al. (2018) [86]		✓			✓		classification	model-specific	multi-modal	objective
Kanehira et al. (2019) [91]		✓			✓		classification	model-specific	multi-modal	objective
Kean (1998) [93]	✓			✓			knowledge engineering	model-specific	numerical	none
Kim et al. (2019) [94]	✓			✓			planning	model-specific	numerical	objective
Laugel et al. (2020) [100]		✓			✓		classification	model-agnostic	numerical	objective
Mothilal et al. (2020) [105]		✓				✓	classification	model-agnostic	numerical	objective
Poyiadzi et al. (2020) [49]			✓		✓		classification	model-agnostic	visual	none
Stepin et al. (2020) [122]		✓			✓		classification	model-specific	linguistic	objective

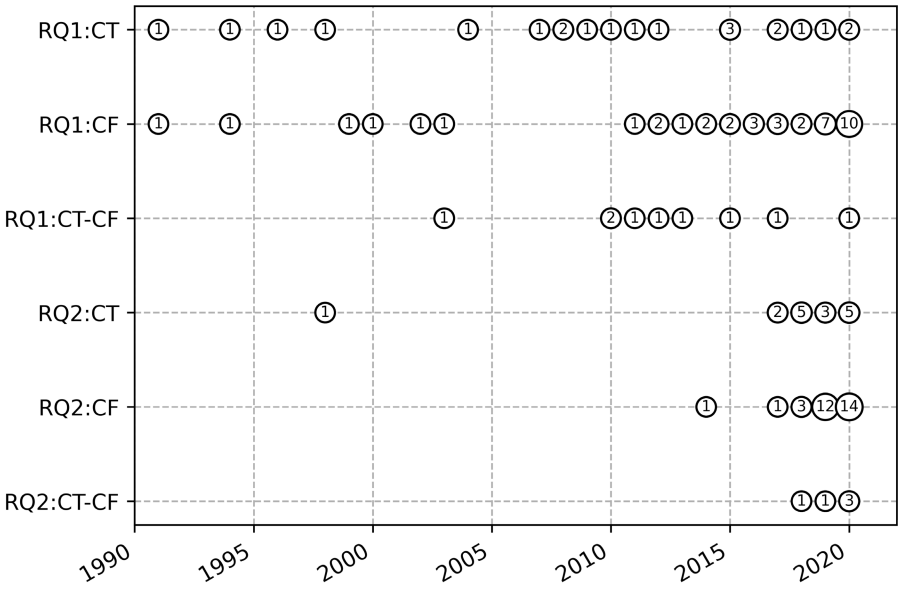

FIGURE 13. Numbers of theoretical and computational contfactual explanation generation frameworks grouped by year of publication. For illustrative purposes, only the studies published from January 1990 to September 2020 are displayed.

pool of publications concerning computational frameworks of contfactual explanation generation. Kean directly refers to a causal account of contrastive explanation to address the problem of abductive reasoning [93]. In addition, Lucic *et al.* [56] explicitly specify that their method is based on previous work on philosophical accounts of contrastive explanation [12] as well as on causal attribution [165], [166]. Kusner *et al.* [55] make use of causal inference models and the corresponding tools provided by Pearl [37]. They assess how discriminatory the generated counterfactual explanations are for the given classification task output. On the other hand, Bertossi redefines the concept of causal explanation [67]. Following a causal account of contfactual explanation, Fernández *et al.* introduce weakly causal irreducible counterfactual explanation [136]. As most of the current ML-tasks are centered around singling meaningful patterns out from unstructured data, establishing causal relations appears to be among the AI problems that are yet to attract global attention. This partly explains why most of the modern contfactual explanation generators focus on feature perturbation when searching for the most relevant contfactuals and not establishing causal relations between them.

At the same time, the other computational frameworks are primarily non-causal. Furthermore, a strikingly low number of such frameworks appear to be rooted in theoretical accounts of explanation due to an imbalance in favor of causality-oriented theoretical accounts. However, the amount of publications for RQ_3 may be somewhat misleading, as contfactual explanations are often redefined without specifically referring to theoretical contributors in explanation. This is

hypothesized to be due to the problem-specific necessities ignored in previous theoretical works from different branches of science. For instance, Laugel *et al.* emphasize that the minimal perturbations required to change the predicted class of a given observation enable a user to understand which features locally impact the prediction and therefore how it can be changed [144]. In this interpretation, counterfactual explanations are conceptually most similar to counterfactuals as defined by Lewis [36]. Indeed, while the concept of "the closest possible world" does not always appear in the related publications, it turns out to be implicitly wired in almost all works. Instead, other considered frameworks do not appeal directly to any of the theoretical accounts of explanation addressed in RQ_1. Hence, it is worthwhile taking a look at how contrastive and counterfactual explanations are redefined in the frameworks not appearing in Section IV-D.

In general, a consensus among researchers has been observed on how contfactual explanations are defined irrespective of the theoretical framework proposed by individual authors. In humanities and social sciences, a major difference in various contfactual theories of explanation is observed to concern the causal nature of explanation and its extrapolation to non-causal cases. In computer science and AI, notions of counterfactual and contrastive explanation are found to be the most dissimilar when applied to non-overlapping problems. Nevertheless, the corresponding line of research in AI makes little use of the rich theoretical background accumulated by now. While some rule-based approaches used in expert systems are justified theoretically (e.g., see [93]), newly emerging tasks present novel challenges for theorists and call for updating the theories developed so far.

More precisely, ML-specific contfactual explanations are designed to answer the question: "Why was the outcome Y observed instead of Y'?" [148]. Anjomshoae *et al.* define finding a contrastive explanation as "contrasting instance against the instance of interest" [129]. Fernández *et al.* specify that a counterfactual is generally regarded as a hypothetical instance similar to an example whose explanation is of interest but with a different predicted class [80]. Also, counterfactual explanations "show a difference in a particular scenario that causes an algorithm to change its mind" [155].

As a majority of the considered frameworks are designed for tackling classification problems, contfactual explanations operate on the notion of a contrast-class (e.g., see [155]) answering the question: "How is the prediction altered when the observation changes, given a classifier and an observation?" Furthermore, these changes are normally expected to be minimal [143].

However, certain application domains as well as the selection of a classifier require researchers to redefine contfactuals imposing task-dependent constraints, which makes it nearly impossible to connect them to any of the existing theories of contfactual explanation. For instance, Martens and Provost define a contrastive explanation for a document classification task to be a minimal set of words such that removing all words within this set from the document changes the predicted class from the class of interest [146]. In addition, Guidotti *et al.* reformulate a counterfactual to be a set of split conditions of a decision tree describing the minimal number of changes in the feature values of a test example [140]. In image classification, it is found necessary to detect specific regions in the given test image. For this type of tasks, the contrastive explanation-seeking question is formulated as follows: "Which parts of the image, if they were not seen by the classifier, would most change its decision? or which inputs, when replaced by an uninformative reference value, maximally change the classifier output?" [133], [139]. Similarly, Dhurandhar *et al.* ask what should be minimally and necessarily present and absent in the given image to justify its classification [135]. Alternatively, counterfactuals are viewed as "solutions that are guaranteed to map back onto the underlying data structure" [153]. Redefined contrastive explanations are also found in the domain of robotics and automatic planning. According to Sukkerd *et al.*, a contrastive explanation answers the question why a generated behavior is optimal with respect to the planning objectives of an autonomous system [157]. Alternatively, contrastive explanations are used to answer why-not questions about the system's behavior in which the consequences of the counterfactuals in question are pointed out [161].

In addition, the nature of the explanation-seeking questions for computational frameworks deserves further discussion. Sokol and Flash distinguish three types of counterfactual explanations: (1) a plain counterfactual ("Why?") generated as the shortest possible class-contrastive counterfactual; (2) a counterfactual explanation not conditioned on the indicated feature(s) ("Why despite?"); and (3) a (partially) fixed counterfactual explanation ("Why given?") which is conditioned on a predetermined set of features [46]. Hilton proposes different types of contrastive questions such as: (1) "Why X rather than not X?"; (2) "Why X rather than the default value for X?" and (3) "Why X rather than Y?" [166]. Following this distinction, Akula *et al.* extend this set of contrastive questions to formulate ten contrastive question types for counterfactual explanation generation [128] (see Section IV-C4). Alternatively, only linguistic templates for such explanations are defined without any theoretical grounding in accordance with any accounts described in Section IV-B. For instance, Sokol and Flash define a counterfactual explanation to be a piece of text following the template: "The prediction is ⟨prediction⟩. Had a small subset of features been different ⟨foil⟩, the prediction would have been ⟨counterfactual prediction⟩ instead" [155].

Remarkably, a wide range of frameworks favor automatic evaluation methods. Thus, they rarely place the end-user in the center of the explanation evaluation process. However, we find an increasing number of interactive frameworks that attempt not only to present the automatically generated explanations to the end-user but also interact with him or her [46], [150], [155]. Promoting interactivity (e.g., by engaging the end-user to participate in an explanatory dialogue with the

system) is expected to make explanation social and further increase user's trust in the system's output.

VI. CONCLUDING REMARKS

In this work, we made two main contributions. First, we provided readers with an overview on theoretical accounts of contrastive, counterfactual, and contrastive-counterfactual explanation as well as frameworks of automatic generation thereof. This overview was based on a systematic literature review. This research methodology allowed us to carry out an unbiased reproducible study from an interdisciplinary topic-specific search in reputable and trustworthy sources. Second, we proposed a two-level taxonomy of the aforementioned types of explanation with the aim of providing a well-established tool that allows us to jointly analyze different proposals in this research field. As a result, this taxonomy facilitates the comparison of approaches and publications. We expect that it raises awareness in researchers in the community about main categories (definitions, practical frameworks, etc.) and subcategories (causal, non-causal, etc.) in the taxonomy. Moreover, we hope that it helps properly characterize the body of work and leverages a deeper collaboration and citation among similar related work.

The findings allow us to draw the following remarks. Contrastive and counterfactual explanations are found to be in great demand across various sub-fields of AI. Mainly applied to a wide range of tasks in computer vision and natural language processing, they present a powerful tool that enhances human-machine interaction and allows for further personalization of the output generated by various AI algorithms, including ML-based black-box algorithms.

In our systematic review, we introduced the term ''contfactual explanation'' to unify the aforementioned families of explanation to subsequently analyse the existing approaches to them from three points of view.

First, we investigated theoretical accounts of contfactual explanation to infer the similarities and differences among the existing theoretical approaches. Contfactuals are found to address both causal and non-causal dependencies. Hence, being a significant challenge, unification of causal and non-causal explanatory engines within a contfactually-driven framework opens new perspectives for the XAI community.

Second, existing computational frameworks for contfactual explanation generation have been inspected. Despite the fact that the notion of contfactual explanation is found to be highly task- and domain-specific, the most commonly accepted definition of a contfactual explanation in the context of XAI refers to a minimal set of feature modifications that makes the model change the prediction. A crucial shortcoming relevant to the inspected frameworks is a lack of standardization with respect to the evaluation methods. While designing a set of standard evaluation metrics is particularly complicated due to a different nature of the tasks that these explanations serve, this is hypothesized to be among major factors preventing researchers from making faster progress in solving the problem of contfactual explanation generation, as it complicates a fair evaluation of newly developed frameworks against the state-of-the-art equivalents. Furthermore, as automatically generated explanations are meant to be user-oriented, more effort is needed to include end-users in the process of assessing the generated explanations.

Third, a synergy between the related theories and computational frameworks has been investigated. We find that a gap between philosophical accounts of contfactual explanation to scientific modeling and ML-related concepts makes the theoretical frameworks poorly applicable to XAI. In addition, the existing methodological differences affect greatly the definition of contfactual explanation found across various approaches. In fact, definitions vary depending on domains of science and even approaches used for solving specific tasks.

Finally, we believe a joint interdisciplinary effort of researchers from both humanities and computational sciences can be particularly fruitful for further progress in contfactual explanation generation.

ACKNOWLEDGMENT

Ilia Stepin is an *FPI* Researcher (PRE2019-090153). Jose M. Alonso is a *Ramon y Cajal* Researcher (RYC-2016-19802). Alejandro Catala is a *Juan de la Cierva* Researcher (IJC2018-037522-I).

REFERENCES

[1] M. Attaran and P. Deb, ''Machine learning: The new 'big thing' for competitive advantage,'' *Int. J. Knowl. Eng. Data Mining*, vol. 5, no. 4, pp. 277–305, 2018.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, '''Why should i trust you?': Explaining the predictions of any classifier,'' in *Proc. 22nd Int. Conf. Knowl. Discovery Data Mining (ACM SIGKDD)*. New York, NY, USA: Association Computing Machinery, 2016, pp. 1135–1144.

[3] N. Wang, D. V. Pynadath, and S. G. Hill, ''Trust calibration within a human-robot team: Comparing automatically generated explanations,'' in *Proc. 11th ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, Piscataway, NJ, USA: IEEE Press, Mar. 2016, pp. 109–116.

[4] C. Rudin, ''Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,'' *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019.

[5] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, ''Explainable agents and robots: Results from a systematic literature review,'' in *Proc. 18th Int. Conf. Auto. Agents MultiAgent Syst. (AAMAS)*. Richland, SC, USA: International Foundation Autonomous Agents Multiagent Systems, 2019, pp. 1078–1088.

[6] S. Wachter, B. Mittelstadt, and C. Russell, ''Counterfactual explanations without opening the black box: Automated decisions and the GDPR,'' *Harvard J. Law Technol.*, vol. 31, no. 2, pp. 841–887, 2018.

[7] G. Schurz, ''Scientific explanation: A critical survey,'' *Found. Sci.*, vol. 1, no. 3, pp. 429–465, Sep. 1995.

[8] J. C. Pitt, *Theories of Explanation*. Oxford, U.K.: Oxford Univ. Press, 1988.

[9] C. B. Cross, ''Explanation and the theory of questions,'' *Erkenntnis*, vol. 34, no. 2, pp. 237–260, Mar. 1991.

[10] T. Lombrozo, ''Explanation and abductive inference,'' in *Oxford Handbook of Thinking and Reasoning*. Oxford, U.K.: Oxford Univ. Press, 2012, pp. 260–276.

[11] T. Miller, ''Explanation in artificial intelligence: Insights from the social sciences,'' *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019.

[12] P. Lipton, ''Contrastive explanation,'' *Roy. Inst. Philosophy Suppl.*, vol. 27, pp. 247–266, Mar. 1990.

[13] C. V. F. Bas, *The Scientific Image*. Oxford, U.K.: Oxford Univ. Press, 1980.

[14] F. Sørmo, J. Cassens, and A. Aamodt, ''Explanation in case-based reasoning—Perspectives and goals,'' *Artif. Intell. Rev.*, vol. 24, no. 2, pp. 109–143, 2005.

[15] N. J. Roese, "Counterfactual thinking," *Psychol. Bull.*, vol. 121, no. 1, pp. 133–148, 1997.
[16] R. M. J. Byrne, "Mental models and counterfactual thoughts about what might have been," *Trends Cognit. Sci.*, vol. 6, no. 10, pp. 426–431, Oct. 2002.
[17] S. Chin-Parker and A. Bradner, "A contrastive account of explanation generation," *Psychonomic Bull. Rev.*, vol. 24, no. 5, pp. 1387–1397, Oct. 2017.
[18] R. Wenzlhuemer, "Counterfactual thinking as a scientific method," *Historical Social Res.*, vol. 34, no. 2, pp. 27–56, 2009.
[19] R. Folger and C. Stein, "Abduction 101: Reasoning processes to aid discovery," *Hum. Resour. Manage. Rev.*, vol. 27, no. 2, pp. 306–315, Jun. 2017.
[20] C. Boutilier and V. Beche, "Abduction as belief revision," *Artif. Intell.*, vol. 77, no. 1, pp. 43–94, Aug. 1995.
[21] P. Lipton, *Inference to the Best Explanation*, 2nd ed. Evanston, IL, USA: Routledge, 2004.
[22] R. M. J. Byrne, "Counterfactual thought," *Annu. Rev. Psychol.*, vol. 67, pp. 135–157, Jan. 2016.
[23] R. M. J. Byrne, "Cognitive processes in counterfactual thinking about what might have been," *Psychol. Learn. Motivat.-Adv. Res. Theory*, vol. 37, pp. 105–154, Oct. 1997.
[24] L. M. Pereira and A. B. Lopes, "Cognitive prerequisites: The special case of counterfactual reasoning," in *Machine Ethics* (Studies in Applied Philosophy, Epistemology and Rational Ethics), vol. 53. Cham, Switzerland: Springer, 2020, pp. 97–102.
[25] J. Paik, Y. Zhang, and P. Pirolli, "Counterfactual reasoning as a key for explaining adaptive behavior in a changing environment," *Biologically Inspired Cognit. Archit.*, vol. 10, pp. 24–29, Oct. 2014.
[26] Y. Zhang, J. Paik, and P. Pirolli, "Reinforcement learning and counterfactual reasoning explain adaptive behavior in a changing environment," *Topics Cognit. Sci.*, vol. 7, no. 2, pp. 368–381, Apr. 2015.
[27] E. Kulakova, M. Aichhorn, M. Schurz, M. Kronbichler, and J. Perner, "Processing counterfactual and hypothetical conditionals: An fMRI investigation," *NeuroImage*, vol. 72, pp. 265–271, May 2013.
[28] G. Grahne, "Update and counterfactuals," *J. Log. Comput.*, vol. 8, no. 1, pp. 87–117, 1998.
[29] M. Ginsberg, "Counterfactuals," *Artif. Intell.*, vol. 30, no. 1, pp. 35–79, 1986.
[30] R. J. Aumann, "Backward induction and common knowledge of rationality," *Games Econ. Behav.*, vol. 8, no. 1, pp. 6–19, Jan. 1995.
[31] W. Spohn, "A ranking-theoretic approach to conditionals," *Cognit. Sci.*, vol. 37, no. 6, pp. 1074–1106, Aug. 2013.
[32] E. Kulakova and M. S. Nieuwland, "Understanding counterfactuality: A review of experimental evidence for the dual meaning of counterfactuals," *Lang. Linguistics Compass*, vol. 10, no. 2, pp. 49–65, Feb. 2016.
[33] N. Hendrickson, *Counterfactual Reasoning: A Basic Guide for Analysts, Strategists, and Decision Makers*. Morrisville, NC, USA: LULU Press, 2011.
[34] H. J. Ferguson and A. J. Sanford, "Anomalies in real and counterfactual worlds: An eye-movement investigation," *J. Memory Lang.*, vol. 58, no. 3, pp. 609–626, Apr. 2008.
[35] D. K. Lewis, *On the Plurality of Worlds*. Oxford, U.K.: Blackwell, 1986.
[36] D. K. Lewis, *Counterfactuals*. Oxford, U.K.: Blackwell, 1973.
[37] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[38] D. Hume, *Enquiry Concerning Human Understanding*. Oxford, U.K.: Clarendon, 1904.
[39] B. Kment, "Counterfactuals and explanation," *Mind*, vol. 115, no. 458, pp. 261–310, Apr. 2006.
[40] D. E. Over, C. Hadjichristidis, J. S. B. T. Evans, S. J. Handley, and S. A. Sloman, "The probability of causal conditionals," *Cognit. Psychol.*, vol. 54, no. 1, pp. 62–97, Feb. 2007.
[41] D. Edgington, "Estimating conditional chances and evaluating counterfactuals," *Studia Logica*, vol. 102, no. 4, pp. 691–707, Aug. 2014.
[42] A. L. McGill and J. G. Klein, "Contrastive and counterfactual reasoning in causal judgment," *J. Personality Social Psychol.*, vol. 64, no. 6, pp. 897–905, 1993.
[43] J. Fang, Z. Huang, and F. van Harmelen, "A method of contrastive reasoning with inconsistent ontologies," in *The Semantic Web*. Berlin, Germany: Springer, 2012, pp. 1–16.
[44] A. Páez, "The pragmatic turn in explainable artificial intelligence (XAI)," *Minds Mach.*, vol. 29, no. 3, pp. 441–459, Sep. 2019.
[45] J. Kunkel, T. Donkers, L. Michael, C.-M. Barbu, and J. Ziegler, "Let me explain: Impact of personal and impersonal explanations on trust in recommender systems," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2019, pp. 1–487.
[46] K. Sokol and P. Flach, "One explanation does not fit all: The promise of interactive explanations for machine learning transparency," in *KI-Künstliche Intelligenz*, no. 2. Berlin, Germany: Springer, 2020, pp. 235–250.
[47] N. Elzein, "The demand for contrastive explanations," *Philos. Stud.*, vol. 176, no. 5, pp. 1325–1339, 2019.
[48] T. Le, S. Wang, and D. Lee, "GRACE: Generating concise and informative contrastive sample to explain neural network model's prediction," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 238–248.
[49] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. D. Bie, and P. Flach, "FACE: Feasible and actionable counterfactual explanations," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Feb. 2020, pp. 344–350.
[50] F. Bergadano, V. Cutello, and D. Gunetti, "Abduction in machine learning," in *Handbook of Defeasible Reasoning and Uncertainty Management Systems: Volume 4 Abductive Reasoning and Learning*. Norwell, MA, USA: Kluwer Academic, 2000, pp. 197–229.
[51] M. T. Keane and B. Smyth, "Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI)," in *Case-Based Reasoning Research and Development*. Cham, Switzerland: Springer, 2020, pp. 163–178.
[52] R. M. J. Byrne, "Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 6276–6282.
[53] C. Molnar, *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*, 1st ed. Fletcher, NC, USA: LULU, Feb. 2019.
[54] K. Sokol and P. Flach, "Counterfactual explanations of machine learning predictions: Opportunities and challenges for AI safety," in *Proc. AAAI Workshop Artif. Intell. Saf.*, 2019, pp. 1–4.
[55] M. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2017, pp. 4069–4079.
[56] A. Lucic, H. Haned, and M. de Rijke, "Why does my model fail?: Contrastive local explanations for retail forecasting," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 90–98.
[57] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.
[58] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Softw. Eng. Group, School Comput. Sci. Math., Keele Univ., Keele, U.K., Dept. Comput. Sci., Durham Univ., Durham, U.K., Tech. Rep. EBSE 2007-001, 2007.
[59] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering—A systematic literature review," *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, Jan. 2009.
[60] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proc. 18th Int. Conf. Eval. Assessment Softw. Eng. (EASE)*, 2014, pp. 1–10.
[61] C. Aguilar-Palacios, S. Munoz-Romero, and J. L. Rojo-Alvarez, "Cold-start promotional sales forecasting through gradient boosted-based contrastive explanations," *IEEE Access*, vol. 8, pp. 137574–137586, 2020.
[62] H. Andreas and L. Casini, "Hypothetical interventions and belief changes," *Found. Sci.*, vol. 24, no. 4, pp. 681–704, Dec. 2019.
[63] E. Barnes, "Why P rather than Q? The curiosities of fact and foil," *Phil. Stud.*, vol. 73, no. 1, pp. 35–53, Jan. 1994.
[64] S. Baron, M. Colyvan, and D. Ripley, "How mathematics can make a difference," *Philosophers Imprint*, vol. 17, no. 3, pp. 1–29, 2017.
[65] S. Baron, "Counterfactual scheming," *Mind*, vol. 129, no. 514, pp. 535–562, Apr. 2020.
[66] S. Baron, M. Colyvan, and D. Ripley, "A counterfactual approach to explanation in mathematics," *Philosophia Mathematica*, vol. 28, no. 1, pp. 1–34, Feb. 2020.
[67] L. Bertossi, "An ASP-based approach to counterfactual explanations for classification," *Rules and Reasoning* (Lecture Notes in Computer Science), vol. 12173. Cham, Switzerland: Springer, 2020, 70–81.
[68] A. Bokulich, "How scientific models can explain," *Synthese*, vol. 180, no. 1, pp. 33–45, May 2011.
[69] G. Botterill, "Right and wrong reasons in folk-psychological explanation," *Int. J. Phil. Stud.*, vol. 17, no. 4, pp. 463–488, Oct. 2009.

[70] S. Boulter, "Contrastive explanations in evolutionary biology," *Ratio*, vol. 25, no. 4, pp. 425–441, 2012.
[71] M. J. L. Bours, "A nontechnical explanation of the counterfactual definition of confounding," *J. Clin. Epidemiol.*, vol. 121, pp. 91–100, May 2020.
[72] R. Briggs, "Interventionist counterfactuals," *Phil. Stud.*, vol. 160, no. 1, pp. 139–166, Aug. 2012.
[73] N. Campbell, "Self-forming actions, contrastive explanations, and the structure of the will," *Synthese*, vol. 197, pp. 1225–1240, Mar. 2018.
[74] A. Chakravartty, "Perspectivism, inconsistent models, and contrastive explanation," *Stud. Hist. Philosophy Sci. A*, vol. 41, no. 4, pp. 405–412, Dec. 2010.
[75] A. Chien, "Scalar implicature and contrastive explanation," *Synthese*, vol. 161, no. 1, pp. 47–66, Mar. 2008.
[76] S. Chin-Parker and J. Cantelon, "Contrastive constraints guide explanation-based category learning," *Cognit. Sci.*, vol. 41, no. 6, pp. 1645–1655, Aug. 2017.
[77] M. Day and G. S. Botterill, "Contrast, inference and scientific realism," *Synthese*, vol. 160, no. 2, pp. 249–267, Jan. 2008.
[78] J. Dickenson, "Reasons, causes, and contrasts," *Pacific Phil. Quart.*, vol. 88, no. 1, pp. 1–23, Mar. 2007.
[79] W. Fang, "An inferential account of model explanation," *Philosophia*, vol. 47, no. 1, pp. 99–116, Mar. 2019.
[80] R. R. Fernández, I. M. de Diego, V. Ace na, J. M. Moguerza, and A. Fernández-Isabel, "Relevance metric for counterfactuals selection in decision trees," in *Intelligent Data Engineering and Automated Learning—IDEAL 2019* (Lecture Notes in Computer Science), vol. 11871. Cham, Switzerland: Springer, 2019, pp. 85–93.
[81] R. R. Fernández, I. M. D. Diego, V. Aceña, A. Fernández-Isabel, and J. M. Moguerza, "Random forest explainability using counterfactual sets," *Inf. Fusion*, vol. 63, pp. 196–207, Nov. 2020.
[82] C. E. Franklin, "Agent-causation, explanation, and Akrasia: A reply to Levy's hard luck," *Criminal Law Philosophy*, vol. 9, no. 4, pp. 753–770, Dec. 2015.
[83] V. Gijsbers, "A quasi-interventionist theory of mathematical explanation," *Logique et Analyse*, vol. 60, no. 237, pp. 47–66, 2017.
[84] A. Ghazimatin, O. Balalau, R. S. Roy, and G. Weikum, "PRINCE: Provider-side interpretability with counterfactual explanations in recommender systems," in *Proc. 13th Int. Conf. Web Search Data Mining*, Jan. 2020, pp. 196–204.
[85] C. Held, "Towards a monist theory of explanation," *J. Gen. Philosophy Sci.*, vol. 50, no. 4, pp. 447–475, Dec. 2019.
[86] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata, "Grounding visual explanations," in *Computer Vision—ECCV 2018* (Lecture Notes in Computer Science), vol. 11206. Cham, Switzerland: Springer, 2018, pp. 269–286.
[87] J. A. Hird, "The political economy of pork: Project selection at the U.S. Army corps of engineers," *Amer. Political Sci. Rev.*, vol. 85, no. 2, pp. 429–456, Jun. 1991.
[88] C. R. Hitchcock, "The role of contrast in causal and explanatory claims," *Synthese*, vol. 107, no. 3, pp. 395–419, Jun. 1996.
[89] J. Hohwy, "Capacities, explanation and the possibility of disunity," *Int. Stud. Philosophy Sci.*, vol. 17, no. 2, pp. 179–190, Jul. 2003.
[90] P. W. Holland, "Causal counterfactuals in social science research," in *International Encyclopedia of the Social Behavioral Sciences*, 2nd ed. Oxford, U.K.: Elsevier, 2015, pp. 251–254.
[91] A. Kanehira, K. Takemoto, S. Inayoshi, and T. Harada, "Multimodal explanations by predicting counterfactuality in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8586–8594.
[92] J. Katz, "Situational evidence: Strategies for causal reasoning from observational field notes," *Sociol. Methods Res.*, vol. 44, no. 1, pp. 108–144, Feb. 2015.
[93] A. Kean, "A characterization of contrastive explanations computation," in *PRICAI'98: Topics in Artificial Intelligence* (Lecture Notes in Computer Science), vol. 1531. Berlin, Germany: Springer, 1998, pp. 559–610.
[94] J. Kim, C. Muise, A. Shah, S. Agarwal, and J. Shah, "Bayesian inference of linear temporal logic specifications for contrastive explanations," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 5591–5598.
[95] R. Knowles and J. Saatsi, "Mathematics and explanatory generality: Nothing but cognitive salience," in *Erkenntnis*. Dordrecht, The Netherlands: Springer, Aug. 2019, pp. 1–19, doi: 10.1007/s10670-019-00146-x.
[96] D. Kostic, "General theory of topological explanations and explanatory asymmetry," *Philos. Trans. Roy. Soc. B, Biol. Sci.*, vol. 375, no. 1796, 2020, Art. no. 20190321.
[97] J. Kuorikoski and P. Ylikoski, "Explanatory relevance across disciplinary boundaries: The case of neuroeconomics," *J. Econ. Methodol.*, vol. 17, no. 2, pp. 219–228, Jun. 2010.
[98] J. Kuorikoski and P. Ylikoski, "External representations and scientific understanding," *Synthese*, vol. 192, no. 12, pp. 3817–3837, Dec. 2015.
[99] D. N. Kutach, "The entropy theory of counterfactuals," *Philosophy Sci.*, vol. 69, no. 1, pp. 82–104, Mar. 2002.
[100] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "Unjustified classification regions and counterfactual explanations in machine learning," in *Machine Learning and Knowledge Discovery in Databases* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence), vol. 11907. Cham, Switzerland: Springer, 2020, pp. 37–54.
[101] N. Levy, "Luck and agent-causation: A response to Franklin," *Criminal Law Philosophy*, vol. 9, no. 4, pp. 779–784, Dec. 2015.
[102] P. Lichterman and I. A. Reed, "Theory and contrastive explanation in ethnography," *Sociol. Methods Res.*, vol. 44, no. 4, pp. 585–635, Nov. 2015.
[103] P. Lipton, *Inference to the Best Explanation* (Philosophical Issues in Science). Evanston, IL, USA: Routledge, 1991.
[104] E. J. Lowe, "What is the source of our knowledge of modal truths?" *Mind*, vol. 121, no. 484, pp. 919–950, 2012.
[105] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 607–617.
[106] R. Northcott, "Degree of explanation," *Synthese*, vol. 190, no. 15, pp. 3087–3105, Oct. 2013.
[107] J. Neyman and K. Iwaszkiewicz, "Statistical problems in agricultural experimentation," *Suppl. J. Roy. Stat. Soc.*, vol. 2, no. 2, pp. 107–180, 1935.
[108] M. Pexton, "Manipulationism and causal exclusion," *Philosophica*, vol. 92, no. 2, pp. 13–51, 2017.
[109] A. R. Pruss and J. L. Rasmussen, "Explaining counterfactuals of freedom," *Religious Stud.*, vol. 50, no. 2, pp. 193–198, Jun. 2014.
[110] A. Reutlinger, "Is there a monist theory of causal and noncausal explanations? The counterfactual theory of scientific explanation," *Philosophy Sci.*, vol. 83, no. 5, pp. 733–745, Dec. 2016.
[111] A. Reutlinger, "Does the counterfactual theory of explanation apply to non-causal explanations in metaphysics?" *Eur. J. Philosophy Sci.*, vol. 7, no. 2, pp. 239–256, 2017.
[112] A. Reutlinger, "Extending the counterfactual theory of explanation," in *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations*. Oxford, U.K.: Oxford Univ. Press, 2018, pp. 74–95.
[113] L. J. Rips and B. J. Edwards, "Inference and explanation in counterfactual reasoning," *Cognit. Sci.*, vol. 37, no. 6, pp. 1107–1135, Aug. 2013.
[114] D.-H. Ruben, "A counterfactual theory of causal explanation," *Nous*, vol. 28, no. 4, pp. 465–481, 1994.
[115] D. B. Rubin, "Bayesian inference for causal effects: The role of randomization," *Ann. Statist.*, vol. 6, no. 1, pp. 34–58, Jan. 1978.
[116] C. Q. Schneider and I. Rohlfing, "Case studies nested in fuzzy-set QCA on sufficiency: Formalizing case selection and causal inference," *Sociol. Methods Res.*, vol. 45, no. 3, pp. 526–568, 2016.
[117] R. Schweder, "Causal explanation and explanatory selection," *Synthese*, vol. 120, no. 1, pp. 115–124, 1999.
[118] C. Seelos and J. Mair, "Organizational closure competencies and scaling: A realist approach to theorizing social enteprise," in *Social Entrepreneurship and Research Methods*, vol. 9. Bingley, U.K.: Emerald Group Publishing Limited, 2014, pp. 147–187.
[119] E. Sober, "A theory of contrastive causal explanation and its implications concerning the explanatoriness of deterministic and probabilistic hypotheses," *Eur. J. Philosophy Sci.*, vol. 10, no. 3, pp. 1–15, Oct. 2020.
[120] R. Stalnaker, "A theory of conditionals," in *Studies in Logical Theory (American Philosophical Quarterly Monographs 2)*. Oxford, U.K.: Blackwell, 1968, pp. 98–112.
[121] A. Steglich-Petersen, "Against the contrastive account of singular causation," *Brit. J. Philosophy Sci.*, vol. 63, no. 1, pp. 115–143, Mar. 2012.
[122] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Farina, "Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2020, pp. 1–8.
[123] M. Strohminger and J. Yli-Vakkuri, "Knowledge of objective modality," *Phil. Stud.*, vol. 176, no. 5, pp. 1155–1175, May 2019.

[124] E. W. K. Tsang and F. Ellsaesser, "How contrastive explanation facilitates theory building," *Acad. Manage. Rev.*, vol. 36, no. 2, pp. 404–419, Apr. 2011.
[125] J. Woodward, *Making Things Happen: A Theory of Causal Explanation*. Oxford, U.K.: Oxford Univ. Press, 2003.
[126] P. Ylikoski and J. Kuorikoski, "Dissecting explanatory power," *Phil. Stud.*, vol. 148, no. 2, pp. 201–219, Mar. 2010.
[127] P. Ylikoski, *Social Mechanisms and Explanatory Relevance*. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[128] A. R. Akula, S. Todorovic, J. Y. Chai, and S.-C. Zhu, "Natural language interaction with explainable AI models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jul. 2019, pp. 87–90.
[129] S. Anjomshoae, K. Främling, and A. Najjar, "Explanations of black-box model predictions by contextual importance and utility," in *Explainable, Transparent Autonomous Agents and Multi-Agent Systems* (Lecture Notes in Artificial Intelligence), vol. 11763. Cham, Switzerland: Springer, 2019, pp. 95–109.
[130] A. Apicella, F. Isgrò, R. Prevete, and G. Tamburrini, "Contrastive explanations to classification systems using sparse dictionaries," *Image Analysis and Processing—ICIAP 2019* (Lecture Notes in Artificial Intelligence), vol. 11751. Cham, Switzerland: Springer, 2019, pp. 207–218.
[131] B. R. Barricelli, E. Casiraghi, J. Gliozzo, A. Petrini, and S. Valtolina, "Human digital twin for fitness management," *IEEE Access*, vol. 8, pp. 26637–26664, 2020.
[132] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati, "Plan explanations as model reconciliation: Moving beyond explanation as soliloquy," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 156–163.
[133] C.-H. Chang, E. Creager, A. Goldenberg, and D. Duvenaud, "Explaining image classifiers by counterfactual generation," in *Proc. 7th Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–13.
[134] S. Dandl, C. Molnar, M. Binder, and B. Bischl, "Multi-objective counterfactual explanations," *Parallel Problem Solving from Nature—PPSN XVI* (Lecture Notes in Computer Science), vol. 12269. Berlin, Germany: Springer, 2020, pp. 448–469.
[135] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives,"in *Proc. Neural Inf. Process. Syst. Found.*, 2018, pp. 592–603.
[136] C. Fernandez, F. Provost, and X. Han, "Counterfactual explanations for data-driven decisions," in *Proc. 40th Int. Conf. Inf. Syst. (ICIS)*, 2019, pp. 1–10.
[137] A. Ferrario, R. Weibel, and S. Feuerriegel, "ALEEDSA: Augmented reality for interactive machine learning," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–8.
[138] O. Gomez, S. Holter, J. Yuan, and E. Bertini, "ViCE," in *Proc. Int. Conf. Intell. User Interfaces (IUI)*, 2020, pp. 531–535.
[139] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 4254–4262.
[140] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini, "Factual and counterfactual explanations for black box decision making," *IEEE Intell. Syst.*, vol. 34, no. 6, pp. 14–23, Nov. 2019.
[141] J. Labaien, E. Zugasti, and X. D. Carlos, "Contrastive explanations for a deep learning model on time-series data," in *Big Data Analytics and Knowledge Discovery* (Lecture Notes in Computer Science), vol. 12393. Berlin, Germany: Springer, 2020, pp. 235–244.
[142] M. Lash, Q. Lin, N. Street, J. Robinson, and J. Ohlmann, "Generalized inverse classification," in *Proc. Int. Conf. Data Mining (SDM)*, 2017, pp. 162–170.
[143] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "Comparison-based inverse classification for interpretability in machine learning," in *Proc. 17th Int. Conf. Inf. Process. Manage. Uncertainty Knowl.-Based Syst. (IPMU)*. New York, NY, USA: Springer-Verlag, 2018, pp. 100–111.
[144] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "The dangers of post-hoc interpretability: Unjustified counterfactual explanations," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 2801–2807.
[145] S. Liu, B. Kailkhura, D. Loveland, and Y. Han, "Generative counterfactual introspection for explainable deep learning," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2019, pp. 1–5.
[146] D. Martens and F. Provost, "Explaining data-driven document classifications," *MIS Quart.*, vol. 38, no. 1, pp. 73–100, 2014.
[147] L. Merrick and A. Taly, "The explanation game: Explaining machine learning models using shapley values," in *Machine Learning and Knowledge Extraction* (Lecture Notes in Computer Science), vol. 12279. Cham, Switzerland: Springer, 2020, pp. 17–38.
[148] J. Moore, N. Hammerla, and C. Watkins, "Explaining deep learning models with constrained adversarial examples," in *PRICAI 2019: Trends in Artificial Intelligence* (Lecture Notes in Computer Science), vol. 11670. Cham, Switzerland: Springer, 2019, pp. 43–56.
[149] F. Mosca, S. Sarkadi, J. M. Such, and P. McBurney, "Agent EXPRI: Licence to explain," *Explainable, Transparent Autonomous Agents and Multi-Agent Systems* (Lecture Notes in Computer Science), vol. 12175. Cham, Switzerland: Springer, 2020, pp. 21–38.
[150] M. A. Neerincx, J. van der Waa, F. Kaptein, and J. van Diggelen, "Using perceptual and cognitive explanations for enhanced human-agent team performance," in *Engineering Psychology and Cognitive Ergonomics* (Lecture Notes in Computer Science), vol. 10906. Cham, Switzerland: Springer, 2018, pp. 204–214.
[151] M. Pawelczyk, K. Broelemann, and G. Kasneci, "Learning model-agnostic counterfactual explanations for tabular data," in *Proc. Web Conf.*, Apr. 2020, pp. 3126–3132.
[152] D. Rajapaksha, C. Bergmeir, and W. Buntine, "LoRMIkA: Local rule-based model interpretability with K-optimal associations," *Inf. Sci.*, vol. 540, pp. 221–241, Nov. 2020.
[153] C. Russell, "Efficient search for diverse coherent explanations," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2019, pp. 20–28.
[154] S. Sharma, J. Henderson, and J. Ghosh, "CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Feb. 2020, pp. 166–172.
[155] K. Sokol and P. Flach, "Glass-box: Explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual assistant," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 5868–5870.
[156] S. Sreedharan, S. Srivastava, and S. Kambhampati, "Hierarchical expertise level modeling for user specific contrastive explanations," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2018, pp. 4829–4836.
[157] R. Sukkerd, R. Simmons, and D. Garlan, "Towards explainable multi-objective probabilistic planning," in *Proc. 4th Int. Workshop Softw. Eng. Smart Cyber-Phys. Syst.* Washington, DC, USA: IEEE Computer Society, May 2018, pp. 19–25.
[158] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2019, pp. 10–19.
[159] J. van der Waa, M. Robeer, J. van Diggelen, M. Brinkhuis, and M. Neerincx, "Contrastive explanations with local foil trees," in *Proc. Workshop Hum. Interpretability Mach. Learn. (WHI)*, 2018, pp. 1–7.
[160] A. White and A. S. D. A. Garcez, "Measurable counterfactual local explanations for any classifier," in *Proc. 24th Eur. Conf. Artif. Intell. (ECAI)*, 2020, pp. 2529–2535.
[161] E. Zhao and R. Sukkerd, "Interactive explanation for planning-based systems," in *Proc. 10th ACM/IEEE Int. Conf. Cyber-Phys. Syst.*, Apr. 2019, pp. 322–323.
[162] N. J. van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," *Scientometrics*, vol. 84, no. 2, pp. 523–538, Aug. 2010.
[163] K. S. Jones and J. R. Galliers, *Evaluating Natural Language Processing Systems: An Analysis and Review*. New York, NY, USA: Springer-Verlag, 1996.
[164] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *J. Artif. Intell. Res.*, vol. 61, pp. 65–170, Jan. 2018.
[165] D. J. Hilton and B. R. Slugoski, "Knowledge-based causal attribution: The abnormal conditions focus model," *Psychol. Rev.*, vol. 93, no. 1, p. 75, 1986.
[166] D. J. Hilton, "Conversational processes and causal explanation," *Psychol. Bull.*, vol. 107, no. 1, pp. 65–81, 1990.

ILIA STEPIN received the Engineer degree (Hons.) in software engineering from the Moscow State University of Instrument Engineering and Computer Science, Russia, in 2012, the bachelor's degree in linguistics from Moscow State Linguistic University, Russia, in 2016, and the M.Sc. degree in computational linguistics from the University of Stuttgart, Germany, in 2018. He is currently pursuing the Ph.D. degree with the University of Santiago de Compostela, Spain. He is currently a Research Assistant with the Research Center in Intelligent Technologies (CiTIUS), Santiago de Compostela, Spain. He is also working on a doctoral project entitled ''Argumentative Conversational Agents for Explainable Artificial Intelligence'' with an emphasis on counterfactual explanation generation. His research interests include (but are not limited to) natural language processing, argumentation theory, and human–machine interaction.

JOSE M. ALONSO (Member, IEEE) received the M.Sc. and Ph.D. degrees in telecommunication engineering from the Technical University of Madrid (UPM), Spain, in 2003 and 2007, respectively. Since June 2016, he has been a Postdoctoral Researcher with the Research Centre in Intelligent Technologies (CiTIUS), University of Santiago de Compostela (USC). He is member of the CiTIUS-USC Research Group on Intelligent Systems. He has published more than 140 papers in international journals, book chapters, and in peer-review conferences. His research interests include explainable artificial intelligence, computational intelligence, interpretable fuzzy systems, natural language generation, and development of software tools. He is also the Deputy Coordinator of the H2020-MSCA-ITN-2019 Project titled ''Interactive Natural Language Technology for Explainable Artificial Intelligence'' (NL4XAI). He is recognized as the Ramon y Cajal Researcher (RYC-2016-19802), the Chair of the Task Force on Explainable Fuzzy Systems in the Fuzzy Systems Technical Committee of the IEEE Computational Intelligence Society (IEEE-CIS), a member of the IEEE-CIS Task Force on Explainable Machine Learning, the IEEE-CIS Task Force on Fuzzy Systems Software, the IEEE-CIS Content Curation Subcommittee, and the IEEE 1855 Working Group for the maintenance and update of the IEEE Standard for Fuzzy Markup Language IEEE Std 1855TM-2016, an Associate Editor of *IEEE Computational Intelligence Magazine*, and a Secretary of the ACL Special Interest Group on Natural Language Generation.

ALEJANDRO CATALA received the Ing., M.Sc., and Ph.D. degrees from the Universitat Politécnica de València (UPV), in 2006, 2008, and 2012, respectively. He is currently a Postdoctoral Researcher awarded with the Juan de la Cierva Fellowship at CiTIUS. He carried out his coBOTnity Project at the Human Media Interaction Laboratory, University of Twente, The Netherlands. His research interests include human–computer interaction, artificial intelligence, and intelligent user interfaces. Along his research career, he has been awarded with several personal grants by the Spanish and Valencian governments as well as the European Union's Horizon 2020 Research and Innovation Program under the Marie Sklodowska-Curie Individual Fellowship Grant (MSCA-IF). He has contributed with more than 70 peer-review publications, and regularly serves as a Reviewer and a Program Committee Member in conferences related to software engineering, artificial intelligence, and human–computer interaction.

MARTÍN PEREIRA-FARIÑA received the B.A. and Ph.D. degrees in philosophy and computational logic from the University of Santiago de Compostela, Spain, in 2007 and 2014, respectively. He is currently a Lecturer with the Department of Philosophy and Anthropology, University of Santiago de Compostela. His research interests include digital humanities, philosophy of language, argumentation theory, and computational models for it.