Machine Reasoning Explainability

Kristijonas Čyras*, Ramamurthy Badrinath, Swarup Kumar Mohalik,
Anusha Mujumdar, Alexandros Nikou, Alessandro Previti,
Vaishnavi Sundararajan, Aneta Vulgarakis Feljan
Ericsson Research

September 2, 2020

Abstract

As a field of AI, Machine Reasoning (MR) uses largely symbolic means to formalize and emulate abstract reasoning. Studies in early MR have notably started inquiries into Explainable AI (XAI) – arguably one of the biggest concerns today for the AI community. Work on explainable MR as well as on MR approaches to explainability in other areas of AI has continued ever since. It is especially potent in modern MR branches, such as argumentation, constraint and logic programming, planning. We hereby aim to provide a selective overview of MR explainability techniques and studies in hopes that insights from this long track of research will complement well the current XAI landscape. This document reports our work in-progress on MR explainability.

1 Introduction

Machine Reasoning (MR) is a field of AI that complements the field of Machine Learning (ML) by aiming to computationally mimic abstract thinking. This is done by way of uniting known (yet possibly incomplete) information with background knowledge and making inferences regarding unknown or uncertain information. MR has outgrown Knowledge Representation and Reasoning (KR, see e.g. [27]) and now encompasses various symbolic and hybrid AI approaches to automated reasoning. Central to MR are two components: a knowledge base (see e.g. [61]; common in Axiom Pinpointing, Automated Theorem Proving (ATP) and Proof Assistants, Non-classical Logic-Based Reasoning (Logic Programming), Argumentation) or a model of the problem (see e.g. [89]; common in Constraint Programming (CP), Planning, Decision Theory, Reinforcement Learning), which

*Corresponding author. Email: kristijonas.cyras@ericsson.com, ORCiD: 0000-0002-4353-8121

formally represents knowledge and relationships among problem components in symbolic, machine-processable form; and a general-purpose inference engine or solving mechanism, which allows to manipulate those symbols and perform semantic reasoning.[1]

The field of Explainable AI (XAI, see e.g. [1, 13, 20, 54, 130, 146, 149, 155, 169, 177, 196]) encompasses endeavors to make AI systems intelligible to their users, be they humans or machines. XAI comprises research in AI as well as interdisciplinary research at the intersections of AI and subjects ranging from Human-Computer Interaction (HCI), see e.g. [149], to social sciences, see e.g. [33, 144]. Explainability of AI is often seen as a crucial driver for the real-world deployment of trustworthy modern AI systems.

According to e.g. Hansen and Rieger in [100], explainability was one of the main distinctions between the 1^{st} (dominated by KR and rule-based systems) and the 2^{nd} (expert systems and statistical learning) waves of AI, with expert systems addressing the problems of explainability and ML approaches treated as black boxes. With the ongoing 3^{rd} wave of AI, ML explainability has received a great surge of interest [13, 54, 149]. By contrast therefore, it seems that a revived interest in MR explainability is only just picking up pace (e.g. ECAI 2020 Spotlight tutorial on Argumentative Explanations in AI[2] and KR 2020 Workshop on Explainable Logic-Based Knowledge Representation[3]). However, explainability in MR dates over four decades, see e.g. [100, 110, 151, 155, 196]. Explainability in MR can be roughly outlined thus.

1^{st} generation expert systems provide only so-called (*reasoning*) *trace explanations*, showing inference rules that led to a decision. A major problem with trace explanations is the lack of "information with respect to the system's general goals and resolution strategy"[151, p. 174]. 2^{nd} generation expert systems instead provide so-called *strategic explanations*, revealing "why information is gathered in a certain order, why one knowledge piece is invoked before others and how reasoning steps contribute to high-level goals"[151, p. 174]. Going further, so-called *deep explanations* separating the domain model from the structural knowledge have been sought, where "the system has to try to figure out what the user knows or doesn't know, and try to answer the question taking that into account."[204, p. 73] Progress in MR explainability notwithstanding, it can been argued (see e.g. [137, 151, 169]) that to date, explainability in MR particularly and perhaps in AI at large is still insufficient in aspects such as justification, criticism, and cooperation. These aspects, among others, are of concern in the modern MR explainability scene (around year 2000 onwards), whereby novel approaches to explainability in various branches of MR have been making appearances. We review some of them here and spell out the explainability questions addressed therein.

[1] See, however, e.g. [25] for an alternative view of MR stemming from a sub-symbolic/connectionist perspective.
[2] https://www.doc.ic.ac.uk/ afr114/ecaitutorial/
[3] https://lat.inf.tu-dresden.de/XLoKR20/

1.1 Motivation

The following summarizes our motivations and assumptions in this work.

1. We appreciate that MR is not yet a commonplace AI term, unlike e.g. ML or KR. We recognize that MR has evolved from KR and comprises other, mostly symbolic, forms of reasoning in AI. However, we here do not attempt to characterize MR, let alone cover all of its branches. Rather, we focus on the MR branches that most prominently exhibit approaches to explainability. Our overview of MR explainability is therefore bounded in scope. Nonetheless, it is dictated by our (invariably limited) professional understanding of the most relevant (e.g. historically important, well established and widely applicable, or trending) MR contributions to XAI.
2. Accordingly, we acknowledge that explainability in AI has been studied for a while and has in time evolved in terms of (a) areas of AI and (b) desiderata. Yet, we maintain that the foundational contributions and the lessons learnt (as well as forgotten) from the research on explainability in KR and expert systems are still very much relevant to both MR explainability in particular, and XAI at large. Specifically:
   (a) Adaptations of long established MR explainability techniques (Classical Logic-Based Reasoning and Non-classical Logic-Based Reasoning (Logic Programming); see Sections 3.1 and 3.2 respectively) have found their ways into newer MR areas (for instance, Axiom Pinpointing in Description Logics and Planning; see Sections 3.1.1 and 3.5 respectively).
   (b) Relatively newer MR branches, such as Answer Set Programming (ASP, Section 3.2.3) and Argumentation (Section 3.3), necessitate and inform newer forms of explainability. In particular, previously established techniques of mostly logical inference attribution do not suffice anymore, see e.g. [82, 123, 151, 204]. On the one hand, some modern MR approaches, such as ASP and Constraint Programming (CP, Section 3.1.2), currently provide techniques that can effectively be considered interpretable but whose workings are nevertheless difficult to explain. On the other hand, XAI desiderata now include aiming for dialogical/conversational explanations, interactivity, actionability as well as causality. These rediscovered concerns are being addressed by modern MR approaches to explainability. Thus, various branches of MR are concerned with explainability anew.
3. We maintain that modern MR systems can hardly be called explainable *just by virtue of being symbolic*[4], in contrast to early expert and intelligent systems often being referred to thus [1]. However, we speculate and contend that MR presumably offers more immediate intelligibility than e.g. ML. Perhaps this is what lends MR explainability approaches to be applied to *both* MR itself and other areas of research, such as ML (e.g. [3, 15, 51, 106, 185]), decision support and recommender systems (e.g. [45, 80, 158, 172]), planning (e.g. [38, 44, 69, 74, 81]), scheduling

[4]Where symbolic entities carry intrinsic semantical meaning and are perhaps more readily interpretable and intelligible than the algebraic-symbolic entities in sub-symbolic/connectionist AI.

(e.g. [58]), legal informatics (e.g. [17, 57]), scientific debates (e.g. [183]). We speculate that it is also potentially the theoretical guarantees often ensured by MR methods that make MR explainability appealing, cf. e.g. [105, 187].

4. Most recent overviews of general XAI appear to focus mostly on ML and somewhat ignore MR, e.g. [1, 13, 96, 146, 149, 177], apart from potentially briefly discussing early expert systems. (See however [158] for a systematic review of explanations in decision support and recommender systems, where MR constitutes majority of the referenced approaches; reviews of explainability in e.g. ASP [78] and Planning [44] are also welcome examples of not-so-common area-specific MR overviews.) We feel that having a broader view of the XAI agenda is crucial in general and that an overview of MR explainability is due for at least the following reasons.
   (a) Explainable MR constitutes a rich body of research whose works span many branches of MR with long histories. A consolidated, even if limited, overview will provide guidance to exploring and building on this research.
   (b) MR explainability techniques are also used for e.g. explainable ML and an overview will help AI researchers to see a bigger picture of XAI as well as to promote innovation and collaboration.
   (c) XAI has arguably started with explainable MR and some of the same conceptual MR explainability techniques can be (and are being) applied to achieve current XAI goals. An MR explainability overview may allow researchers to rediscover known problems and solutions (potentially saving from reinventing things) and to give credit where it is due.

1.2 Contributions

Our contributions in this report are as follows.

1. Building on conceptual works on XAI as well as XAI overviews, we propose a loose categorization of MR explainability approaches in terms of broad families of explanations, namely *attributive*, *contrastive* and *actionable*.
2. We provide an overview of some prominent approaches to MR explainability, differentiating by branches of MR and indicating the families that such explanations roughly belong to.
3. We indicate what kinds of questions the explanations aim to answer.

This is not a systematic review, but rather an overview of conceptual techniques that MR brings to XAI. We do not claim to be exhaustive or even completely representative of the various MR approaches to explainability, let alone to characterize what counts as a branch of MR. Rather, we hope to enable the reader to see a bigger picture of XAI, focusing specifically on what we believe amounts to MR explainability.

1.3 Omissions

In this report, we omit the following aspects of XAI.

- We leave out the following research areas that could potentially be considered related to MR.
  1. Rule-based classification/prediction, e.g. [209], which comprises of largely ML-focused approaches for associative/predictive rule generation and/or extraction. Overviews of such rule-based ML explainability can be found in e.g. [1, 9, 13, 96].
  2. Neuro-symbolic computing and graph neural networks, see e.g. [129] for an overview, where the use of classical logics alongside ML techniques has recently been argued as a factor enabling explainability. Other works use classical logic for approximation and human-readable representation of ML workings. For instance, the authors in [47, 48] use neural networks to learn relationships between ML classifier outputs and interpret those relationships using Boolean variables that represent neuron activations as well as predicates that represent the classification outputs. These then yield human-readable first-order logic descriptions of the learned relationships. Neuro-symbolic computing comprises heavily of ML-focused approaches that we deem beyond the scope of this report on MR explainability.
  3. Probabilistic reasoning, which is a field (of Mathematics as well as Computer Science) in itself, and spans AI at large. We meet several explanation-oriented approaches when discussing various MR branches, but we do not separately consider Bayesian Networks or Probabilistic Graphical Models among MR branches with a focus on explainability.
  4. Game theory, which is a field of Mathematics that studies strategic interaction between rational decision makers. In AI, game theory provides foundations to reasoning with preferences, multi-agent systems (MAS) and mechanism design among others, possibly with the benefit of enabling explainability. SHAP (Shapley additive explanations) [140] is a well-known example of application of game theory to explainability in ML as well as MR (e.g. [126]). We do not think that game-theoretic approaches constitute a branch of MR, or at least not one with a particular focus to explainability. We do, however, review some related approaches to explainability from Decision Theory.
  5. Robotics, which is an interdisciplinary branch of Engineering and Computer Science. In terms of explainability, the cornerstone notion there is explainable agency [130], which amounts to an autonomous agent being able to do the following:
     (a) explain decisions made during plan generation;
     (b) report which actions it executed;
     (c) explain how actual events diverged from a plan and how it adapted in response;
     (d) communicate its decisions and reasons.

     In [10] Anjomshoae et al. review explainable agency approaches where goal-driven agents

and robots purport to explain their actions to a human user. We encounter some approaches to explainability of autonomous agents which belong to specific branches of MR such as Planning. Generally, however, explainable agency considers human-AI interaction aspects which we do not cover in this report.

- Medium of explanations, i.e. whether explanations are textual, graphical, etc., see e.g. [93, 128, 149, 189]. However important, these are largely human-AI interaction aspects, beyond the scope of this report.
- Evaluation of explanations is beyond the scope of this report too. We believe that, on the one hand, well-established computational metrics for systematic comparison and evaluation of explanations are generally lacking [97],[5] which we believe is especially true with respect to MR approaches to explainability. On the other hand, we believe that research on evaluation of, particularly, MR explainability approaches via human user studies is complicated and not yet mature either, see e.g. [149].

We hope these omissions do not lessen our contribution of a conceptual overview of MR explainability.

2 Explainability

We briefly discuss here the main purposes of explanations in AI systems and present a categorization of explanations. We propose that, intuitively, the main purpose of explanations in XAI is to enable the user of an AI system to not only understand (to a certain extent) the system, but also to do something with the explanation. We suggest this entails that the explanations answer some sort of questions about the AI system dealing with the problem at hand. The kinds of answers provided by the explanations will help us to loosely categorize them, enabled by approaches to explainability in MR.

2.1 Purpose of Explanations

At large, the purpose of explainability of an AI system can be seen to be two-fold, to quote from [110, p. 160]:

> The system either provides knowledge and explanations necessary for the user to carry out his or her task, or alternatively, the system carries out some action and then explains the need and reason for the action the system itself has taken to the user.

Borrowing from [13, p. 85, our emphasis], we stipulate what an explainable AI system entails:

> Given an audience, an explainable Artificial Intelligence is one that *produces details or reasons* to make its functioning clear or easy to understand.

[5] Though the literature on computational measures of explainability in ML is expanding, see e.g. [37, 149].

In other words, an explainable AI system has to first **produce details and reasons underlying its functioning** – call this an **explanation**. It is then upon the explainee (i.e. the recipient or user of the explanation, also called the audience) to take in the explanation. Thus, we are interested in the purpose of an explanation (see e.g. [149, 158, 176, 194]) from the explainee point-of-view:

> "What will I (i.e. the agent, human or artificial) do with the explanation?"

We do not attempt a representative list of the purposes of explanations, but nevertheless give some examples. For instance, an expert in the domain in which an AI system is applied may want to do any one of the following:

a) Understand how the system functions in general, in mundane situations, whence they expect explanations to pertain to certain known aspects of the domain;
b) Learn how the system outputs aspects of the domain that are unexpected to the expert;
c) Confirm and compare the system's behavior in both expected and unexpected situations;
d) Act on the produced knowledge, whence they expect guidance towards desirable results.

Different purposes of explanations are well reflected by what are called XAI design goals in [149], that amount to "identifying the purpose for explanation and choosing what to explain for the targeted end-user and dedicated application." See also e.g. [158, 194] for purposes of, particularly, human-centric explanations in AI. Note, however, that we do not limit ourselves to human users of AI systems. Instead, we allow AI (or otherwise intelligent machine) agents themselves to require explanations from AI systems. Considering the progress in developing intelligent machines and autonomous systems, it seems natural to foresee situations where an AI agent probes another for e.g. a justification of the latter's (intended) actions in order to achieve or negotiate common goals.

We do acknowledge that the intended purpose is part of the explanation's context, among other factors such as the explainer, the explainee and the communication medium [194]. The context, specifically the explainee, may inform or correlate with the purpose of explanations, especially in cases of human audiences [13, 149, 198, 214]. But that need not generally be the case: if the explainees are people, then it is "people's goals or intended uses of the technology, rather than their social role" [214] that shape the explainability desiderata. Note that there can be both AI-to-human and AI-to-AI explainability, but we consider the purpose of explanations, rather than the nature of the explainee, to be primary.

Finally, we recognize that various usability desiderata [189], including actionability (see e.g. [59, 123, 189]), are key to the purpose of explainability. We thus maintain that irrespective of the purpose of explanations, for the explainee to consume and act upon, i.e. "to do something with" an explanation, *the explanation has to answer some question(s) about the AI system and its functioning*. We discuss next how explanations in MR (and potentially in AI at large) can be categorized according to the questions they aim to answer.

2.2 High-level Questions

Intuitively, answering *what*-, *why*-, *when*-, *how*-types of questions about an AI system's functioning falls under the purview of explainability [81, 137, 149, 155]. At a high-level, we are interested in the following questions:

- Q1 Given a representation of the problem and given a query, e.g. a decision taken or an observation, what are the reasons underlying the inference pertaining to the query?
- Q2 Given a representation of the problem and its solution or an answer to a query, why is something else not a solution or an answer?
- Q3 Given different information or query, e.g. a different decision or observation, would the solution/answer change too, and how? (Purely based on explanation, without recomputing!)
- Q4 Given a representation of the problem and its current solution, and given a representation of a desired outcome, what decisions or actions can be taken to improve the current solution and to achieve an outcome as close as possible to the desired one?

These questions need not be limited to MR, but may well apply to e.g. ML too.

Questions of type Q1 pertain to inputs to the system (see e.g. [137, p. 14]), definitions and meaning of internal representations (see e.g. [155, p. 387], [128, p. 110], [32, p. 164]) and attribution of those to the outputs (see e.g. [149]). Such questions aim at soliciting "insight about the information that is utilized and the rules, processes or steps that are used by the system to reach the recommendation or the outcome it has generated for a particular case."[93, p.1483] Answers to Q1 type of questions thus "inform users of the current (or previous) system state" and how the system "derived its output value from the current (or previous) input values"[137, p. 14].

Questions of type Q2 and Q3 pertain to reasons against the current outcome and in favor of alternative outcomes as well as to alternative inputs and parameters (see e.g. [32, 81, 128, 137, 149, 155]). Such questions solicit characterization of "the reasons for differences between a model prediction and the user's expected outcome."[149] Answers to Q2 and Q3 types of questions thus "inform users why an alternative output value was not produced given the current input values" and "could provide users with enough information to achieve the alternative output value"[32, p. 167].

Questions of type Q4 pertain to changes to inputs, modelling or the problem itself that would lead to user-desired outputs (see e.g. [137, 149]). Such questions aim at soliciting "hypothetical adjustments to the input or model that would result in a different output"[149]. Answers to Q4 type of questions thus provide guidelines that help the user to achieve a (more) desired outcome.

We next propose a loose categorization of explanations in MR (and potentially AI at large) that places explanations in families of types of explanations aiming to answer the high-level questions.

2.3 Categorization for Explanations

The driving factors of our categorization are the questions that explanations specifically in various MR approaches aim to answer. Building on other works that aim at classifications/categorizations of explanations, see [93, 100, 110, 123, 128, 137, 144, 149, 151, 152, 169, 177, 189, 207], we distinguish three families/types of explanations: **attributive** (e.g. logical inference attribution, feature importance association), **contrastive** (e.g. reasons con as well as pro, counterfactuals), **actionable** (e.g. guidelines towards a desired outcome, realizable actions). We characterize these families of explanations using the following notions.[6]

Abstractly, we assume at hand

an AI system $\mathbb{S}$ that given information i yields outcome o.

For instance, in some forms of MR, the system can be instantiated by a knowledge base KB consisting of background knowledge, that given information (e.g. a query) i represented in symbolic form entails (for some variant of formally defined entailment $\models$) inference o:

$$KB \cup \{i\} \models o$$

In some forms of ML, the role of the system can be taken by a model or function $f : X \to Y$ (trained on some set $X' \subseteq X$ of data), that given instance $i \in X$ assigns to it a label $o \in Y$:

$$f(i) = o$$

The above are but examples of high-level descriptions of AI methodologies. They are nevertheless helpful to find intuitions behind the explanation categories suggested as follows.

2.3.1 Attributive Explanations

Attributive explanations (see e.g. [140, 144, 151]) rely on the notions such as trace, justification, attribution and association, widely used in MR and ML literature alike. At a high level, they aim to answer the following question:

Q1 Given a representation of the problem and given a query, e.g. a decision taken or an observation, what are the reasons underlying the inference pertaining to the query?

Or, using the notions above:

What are the details and reasons for system $\mathbb{S}$ yielding outcome o, given information i?

[6] Note that these are not formal definitions or descriptions. We believe that due to diversity of MR a formalization of such terms would be incomplete, controversial and unhelpful, if not impractical or outright impossible. Instead, we use natural language descriptions to practically convey our ideas and supply intuition.

Attributive explanations justify o by attributing to/associating with o (parts of) $\mathbb{S}$ and i. If applicable, attributive explanations may contain a trace of such attribution/association. In MR, for instance, using the above notation, an attributive explanation for o given i can be

$$KB' \subseteq KB \text{ such that } KB' \cup \{i\} \models o$$

Normally, if applicable, such KB' would be required to be minimal (for some form of minimality, e.g. subset inclusion), consistent (for some notion of consistency or non-contradiction) and informative (or non-trivial), see e.g. [73] for formalization in logical terms. As a trace, KB' can contain e.g. inference rules, proof or some sort of argument for o given i. In the case of ML, an attributive explanation can be some minimal part $\hat{i}$ of instance i that suffices for f to yield o (irrespective of the rest of i), see e.g. [105] for examples of such explanations for ML classifications, formalized as well in logical terms. In another form, for instance in planning, an attributive explanation could exhibit (the relevant part of) the model of the problem or the state the model is in, given i, together with relational or causal links to o.

Attributive explanations are easy to design and are therefore commonplace [93, p.1483]. They are prevalent in all MR branches we consider in this report, perhaps most prominently in Classical Logic-Based Reasoning (Section 3.1). That they are prominent among classical-logic based approaches to explainability is not surprising because, as we have seen, explainability started with early AI systems which were largely based on classical logics. That attributive explanations are prevalent in all MR approaches to explainability is not surprising either, because attribution can be seen to be necessary for building the more advanced contrastive and actionable explanations (see e.g. [149]).

2.3.2 Contrastive Explanations

Contrastive explanations (see e.g. [110, 120, 144, 149, 188, 191]) pertain to the notions such as criticism, contrast, counterfactual and dialogue. At a high level, they aim to answer the following questions:

- Q2 Given a representation of the problem and its solution or an answer to a query, why is something else not a solution or an answer?
- Q3 Given different information or query, e.g. a different decision or observation, would the solution/answer change too, and how? (Purely based on explanation, without recomputing!)

In other words:

> What are the details and reasons of system $\mathbb{S}$ yielding outcome o rather than different outcome o', given information i?
> And supposing that the given information is not i (but some different i'), would o' be the outcome?

Contrastive explanations address potential criticisms of the yielded outcome o given information i by dealing with contrastive outcomes o' and information i'.

On the one hand, in addition to an attributive explanation as to why $\mathbb{S}$ yields o given i, contrastive explanations can provide counterexamples attributing to (parts of) $\mathbb{S}$ and i why o' is not possible. For instance, consider a classical logic-based setting where background knowledge KB is queried with i, expecting a 'yes' or 'no' outcome $o \in \{y, n\}$ (regarding e.g. provability of i from KB). If $o = n$, then a contrastive explanation as a counterexample to the alternative outcome $o' = y$ could be a model M in which $KB \cup \{\neg i\}$ hold true. In forms of ML, a contrastive explanation as a counterexample can be some minimal part $\hat{i}$ of instance i that prevents f from yielding o (irrespective of the rest of i without $\hat{i}$). In, say, planning, a contrastive explanation as a counterexample can be some part of the model that together with the given information makes some goal unachievable.

On the other hand, contrastive explanations can work via counterfactuals, see e.g. [33, 90, 203]. Taken plainly, a "counterfactual is a statement such as, 'if p, then q,' where the premise p is either known or expected to be false."[90, p. 35] In contrast to attributive explanations and counterexamples, "counterfactuals continue functioning in an end-to-end integrated approach"[203, p. 850], indicating consequences of changing the given information. So counterfactual contrastive explanations are about making or imagining different choices and analyzing what could happen or could have happened. Often, counterfactual contrastive explanations address changes with respect to more desirable outcomes than the one yielded by the AI system [33, 90, 203].

In our terms, where i is information at hand, a counterfactual invites one to consider what happens if different (and thus currently false) information i' were at hand, speculating that it may lead to a different, perhaps more desirable, outcome o'. In more conventional MR terms then, a contrastive explanation as a counterfactual can for instance be a

$$\text{modification } i' \text{ of } i \text{ such that } KB \cup \{i'\} \models o'$$

Taken together with the system $\mathbb{S}$ yielding outcome o given information i, this expresses that 'if i' rather than i was given, then o' rather than o would be the outcome'.

Note though that it may be that no such modification is achievable or desirable, perhaps due to restrictions placed by the underlying application. However, if it is, then it may be reasonable to strive for some minimal modification i' that is in some sense most similar to i. In addition, an attributive explanation could be incorporated by, for instance, exhibiting some minimal $KB' \subseteq KB$ such that $KB' \cup \{i'\} \models o'$, so that the counterfactual explanation indicates some minimal modification of the given information which together with some background knowledge suffices to yield a more desirable outcome. Similarly, in ML terms, if at least part $\hat{i}$ of instance i needs to be changed for f to yield o' instead of o, then a contrastive explanation as a counterfactual can be some minimal modification i' of i (if it exists) that necessitates f to yield o'. In planning, a given difficult goal can be reduced to

a sub-goal by considering counterfactual "if only thus-and-so were true, I would be able to solve the original problem" as a contrastive explanation, which then entails "arranging for thus-and-so to be true"[90, p. 36], if possible.

Another variant of contrastive explanations in MR takes form in graph-like representations of pros and cons of reasoning outcomes. Explanations of this type amount to defining a formal structure, which can usually be represented as a graph, that consists of information $KB' \subseteq KB$ most relevant for yielding the outcome o given i, together with relationships among $KB' \cup \{i, o\}$ revealing information dependencies and/or which information is in favor or against o. Such graph-like explanations are popular in Answer Set Programming (ASP), where they can be comprised of the (positive and negative) literals and rules considered in deriving the answer o to the query i from a logic program KB, see Section 3.2.3. They are also prevalent in Argumentation, where graphs capture the relevant supporting and conflicting information from $KB \cup \{i\}$ that allows to contrast the justifications of, and counterexamples to o. Such graph-based explanations provide basis for dialogical explanations that formalize criticism and defense of the yielded outcome. Just as counterfactual explanations, dialogical ones allow the explainee to be engaged, rather than simply presented, with explanations.

Contrastive explanations of these and similar various forms appear in Constraint Programming (CP), Automated Theorem Proving (ATP) and Proof Assistants, forms of abductive reasoning, Answer Set Programming (ASP), Argumentation, Planning. They are non-trivial to define and design, given the usually numerous contrastive situations, alternative answers and solutions. Overall, contrastive explanations, especially counterfactual and dialogical ones, are strongly related to actionable explanations, in that contrastive explanations can support provision of actions and guidelines following which the AI system will yield a desired outcome.

2.3.3 Actionable Explanations

We maintain that *actionable explanations* (see e.g. [59, 123, 189]) should be interventional, interactive, collaborative, pedagogic. At a high level, they aim to answer the following question:

Q4 Given a representation of the problem and its current solution, and given a representation of a desired outcome, what decisions or actions can be taken to improve the current solution and to achieve an outcome as close as possible to the desired one?

In other words:

> What can be done in order for system $\mathbb{S}$ to yield outcome o, given information i?

Actionable explanations address potential interventions that may yield a desired outcome o, given information i. They entail both interaction and collaboration between the system and its user in that actionable explanations guide or teach the user on what actions/changes can be taken/made and the user may choose to follow them or not to. Importantly, actionable explanations allow to take actions or

make changes that alter the system and possibly the problem themselves. For instance, in MR terms, an actionable explanation can be a

$$\text{modification } KB' \text{ of } KB \text{ such that } KB' \cup \{i\} \models o$$

As with contrastive explanations, it would obviously be normal to require some minimality of the modification KB' and its similarity to KB (note that minimality may be a complicated aspect especially with respect to non-monotonic entailment). Additionally, an actionable explanation can supply a modification i' of i such that $KB' \cup \{i'\} \models o$, similarly to a counterfactual contrastive explanation.

In forms of ML, an actionable explanation can be some designation of the model's f parameter changes that result in a modified model f' such that $f'(i) = o$. (This can be achieved e.g. by directly modifying the weights, or the classification thresholds or even the training set, see e.g. [123] for examples of such actionable explanations for Naive Bayes Classifiers). In planning or scheduling, an actionable explanation can be some minimal change of goals or resources (i.e. modification of the problem and hence the solver's model) that are needed to attain a (solvable problem and its) solution satisfying as much as possible the initial goals and constraints.

Actionable explanations also apply to the situations where the world itself changes (i.e. not necessarily as a consequence of the user's actions) and the previous outcomes/solutions are no longer good or do not exist at all, so that explanations provide guidance in order to obtain new or better outcomes/solutions. For instance, the background knowledge KB or the input space X of the model f may change and the previous queries or inferences become invalid, whence actionable explanations answer why and how to react. Ideally, actionable explanations would enable a meaningful interaction between an AI system and its user (again, human or machine, indifferently) leading to a fruitful collaboration. Such an interaction could be for instance conversational, formalized as dialogue between the user and the system, see e.g. [18, 51, 57, 101, 110, 121, 128, 144, 146, 150, 151, 186, 188, 190, 204].

Perhaps due to their inherent complexity of taking into account arguably more consequential changes as well as attributions and contrasts, actionable explanations are not very prominent in MR. We believe that they exist perhaps to a limited extent in e.g. Argumentation.

3 MR Branches and Explanations

We here overview branches of MR where explainability is studied. We do not claim to define what counts as MR but offer our perspective. We think that assignment to MR of some approaches, particularly those falling into Classical Logic-Based Reasoning, Non-classical Logic-Based Reasoning (Logic Programming) and Argumentation, will not be controversial. We likewise feel that Decision Theory and Planning belong well to MR too. However, we admit that Multi-Agent Systems and Causal Approaches can be regarded as more interdisciplinary branches, but we maintain to consider

approaches within the realm of MR. Finally, Reinforcement Learning is traditionally an area within ML, but there are some approaches (that we focus on) that use MR techniques (as understood in this report) for explainability.

3.1 Classical Logic-Based Reasoning

Arguably the most well established and far-back dating MR explainability techniques rest on classical logic-based derivation/deduction. Falappa et al. outline well in [73] the notion of explanation used in logic-based reasoning formalisms as follows. An explanation for a sentence a is a minimal, consistent and informative set A of sentences deducing a (where minimality is with respect to subset inclusion $\subseteq$, consistency and deduction are formalized with respect the underlying logic and informativeness requires the consequences of A to not be properly included among the consequences of a):

$$A \vdash a,$$
$$A \nvdash \bot,$$
$$B \nvdash a \text{ for } B \subsetneq A,$$
$$Cn(A) \nsubseteq Cn(a).$$

At large, such logic-based explanations aim to answer the following kinds of questions:

- What explains a given observation (o)?
- Which information logically entails a given inference (o)?
- Why is a given formula derived or not?
- Why is a set of formulas a solution?

Since such and similar explanations in logic-based reasoning are generally non-unique, preferred explanations are often selected using some specific ordering criteria [165] with various forms of minimality (e.g. cardinality, depth of proof/inference) being a frequent choice. Different works apply these logic-based explanation ideas for explainability in various settings, including the so-called Model Diagnosis (e.g. [60, 167, 175]), abductive reasoning (see Sections 3.1.5, 3.2.1) and non-classical logic-based approaches such as Logic Programming (LP, see Section 3.2). In this section we review some prominent MR explainability techniques from various classical logic-based approaches to reasoning.

3.1.1 Axiom Pinpointing

Axiom pinpointing [164] in description logics (which are decidable fragments of first-order logic) is a very good example of the well established logic-based explanation concepts still being very much relevant in modern MR explainability. Axiom pinpointing amounts to finding axioms in a knowledge base that entail or prevent a given consequence/query, whereby minimal such sets of axioms are taken

as justifications/explanations. Such explanations aim to answer the following question:

- given a knowledge base KB (possibly inconsistent, with an inconsistency-tolerant semantics) and a query Q (e.g. a Boolean conjunctive), why is Q (not) entailed by KB? [12, 19]

Works of [19, 39, 104, 114, 139] are examples of explaining knowledge base query answering where explanations are defined as (minimal) subsets entailing or contradicting a given query with respect to a (consistent or inconsistent) knowledge base. Roughly then, in our terms, given a knowledge base KB and query i, an explanation for the outcome $o \in \{y, n\}$ (representing 'yes' for 'entailed' and 'no' for 'not entailed', respectively) is a $\subseteq$-minimal $KB' \subseteq KB$ such that $KB' \models i$, if $o = y$, and a $\subseteq$-minimal $KB' \subseteq KB$ such that $KB' \cup \{i\} \models bot$, if $o = n$. Such explanations are **attributive**.

3.1.2 Constraint Programming (CP)

Constraint Programming (CP) [178] is a paradigm for solving combinatorial search problems, often called Constraint Satisfaction Problems (CSPs). CSPs are represented in terms of decision variables and constraints, usually in classical logic vocabulary, and solving them amounts to finding value assignments to variables that satisfy all the constraints.[7]

Already in mid-90s it was understood that explanations in CP cannot simply amount to tracing of solutions but need to be much more sophisticated. To quote from [82, p. 4860],

> A natural approach to providing a richer explanation of a solution would be to 'trace' the program's solution process. However, constraint solvers generally employ search, and tracing search tends not to provide a very satisfying explanation. For backtrack search: "I tried this and then that and hit a dead end, so I tried the other instead". Even worse, for local search: "I kept getting better, but then I tried some other random thing".

Apart from search, inference is a critical part of constraint solving and was used early for explainability. For instance, in [192], a "trace of the inference, with some rudimentary natural language processing, provided explanations for puzzles taken from newsstand puzzle booklets that were reasonably similar to the answer explanations provided in the back of the booklets."[82, p. 4860] Overall in CP, "much of the work on explaining failure actually is focused on programs explaining intermediate failures to themselves in order to reach a solution more efficiently."[82, p. 4860]. The questions that explanations in CP aim to answer include the following.

- Why is there a conflict between these parts of the system? [83]
- Which constraints result into failure? [111]
- Why does this parameter have to have this value? [83]
- What are the next propagation steps? [65, 208]
- Which choices should I relax in order to recover consistency? [7, 82]

[7] We note that it is also very natural to use LP (see Section 3.2) in solving CSPs.

- Which choices should I relax in order to render such a value available for such a variable? [7, 82]
- From which subsets of current choices did inconsistency follow? [7, 82]
- Why is this value not available any longer for this variable? [7, 82]

We review below several approaches in CP to devising explanations that answer such questions.

In his seminal paper [111] from 2004, Junker proposed QuickXplain – a general purpose technique for explainability in constraint programming. In a general setting of constraints (including e.g. CP, Satisfiability (SAT) and Beyond, description logics), relaxations (resp. conflicts) are defined as sets of constraints for which (resp. no) solution exists. Conflicts thus explain solution failure, relaxations restore consistency. However, both are generally exponential to construct and present, whereas a user may desire explanations pertaining to the most important constraints. Thus, preferred relaxations and preferred conflicts are defined to be minimal with respect to lexicographic orderings (over relaxations and conflicts) defined using any total order over constraints. In practice, this amounts to successively adding the most preferred constraints until they fail and then removing the least preferred constraints as long as that preserves failure. Explainability-wise, "preferred conflicts explain why best elements cannot be added to preferred relaxations".[111, p. 169][8] Further, the described "checking based methods for computing explanations work for any solver and do not require that the solver identifies its precise inferences."[111, p. 172] In a more general CP setting [162] similar to [111], O'Sullivan et al. propose representative explanations (and algorithms thereof) in which "every constraint that can be satisfied is shown in a relaxation and every constraint that must be excluded is shown in an exclusion set."[162, p. 328]

Other works, notably alldifferent [65], produce explanations for improving solver strategies. The motivation of Downing et al. is as follows [65, p. 116, emphasis original]:

> Whenever a propagator changes a domain it must explain how the change occurred in terms of literals, that is, each literal l that is made true must be explained by a clause $L \rightarrow l$ where L is a (set or) conjunction of literals. When the propagator causes failure it must explain the failure as a *nogood*, $L \rightarrow \bot$, with L a conjunction of literals which cannot hold simultaneously.

Roughly then, the propagator's actions can be explained using cut-sets, where an explanation is effectively a logical constraint on (the values of) variables. Similar ideas using a lazy clause generation solver for explaining propagation via constraints from which nogoods can be computed can be applied to string edit distance constraints, e.g. in [208] Winter et al. use explanations that consist of literals which logically entail the truth of a Boolean variable that encodes propagation of some variable's value.

[8] Effectively, such explanations can be seen as a form of Brewka's preferred subtheories [29] in the language of constraints.

In [7] Amilhastre et al. provide "explanations for some user's choices and ways to restore consistency", whereby "the user specifies her requirements by interactively giving values to variables or more generally by stating some unary constraints that restrict the possible values of the decision variables." The goal is to "provide the user with explanations of the conflicts" in terms of "(minimal) inconsistent subsets of the current set of choices." From a technical point-of-view, "in order to circumvent intractability from the practical side, [the] approach relies on a compilation of the original problem into a data structure from which much better performances can be obtained", specifically an "automaton that represents the set of solutions of the CSP."

O'Callaghan, O'Sullivan, and Freuder argue in [160] that "desirable is an explanation that is corrective in the sense that it provides the basis for moving forward in the problem-solving process" and formally define a corrective explanation intuitively as "a reassignment of a subset of the user's unary decision constraints that enables the user to assign at least one more variable", providing as well an algorithm for computing corrective explanations of minimal length.

Recently, there have also been works that consider a constraint solver assisting a human user in solving some logical problem. For instance, in [21] Bogaerts et al. "the propagation of a constraint solver through a sequence of small inference steps". They use minimal unsatisfiable sets of constraints for generating explanations of the solver's individual inference steps and the explanations can overall be seen as proofs or traces of the solver's working towards a solution.

Overall, explanations in CP can be roughly described as follows. Given a set KB of constraints, with i being the latest propagation, variable assignment or user added constraint(s), an explanation for the latest inference o (variable value restriction, e.g. $l = \top$ or $v \in [1, 3]$, or failure $\bot$) is a $\subseteq$-minimal $KB' \subseteq KB$ such that $KB' \cup \{i\} \models o$. This falls into the category of **attributive** explanations.

Instead, **contrastive** explanations that answer questions pertaining to indication of constraints or variable values leading to inconsistencies and consistency restoration can be defined as follows. Given $KB \cup \{i\} \models o$ and an alternative outcome $o' \neq o$, an explanation as a counterfactual can be a modification i' of i such that $KB \cup \{i'\} \models o'$. In particular, if the outcome $o = \bot$ means that i cannot be satisfied given KB (i.e. i cannot be extended to a solution of the CSP) then an explanation for an alternative outcome (i.e. where a solution can be found) is a minimal $KB' \subseteq KB$ such that $KB' \cup \{i\} \models \bot$ and a modification i' of i such that $KB \cup \{i'\} \not\models \bot$.

3.1.3 Satisfiability (SAT) and Beyond

Solving Satisfiability (SAT) problems [62] is a special case of CSP solving. Satisfying assignments are also referred to as *implicants*. *Prime implicants* are then implicants of minimal size, in the sense that they satisfy the formula using a minimal set of assigned variables. Prime implicants can thus be seen as **attributive** explanations. Indeed, Ignatiev, Narodytska, and Marques-Silva exploit prime implicants in [107] to define logical explanations and counterexamples with respect to ML model

classifications. Specifically, given a Mixed-Integer Linear Programming (MILP)/Satisfiability Modulo Theories (SMT) encoding of an ML model (exactly representing its behavior), an assignment to variables corresponds to an input/instance, with each variable corresponding to a feature. A satisfying assignment, for an ML model in which the prediction has been fixed, represents an instance which maps to that prediction. For a given instance then, a corresponding prime implicant explains the model's prediction/inference.[9]

As with general CP, another challenge is that of explaining inconsistency or unsolvability. Given an unsatisfiable formula, modern SAT solvers are able to report a proof of its unsatisfiability together with its support, also known as (unsatisfiable) core. A core is a subset of the original formula which is itself unsatisfiable. Although usually smaller than the original formula, a core might not yet be minimal. Instead, Minimal Unsatisfiable Subformulas (MUSes) are cores of minimal size, in the sense that every subset of a MUS is satisfiable. MUSes can thus be regarded as explanations of unsatisfiability [170] since they represent the culprits generating an inconsistency, and in fact suffice to generate a conflict in the formula, regardless of the rest. Multiple MUSes can exist in a formula and so multiple explanations abound, giving rise to both **attributive** and **contrastive** explanations, answering the following questions.

- Which information minimally entails a given inference?
- Which information prevents a given logical expression to be satisfied?

Analysis of inconsistent formulas plays an important role in MR explainability, since many, if not all, of the approaches for extracting classical logic-based explanations can be reduced to this problem. In addition, MUSes are widely used as building blocks of explanations in other MR approaches, for instance Planning (see e.g. [69, 70]), Argumentation (see e.g. [157]), Decision Theory (see e.g. [22]).

3.1.4 Automated Theorem Proving (ATP) and Proof Assistants

Automated Theorem Proving (ATP) is a procedure whereby a tool known as a "theorem prover" is provided a proposition, and it returns 'true' or 'false' (or runs out of time). Theorem provers have matured tremendously in recent years, and are now used in many settings, not only for arithmetic. In general, SAT solvers, SMT solvers, and model checkers also fall under the ambit of theorem proving. If a theorem prover returns 'false', it also generates a falsifying counterexample, as a contrastive explanation to the user as to why the proposition does not hold. However, most theorem provers provide no explanation for a proposition which is verified to be true [79, 195].

Proof assistants are special kinds of theorem provers (often called interactive theorem provers). These are hybrid tools that automate the more routine aspects of building proofs while depending on human guidance for more difficult aspects. One can write a theory of one's choice, and verify

[9] In the basence of logical representation, one can instead define prime implicant explanations for ML classification directly as minimal sets of features of an instance that suffice to yield the classification, see e.g. [185].

whether or not a proposition holds of this theory. Well known examples of general proof assistants are Isabelle[10], F*, Coq[11] etc. [16]. Specialized proof assistants exist as well, such as Tamarin for security protocol verification. The user can write any theory supported by a proof assistant and verify a proposition in that theory, rather than being limited to some theory of the designer's choice. Proof assistants are being widely used for a variety of applications, including building verified compilers [122] and verified implementations of processors [171].

However, a proof assistant might return a true/false answer, run out of time, or stop at a subgoal that it does not "know" how to solve. In the last case, the user might need to write helper code and expand the theory. In essence, proof assistants are interactive, that is, they often require more input from the user to solve "difficult" goals. If a proof assistant returns 'true', it generates a sequence of steps that one can then use to replicate the proof by hand for better understanding. Some specialized proof assistants generate counterexamples if the proposition is found to be 'false', but not all of them do – Coq, for instance, would just show the user some pending goals which it cannot solve (and the statements corresponding to these goals might themselves be false), or throw an error message to say that the proposition cannot be proved true by some underlying decision procedure.

In sum, explanations from theorem provers aim at answering the following questions.

- If a proposition is true, what is a (shortest/most readable) proof?
- If a proposition is false, what is a counterexample?

The first type of explanations are **attributive**, whereas the second are **contrastive** (as counterexamples).

There are obviously open problems pertaining to explainability in ATP, to name a few. 1. Can a theorem prover/proof assistant also be optimized to provide the "best" counterexample, according to some measure? 2. Can non-interactive theorem provers also provide explanations/proof descriptions for propositions that are verified to be true? What would these explanations look like? 3. If a (sub)goal involves an unsolvable loop which leads to a timeout, can this be output as an explanation instead of (or in addition to) timing out? Furthermore, currently the explanations generated by most theorem provers and proof assistants are not very human-friendly. Some proofs/counterexamples generated by some theorem provers can take hundreds of lines, making it difficult for a human to use these to understand the underlying (mal)functioning of the system. There is some work [85] along the lines of designing provers which produce proofs that look like ones humans might write, but only for very specific domains.

In terms of AI-to-AI or machine-to-machine explanations, one can consider the example of proof-carrying code, where an application downloaded from an untrustworthy location comes with a proof of its "correctness" for the host system to verify before installation. No human intervention is needed in order to either generate the proof or to verify it. This can be considered an example of the general

[10] https://isabelle.in.tum.de/
[11] https://coq.inria.fr/

idea of an interactive proof system [92].

An interactive proof system models computation as the exchange of messages between two parties: a prover and a verifier. The prover has vast computational resources, but cannot be trusted, while the trusted verifier has bounded computation power. As the name suggests, the prover tries to prove some statement to the verifier. Interactive proofs often proceed in "rounds", where the parties send messages to each other based on previous messages they have received, till the verifier is "convinced" of the truth of the statement. One can also have one-round (often referred to as non-interactive) proofs where the prover only needs to send one message to convince the verifier of a true statement, and no further rounds of interaction are necessary.

Interactive proof systems have two requirements: soundness and completeness. Essentially, soundness claims that no prover – even a dishonest one – can convince the verifier of a false statement. Completeness says that for every true statement, there is a proof that the prover can produce to the verifier to convince it. The verifier may choose to "probe" various parts of the proof sent by the prover to convince itself of the verity of the statement. This is often done in an efficient manner by picking random bits and using them to identify which parts of the proof to inspect. Depending on what the abilities of the prover and the verifier are, one can get different classes of proofs. One can get (slightly) different systems based on whether the random values chosen by the verifier are made public or kept private. If one assumes the existence of special objects like one-way functions, one can construct "zero-knowledge proofs", where the verifier is convinced exactly of the intended statement, but no further information about said statement is revealed. One can also have systems where multiple provers can interact with the verifier (but not with each other) to prove a statement.

Interactive proof systems can be seen as an excellent example of AI-to-AI explainability approaches that have existed for a long time. However, as previously noted, there are still plenty of interesting open problems in this area.

3.1.5 Abduction

Abductive reasoning, e.g. [113, 118, 133], aims at explaining phenomena such as observations, abnormalities, anomalies, false predictions. The notion of an explanation follows closely that of explanations in classical logic-based reasoning delineated above, but in addition allows for inventing, i.e. abducing, additional knowledge. This abduced knowledge, possibly together with existing background knowledge, is seen as an explanation in that it allows to deduce a given phenomenon. Minimal (in various forms) sets of abducibles as explanations can be used for explaining abductive reasoning via causal knowledge graphs [200], or, as in e.g. [15, 107], for explaining ML model classifications by logically encoding ML models and abducing minimal assignments representing feature attributions that guarantee classifications.

Explanations via abducibles aim at answering the following kinds of questions.

- What information would entail this observation? [118]
- Why was the wrong belief or expectation formed? [133]

Abducibles as explanations are thus largely **attributive**: given background knowledge KB and a (possibly hypothetical) inference o, the abducible knowledge ab is an explanation for o just in case $KB \cup \{ab\} \models o$. However, hints of contrastive and actionable explanations appear in abducibles in that they indicate knowledge that, if present, would allow to achieve a hypothetical/alternative outcome/inference.

3.2 Non-classical Logic-Based Reasoning (Logic Programming)

Logic Programming (LP, see e.g. [11, 119]) is both a knowledge representation formalism and computational mechanism for non-classical, particularly non-monotonic, reasoning. Often, explainability in LP is enabled by the declarative reading of rules in logic programs, which allows for **attributive** explanations in terms of knowledge and rules that yield a specific inference. Further, if the knowledge and rules carry immediate semantic meaning, then deductive inference paths or proof trees are readily interpretable and can be translated into natural language, as e.g. in Rulelog [94, 95]. Still further, **contrastive** explanations can be obtained by inspecting conflict resolution strategies.

In general, LP takes various forms, most notably abductive, inductive and answer set programming, each with different computational procedures and different approaches to explainability. We discuss some of these below.

3.2.1 Abductive Logic Programming (ALP)

Abductive Logic Programming (ALP) [71, 113] is a form of abductive reasoning expressed using LP vocabulary. Abductive logic programs are used for knowledge representation and abductive proof procedures for automated reasoning. Abductive proof procedures interleave backward and forward reasoning and can be used for checking and enforcing properties of knowledge representation (via queries or inputs to the program) as well as for agents to abduce and/or explain actions required to check/enforce such properties.

Formally, given an abductive logic program P, an abductive explanation for an observation o (conjunctive formula) is a pair (D, θ) consisting of a (possibly empty) set D of abducibles and a (possibly empty) variable substitution θ for the variables in o such that P together with abducibles D entail $o\theta$ and satisfy integrity constraints. Different notions of entailment and satisfaction can be adopted, for example classical first-order entailment $P \cup D \models o\theta$ and consistency $P \cup D \not\models \bot$.

Intuitively, abductive explanations answer the following questions [113, 199]:

- Why did this observation occur?
- What explains this observation?
- How to reach this goal/query?

Accordingly, as in Section 3.1.5 Abduction, abductive explanations in ALP are **attributive**, with flavors of actionability: "explanations can be thought of as data to be actually added to the beliefs of the agents and actions that, if successfully performed by the agents, would allow for the agents to achieve the given objectives while taking into account the agents' beliefs, prohibitions, obligations, rules of behavior and so on, in the circumstances given by the current inputs."[199, p. 100]

3.2.2 Inductive Logic Programming (ILP)

Inductive Logic Programming (ILP) [153] studies the inductive construction of logic programs from examples and background knowledge. Briefly, "given a set of positive examples, and a set of negative examples, an ILP system constructs a logic program that entails all the positive examples but does not entail any of the negative examples."[72, p. 1] The set of induced rules, called hypotheses, possibly together with a proof trace, is viewed as an explanation of the examples in the context of the background knowledge: given background knowledge KB and sets P and N of positive and negative examples, respectively, a set R of clauses (i.e. hypotheses) is an explanation for given examples just in case $KB \cup R \vdash p$ for all $p \in P$ and $KB \cup R \nvdash n$ for all $n \in N$. Hypotheses as explanations thus aim to answer the following question:

- What general hypothesis best explains the given specific examples/observations?

As such, explanations in ILP are closely related to explanations in ALP. The two LP techniques can indeed be considered complementary [153], noting that "abduction is the process of explanation – reasoning from effects to possible causes, whereas induction is the process of generalization – reasoning from specific cases to general hypothesis."[132, p. 205] Some recent works, e.g. [72, 145], use integrated ILP and ML techniques to learn explanatory logic programs from non-symbolic data. Overall, ILP explanations as hypotheses or proof paths are **attributive** in nature.

3.2.3 Answer Set Programming (ASP)

Fandinno and Schulz provide an excellent overview in [78] of explanations in Answer Set Programming (ASP) [30, 136], which have been researched for some 25 years. There are two main families of approaches, namely justification and debugging. The respectively aim at answering these kind of questions:

- Why is a literal (not) contained in an answer set?
- Why is an unexpected or no answer set computed?

Both families first-and-foremost provide **attributive** explanations, albeit with different flavors: justifications can be inspired by, for instance, causal or argumentative reasoning; debugging can be based, for instance, on reporting unsatisfied rules or unsatisfiable cores. Both justification and debugging approaches may also supply **contrastive** explanations by including conflicting information and revealing conflict resolution that takes place in reasoning. We briefly summarize the main ideas below,

following [78].

Justification approaches by and large concern consistent logic programs and provide "somewhat minimal explanation as to why a literal in question belongs to an answer set". Off-line justifications [166] are graph structures describing the derivations of atoms' truth values via program rules and can be seen to provide traces of dependencies. Labelled Assumption-Based Argumentation (ABA)-based answer set justifications (LABAS) [180, 181] abstract away from intermediate rule applications and focus on the literals occurring in rules used in the derivation and can be seen to provide traces via (supporting and attacking) arguments (see Argumentation), thus exhibiting reasons pro and con the inference in question.

In a similar vain, causal graph justifications [34] associate with each literal a set of causal justifications that can be graphically depicted and can be seen as causal chain traces. Causal graph justifications are inspired by why-not provenance [202], which itself provides non-graphical justifications expressing modifications to the program that can change the truth value of the atom in question. (By extension, justifications for the actual truth values do not imply any modifications.) These can be seen to approach the realm of actionable explanations, though unlike causal graph justifications, why-not provenance does not discriminate between productive causes and other counterfactual dependencies (see e.g. [98, 99]).

Debugging approaches instead by and large concern inconsistent logic programs and provide explanations as to why a set of literals is not an answer set. The spock system [28, 87] transforms a logic program into a meta-(logic)program expressing conditions for e.g. rule applicability and whose answer sets capture violations of the given candidate answer set of the original program in terms of rule satisfaction, unsupported atoms or unfounded loops. The Ouroboros system [161] extends these ideas to logic programs possibly with variables, tackling also the issue of multiple explanations by requiring the user to specify an intended answer set. Instead, [64, 184] propose interactive debugging whereby the user is queried for about specific atoms to produce the relevant explanations of inconsistencies.

The above summary shows the wealth of explainability techniques in ASP, but there are also works that use ASP for explanations in other areas. For instance, in [35] Calegari et al. map decision trees (DTs) into logic programs to explain DT predictions via natural language explanations generated from logic program rules.

In another recent research direction [15], Bertossi uses ASP to generate **contrastive** explanations for discrete, structured data classification. Briefly, there: a) causal explanations are sets of feature–value pairs such that changing at least one value changes the classification label; b) counterfactual (value-)explanations are individual feature-value pairs such that changing the value changes the classification; c) actual (value-)explanations are individual feature-value pairs together with a set of feature-value pairs such that changing the value of the former does not suffice to change the classification, but changing the values of both the former and the latter does. Actual and counterfactual explanations can be assigned explanatory responsibility measure, which amounts to the inverted size of the small-

est set accompanying an actual explanation, with the explanatory responsibility of a counterfactual explanation always being 1 (because it suffices to flip only that feature-value to change the classification). Overall, Bertossi proposes to encode the inputs-outputs of an ML classifier into an ASP program, together with predicates and constraints capturing interventions (i.e. feature-value flipping) to extract the various explanations defined by computing answer sets.

3.3 Argumentation

"Computational Argumentation is a logical model of reasoning that has its origins in philosophy and provides a means for organizing evidence for (or against) particular claims (or decisions)."[186, p. 277] Moulin et al. review in [151] a large body of literature on MR explainability and argue for the use of argumentation to support interactive and collaborative explanations of reasoning that take into account the aspects of justification as well as criticism of claims/decisions/solutions. Indeed, "[t]here is a natural pairing between Explainable AI and Argumentation: the first requires the need to explain decisions and the second provides a method for linking any decision to the evidence supporting it."[186, p. 277]

By and large, data and knowledge in argumentation formalisms can be represented using various forms of directed graphs, whereby nodes represent arguments modelling individual pieces of information and edges represent relationships among arguments (e.g. attacks for conflicting information, supports for supporting information). Reasoning then amounts to find the sets of "good" arguments, for instance arguments acceptable under some semantics, i.e. a collection of formal criteria such as not attacking each other and defending against all attackers. (See Figure 1 (i).)

A common approach to explaining decisions in argumentation, one which encompasses both justification and criticism, essentially amounts to traversing (a part of) an argument graph to show the attacking and defending arguments (together with their relationships) relevant to the decision, where the decision essentially amounts to accepting (resp. rejecting) an argument as (resp. not) "good". In addition, one can speculate what kind of changes to the argument graph, such as additions or removals of arguments or relationships, would result into different acceptance status(es) of the argument(s) in question. Argumentative explanations thus aim to answer the following questions:

- Given a set of arguments, why is a particular argument a "good"? [76]
- What are the reasons (i.e. other arguments) for and against accepting a?
- Can the given reasons for acceptance of a be contested?
- Which arguments/relationships should be removed or added to accept argument a? [77, 179]

Intuitively, to explain acceptability of an argument x, "one would need to show how to defend x by showing that for every argument y that is put forward (moved) as an attacker of x, one must move an argument z that attacks y, and then subsequently show how any such z can be reinstated against attacks (in the same way that z reinstates x). The arguments moved can thus be organised into a graph

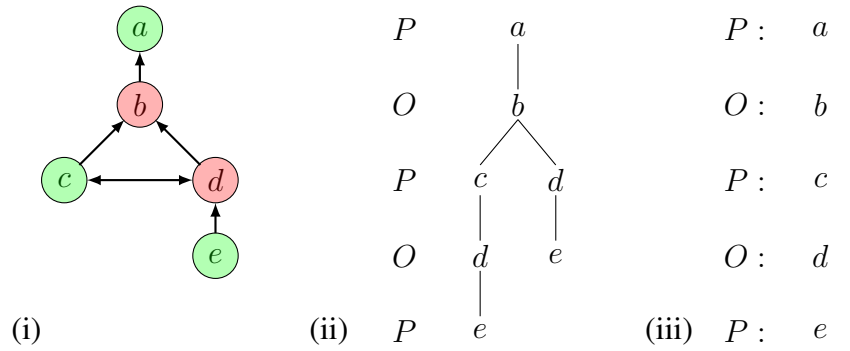

Figure 1: Adapted from [147, Fig. 6.3, p. 114]. (i) An AA graph with nodes as arguments, labelled a, b, c, d, e, and directed edges as attacks. Arguments accepted (resp. rejected) under the grounded extension semantics [66] are colored green (resp. red). (ii) Grounded dispute tree between the proponent P and opponent O, for the topic argument a. The proponent can successfully defend the claimed argument a against all the counterarguments. (iii) A dialogue explanation as to why a is accepted, to be read thus: "a is objected only by b, but is in turn defended by c; c is objected only by d, but is defended by the unobjectionable e". The explanation reveals all the pros and cons relevant for acceptance of a.

of attacking arguments that constitutes an explanation as to why x is [acceptable]"[148, p. 109].

So, given a graph G representing an argumentation framework and a query regarding a statement (e.g. (the claim of) an argument) i, an explanation for the acceptance status o of i is a minimal (in some sense) subgraph $G' \subseteq G$ consisting of arguments for and counterarguments against o that (fully or partially) determine o. The subgraph G' can be turned into a formal dispute tree [67, 148] where two fictional players – proponent P and opponent O – put forward arguments in favor (pro) and against (con) a topic argument, see Figure 1 (ii). This can accompanied by dialogical explanation to the effect of establishing o by using the arguments for (pros) to defend from the counterarguments (cons), see Figure 1 (iii). Such explanations are **contrastive** and answer the first two questions above. Relatedly, an explanation can be a modification G' of the original argument graph G such that the desired acceptance status o of i is achieved. Such explanations can be seen as **actionable** and answer the last two questions above.

In practice, constructing argumentative explanations amounts to one or both of the following two phases.

- Extract a connected subgraph (e.g. a path, branch, cycle; possibly depth-, width- or otherwise bounded) that consists of the relevant arguments and counterarguments together with their acceptability statuses, and potentially indicate arguments and/or relationships that if added and/or removed would change the acceptability statuses. Examples include [58, 59, 75, 77, 115, 172, 180, 181, 183, 210, 212].
- Construct a formal dialogue [168] or an argument game [143, 148] in "which participants engage in structured, rule-guided, goal-oriented exchange [of arguments] according to specific protocols."[186, p. 208] Such dialogue games underly argumentation-supported dialogi-

cal/conversational explanations [141, 186, 205, 206]. The "winning" arguments are determined by argumentation semantics, so such dialogue games show how one player justifies the decision while the other criticizes it. Examples include [12, 50, 51, 55, 57, 67, 74, 76, 77, 148, 181].

These ideas are not limited to various argumentation formalisms. On the one hand, they can be directly applied to reasoning formalisms that can be captured or reinterpreted in argumentative terms:

1. In [24] Booth et al. construct explanatory dialogues for logic programs using abstract argumentation (AA, [66]).
2. In [180, 181] Schulz and Toni use correspondence between answer sets in Answer Set Programming (ASP) and stable extensions in Assumption-Based Argumentation (ABA, [23, 56]) to explain why a literal is or is not contained in an answer set of a logic program. The idea behind an explanation, or rather a justification, "for a literal l is to find the underlying literals, necessary to derive l, as well as conflicts with other literals which influence whether or not l is part of an answer set."[180, p. 3]
3. In [86] García and Simari define argumentative dialogical explanations for queries in Defeasible Logic Programming (DeLP).
4. In [213] Zhong et al. map decision frameworks into ABA frameworks so that best decisions in the former correspond to acceptable arguments in the latter, and use ABA dispute trees as explanations.
5. In [74] Fan captures planning problems in ABA and extracts explanations pertaining to actions and/or their preconditions and goals relevant to the (in)validity of a plan.

On the other hand, ideas based on argumentative and dialogical explanations can be applied to explain reasoning in other fields:

6. In [183] Šešelja and Straßer extend AA frameworks with explanatory and incompatibility relations to define explanations as explanatory paths in conflict- and incompatibility-free preferred extensions, applying this to scientific debates.
7. In [31] Briguez et al. propose an argumentation-supported recommender system using DeLP, which allows to naturally extract explanations in terms of reasons for and against a recommendation. Similarly, in [50, 172] the authors propose and evaluate empirically a gradual argumentation-based recommender system where items and their aspects are captured via arguments, gradual semantics is used to determine the recommendations and explanations amount to argument subgraphs that are turned into natural language explanations.
8. In [12, 101] the authors use variants of dispute trees and argumentation dialogues for explaining answers to queries in ontology-based knowledge bases.
9. In [197] Timmer et al. extract probabilistically supported arguments from a Bayesian network to construct a support graph and, given a set of observations, build arguments from that support graph. Such arguments can facilitate the correct interpretation and explanation of the evidence modelled in Bayesian networks.

10. In [186] Sklar and Azhar illustrate argumentative natural language explanations arising from argumentation dialogue games in a collaborative human-autonomous agent scenario.
11. In [49] Cocarascu et al. use machine learning (autoencoders) to populate case-based reasoning (CBR)-inspired argumentation frameworks for data classification/prediction. There, explanations can be extracted in two equivalent ways: dialectical explanations as subgraphs with arguments comprising relevant data instances with known labels that support and/or contrast with the prediction; rule-based explanations as logic programming (LP) rules (with exceptions) comprised of features. Cocarascu et al. further such dialectical explanations in [51] to classification of categorical data, annotated images and text.
12. In [182] Sendi et al. extract probabilistic classification rules from deep neural networks which then constitute arguments explaining any given classification.
13. In [57] Čyras et al. use CBR-inspired, AA-driven reasoning formalism to classify and explain legislative data.
14. In [58] Čyras et al. use AA frameworks to capture schedules and their properties for defining, extracting and presenting [59] explanations in makespan scheduling.

Other forms of argumentation-enabled explainable reasoning have been proposed too: (i) In [52] Collins et al. propose structured argumentation as a good candidate for representing causality and forming as well as communicating explanations in the planning domain; (ii) in [173] Raymond et al. propose an argumentation-based architecture for designing explainable human-agent systems for deconfliction environments.

3.4 Decision Theory

Decision theory gathers different domains such as Multi-Criteria Decision Making, Decision Making under Uncertainty and Computational Social Choice (SC). "The typical decision problem studied in decision theory consists in selecting one alternative among a set X of candidate options, where the alternatives are described by several dimensions. This selection is obtained by the construction of a preference relation over X."[125, p. 1411] The final part of the decision process, i.e. explaining the outcome of the decision model to the user, is by and large not readily supported by decision theory models due to their complexity. It is nonetheless arguably as important as the recommendation of the outcome itself (see e.g. [14, p. 152]) and attempts at answering the following kind of question (see e.g. [22, 125]):

- Can this recommendation (i.e. the chosen alternative) be justified under the given preference profile?

Explainability in decision theory has been researched in modern MR. On the one hand, forms of Argumentation have been proposed for explainable decision making, see e.g. [5, 6, 212, 213]. Specifically, in [5, 6] the authors propose to use abstract argumentation with preferences for multiple

agents to argue about acceptable decisions. There, explainability is assumed to arise naturally from the argumentative process, but is not specified at all. Instead, Zhong et al. map decision frameworks into Assumption-Based Argumentation and use dispute trees (see Section 3.3) to generate explanations as to why a decision is best, better than or as good as others. They also translate such **contrastive** dialogical explanations into natural language and perform an empirical user study in a legal reasoning setting to evaluate their approach. These works effectively define argumentation-supported decision making procedures where explainability arises from the argumentative methods employed.

On the other hand, Labreuche in [125] provides solid theoretical foundations to produce explanations for a range of decision-theoretic models via provision of arguments, but not using formal argumentation. There, arguments, and therefore explanations, essentially are "based on the identification of the decisive criteria."[125, p. 1415] They are thus **attributive**, supplying justifications in terms of higher-level criteria that support the recommended decision.

Similar in spirit, Labreuche, Maudet, and Ouerdane in [127] consider the setting of qualitative multi-criteria decision making with preferential information regarding the importance of the criteria and preference rankings over different options (choices). They define explanations as collections of factored preference statements (roughly of the form 'on criteria I, option o is better than options P') that justify the choice of the weighted Condorcet winner. Such explanations pertain to the relevant data (problem inputs) supporting a proof that the recommended decision/choice is the best one.

Some more recent works essentially apply to various settings the same conceptual idea that explanations of decisions made pertain to important criteria or principles based on which the model makes a decision. For instance, in [159] explaining amounts to identification of decisive criteria as sets of attributes that are most important for preferring one option over another. Nunes et al. thus propose attributive explanations via several decision criteria in a quantitative multi-attribute decision making setting. Belahcene et al. deal with incomplete preference specifications in [14] and thereby use preference swaps to explain/justify decisions. Intuitively, their explanations transform a complex preference statement (over many attributes) that needs to be understood by the user into a series of simpler preference statements (over few attributes). In [126], Labreuche and Fossier consider hierarchical models of multi-criteria decision aiding. They use Shapley values to define axiomatic indices regarding the influence of different criteria and provide attributive explanations pertaining to the importance of different criteria. In a setting where different audiences may adhere to different norms [22], Boixel and Endriss explain the decision making outcome by presenting axioms (with respect to audience's norms) for which no voting rule would yield a different outcome. We deem such explanations pertaining to the importance of decision-making criteria to be **attributive**.

3.5 Planning

Automated planning (or AI planning) is a class of decision making techniques that deals with computing sequences of actions towards achieving a goal. The solution to a planning problem is found by a planner, typically one based on heuristic search. Chakraborti et al. review in [44] the most recent advances in explainable AI planning. They emphasize the importance of user modelling, or "persona" of the explainee. Although their classification is broad – in terms of end-user, algorithm designer and model designer personas – it is clear that more granular models are possible. Their work reinforces Miller's view on the characteristics of effective explanations [144], namely that explanations should be "social in being able to model the expectations of the explainee, selective in being able to select explanations among several competing hypothesis, and contrastive in being able to differentiate properties of two competing hypothesis."[44, p. 4805] Further, Chakraborti et al. also point to the use of abstractions as a means to provide effective explanations. They go on to provide a useful categorization of recent approaches to explanations in AI planning, based on whether explanations reveal the working of an algorithm, details of the underlying model, or attempt to reconcile the differences between the user and system's mental models. The kinds of questions that explanations in planning aim to answer pertain to goodness of plans, alternative choices and unsolvability, and are as follows:

- What changes (in the current state) would make this solvable? (I.e. making excuses.) [91]
- What are possible reasons you could not compute a plan from your current state to the given goal state? [91]
- Why a plan fails to be a solution, which actions are invalid? [74]
- Why fail and what (temporal) repercussions does the first failure have? [201]
- Why did the agent take action A (at that time) rather than action B (resp. earlier or later)? [38]
- Why was a particular predicate (agent/object) involved in the plan? [38]
- How good is a given plan from the point of view of the human observer (their computational model)? [42]

To answer questions about unsolvability and failures, some approaches provide **contrastive** explanations in terms of a part KB' of the model KB that together with the initial state i and the desired goal(s) o lead to unsolvability $\perp$. For instance, Fan in [74] captures STRIP-like planning in Assumption-Based Argumentation and extracts explanations pertaining to actions and/or their preconditions and goals relevant to both validity and invalidity of a plan. The unsolvability of planning tasks can also be explained to the user by pointing out unreachable but necessary goals in terms of propositional fluents, assuming appropriate abstractions of the user's model of the problem, e.g. [193]. Similarly, pointing out actions executed in a faulty node and their propagations (subsequent failed actions and their relationships) can be used to define explanations in temporal multi-agent planning [201]. Such explanations mostly work by attributing counterexamples to solvability in the current system $\mathbb{S}$. However, counterfactual contrastive explanations in terms of excuses – minimal, restrictive

changes to the initial state i that would allow to reach the desired goal o – are also possible [91].

Regarding alternative courses of action, **contrastive** explanations can be supplied that pertain to higher-level properties satisfied or not by the (current or alternative) plan. The general idea in answering questions about a given plan π not satisfying some property p is to produce a new plan π' that does satisfy p and compare the two plans [69]. For instance, in [68] Eifler et al. develop a means to qualify how good or poor a plan is, beyond the obvious properties such as cost or plan length. Such plan-property dependencies allow oversubscribed goals to be reasoned over, and explained by an agent. Specifically, in [69] Eifler et al. define explanations via plan-property entailment of soft goals so that an answer to the question "why is a property (set of soft goals) not achieved?" amounts to exhibiting other properties (sets of soft goals) that would not be achieved otherwise. In the specific case of goal exclusions, explanations amount to $\subseteq$-minimal unsolvable goal subsets. [70] is an extension of this approach to plan properties formulated in Linear Temporal Logic (LTL).

Another prominent challenge in explaining planning that has recently attracted a fair amount of research interest is that of model reconciliation [40, 41, 42, 43, 124], aiming to answer the last question above. Here the assumption is that humans have a domain and task model that differs significantly from that used by the AI agent, and explanations suggest changes to the human's model to reconcile the differences between the two. The objective of model reconciliation is not to completely balance the information asymmetry, but is selective to knowledge updates that can minimally cause human-computed plans to match those computed by the system. Such explanations are **contrastive** too [44], in that they contrast the models of the AI system and its user and aim to bring modification to the latter so as to convince the user of the goodness of the plan. Relatedly but orthogonally, Krarup et al. show in [120] how user constraints can instead be added to the formal planning model so that contrastive explanations as differences between the solutions to the initial and the new model can be extracted. We note that model reconciliation is in some aspects closely related to plan/goal recognition [36]. Essentially, recognizing the AI agent's goals may be seen as a prerequisite to providing rationale for its behavior. And the other way round, recognizing the user's plans and goals can serve as a means to improve explanations.

3.6 Multi-Agent Systems

Explainability in AI-supported Multi-Agent Systems (MAS) (sometimes also called Distributed AI [134]) concerns interactions among multiple intelligent agents so as to agree on and explain individual actions/decisions. Roughly, the questions of interest can be posed as follows:

- How were the user's (and the interacting agents') preferences taken into account when making a decision?
- What is the user's satisfaction with the decision?

There are a few works in MR, specifically in Argumentation (see Section 3.3), that aim to address

at least the first question above. On the one hand, in [5, 6] the authors propose to use abstract argumentation with preferences for multiple agents to argue about acceptable decisions. Explainability is assumed to arise naturally from the argumentative process, via relationships among arguments, given preferences and argumentation semantics. On the other hand, Raymond et al. propose in [173] an argumentation-based architecture for designing explainable human-agent systems for deconfliction environments. There, agents exchange conflict-free sets of arguments/rules in a dialogue D and an explanation of some topic argument a is a set S of related-admissible arguments [76] that recursively defend a. The approach focuses "on generating post-hoc explanations derived from the history of a dialogue D". These are examples of **contrastive** argumentative explanations in MAS settings.

However, more generally and with respect to the user's satisfaction, Kraus et al. stipulate in [121] that there has so far been little explainability in multi-agent environments. They claim explainability in MAS is more challenging than in other settings because "in addition to identifying the technical reasons that led to the decision, there is a need to convey the preferences of the agents that were involved." Murukannaiah, Ajmeri, Jonker, and Singh echo this concern in [154] from the points-of-view of multi-agent ethics, fairness etc. Several recent works suggest ways to address the challenges. At a high-level, in [46] Ciatto et al. propose to integrate symbolic and connectionist approaches via a multi-agent system to achieve explainability. More specifically, Kraus et al. propose to use ML for generating "personalized explanations that will maximize user satisfaction", which necessitates collecting "data about human satisfaction from decision-making when various types of explanations are given in different contexts."[121, pp. 13534-13535] Along similar lines, in [8] Amir et al. suggest research directions for agent strategy summarization to complement explainability in MAS. It remains to be seen what lines of research will be instigated by these recent calls to renew interest in MAS explainability, and whether MR, for instance the argumentation-based approaches discussed above, will play a significant role.

3.7 Reinforcement Learning

Reinforcement learning (RL) has recently become visible as a promising solution for dealing with the general problem of optimal decision and control of agents that interact with uncertain environments. Application areas range from telecommunication systems, traffic control, autonomous driving, robotics, economics and games. In the general setting, an RL agent is usually operating in an environment repetitively applying one of the possible available actions and receives a state observation and a reward as feedback. The goal of the RL framework is to maximize the overall utility over a time horizon. The choices of right actions are critical; while some actions exploit the existing knowledge, some actions explore to how to increase the collected reward in future, at the cost of performing a locally sub-optimal behavior.

Explaining control decisions produced by RL algorithms is crucial [2], since the rationale is often

obfuscated, and the outcome is difficult to trust for two reasons: 1. lack of coverage in the exploration, and 2. generalizability of the learned policies. Explainability in RL is complicated due to its real-time nature, since control strategies develop over time, and are typically not evaluated over snapshots. Several techniques for explanations are proposed in [2] such as Bayesian rule lists, function analysis, Grammar-based decision trees, sensitivity analysis combined with temporal modeling using long-short term memory networks, and explanation templates. Albeit such techniques are relevant as early attempts towards explaining RL decisions, we do not review them in this paper since we do not think they fall within MR. Instead we next briefly discuss some approaches that use symbolic techniques for explainability in RL.

Approaches in the literature that integrate the RL framework with explanations have been investigated in [84, 112, 117]. In particular, the authors in [84] introduce a framework of instructions-based behavior explanation in order to explain the future actions of RL agents. In this way, the agent can reuse the instructions from the human which leads to faster convergence. In [112], the method of reward decomposition is proposed in order to explain the actions taken by an RL agent. The idea is to split the reward in semantically meaningful types such that such that the action of the RL agent can be compared with reference to trade-offs between the rewards. The study in [117] deals with the general framework of explaining the policies over Markov Decision Policies (MDP). Under the assumption that the MDP is factored, a subset a minimum set of explanations that justify the actions of MDP is proposed. We believe these type of justifying explanations are a form of **attributive** explanations.

There is recent work on **contrastive** explanations in RL too. Specifically, in [142] Madumal et al. define causal and counterfactual explanations for MDP-based RL agents given their action influence models (which extend structural causal models [163] with actions). Specifically, they define a (complete) causal explanation for an action A "as the complete causal chain from A to any future reward that it can receive", with a minimal such explanation omitting intermediate nodes in such a chain, leaving source and destination nodes only. Further, they define a (minimally complete) explanation as the difference between the actual causal chain for the taken action A, and the counterfactual causal chain for some other action B. They show experimentally with human users that their explanations are subjectively good enough and help the users to better understand the RL agent's actions. However, the method requires a correct model of the world to be given upfront.

3.7.1 Constrained RL

State-of-the-art RL is associated with several challenges: guaranteed safety during exploration in the real world is one of them, and intricacy of reward construction is another. Many recent works introduced preliminary results on mitigating these challenges through formal methods [4, 108, 109, 135]. The most related ones to MR focus on shielding or constraining exploration in general or in various specific contexts, presenting objectives in the form of a linear temporal logic (LTL) formula

[4, 109]. Recent works also include the design of control policies using RL methods to find policies which guarantee the satisfaction of properties described in temporal logic [26, 108, 135].

3.7.2 Multi-Agent RL (MARL)

When it comes to Multi-Agent RL (MARL) frameworks, the main challenge to be addressed is the dependency between the action and rewards of different agents in the environment [131, 211]. It is natural when two or more agents are trying to optimize their local behavior over a horizon of time, and conflicts might occur with respect to the team or global behavior of the agents. In such a setting, different local optimal actions might lead to conflicting collaborative behavior. Thus, such scenarios render the collaborative reward design challenging, and new algorithms should be designed in order to address such problems. The MARL scenario imposes additional constraints and an efficient way to handle them is to use a symbolic framework by presenting the constraints in a more convenient ways, such as logical description. Kazhdan, Shams, and Liò in [116] provide such model extraction techniques that enhance explainability of MARL frameworks.

3.8 Causal Approaches

Causal models (see e.g. [88, 128, 163]) are useful in guiding interventional decisions and analyzing counterfactual hypothetical situations. Using causal models one can not only provide a decision but also provide a basis for what-if analysis, thus providing explanations. Lacave and Díez provide in [128] a summary of the work done on explaining AI models where the models are causal, quite often Bayesian networks. The authors distinguish three classes of explanations. 1. Explanation of evidence – this is basically abduction where the explanation is finding most probably explanations of variables not directly observable, based on the observed evidence variables. This can be seen as a form of **attributive** explanations. 2. Explanation of the model – this is simply a description (graphically or in text) the causal model. 3. Explanation of reasoning – here the objective is to explain the reasoning process for the result obtained, for specific results not obtained, or for counterfactual reasoning. We see these largely as **contrastive** explanations.

As regards the first class, [88] can be seen as an early example. There, Geffner equates "being caused" with "being explained". In contrast to earlier explanation techniques that used logical derivations (roughly as antecedents of rules that explain the consequents), Geffner augments default theories [174] with a causality/explanations operator C which is used to define an order over classes of models of default theories in terms of explained abnormalities (ab predicates). The kind of rule-based attributive explanations therefore have an abductive flavor. [156] is instead a modern MR example to explaining causality. Nielsen et al. show that given a set of variables, the values of which an explanation is sought after, it is possible to determine a set of other variables (explanatory variables)

which probabilistically explain the given set of observed variables. Some of these are possibly observed while the others are not. Indeed this includes abducing some variables from others. The use of a causal Bayesian network to trace these other variables is particularly interesting in the case of causal explanations.

Nielsen et al. further show how to use the interventional distribution on a Causal Bayesian network to compute a causal explanation, thus approaching the realm of contrastive explanations. Another recent example of attributive explanations with a counterfactual flavor is the work of [142] which uses structural causal models to explain actions of RL agents. In principle, to support counterfactual reasoning and thus contrastive explanations, one can use pure regression techniques factoring in time to do some causal correlation For instance, in Granger-causal inferencing the idea is to use time series data analysis and hypothesis testing to extract possible causal influences. However, much more informative models are structural equation models (see e.g. [163]) and causal Bayesian networks (see e.g. [103]).

Structural equation models allow specification of equations that denote the effect of one variable on the other. That is most helpful in doing interventional analysis. The model also supports a logic with an algebra that allows counterfactual analysis. On the other hand, Bayesian networks allow for probabilistic relations between variables and thus also enable interventional and counterfactual analysis. The Structural Causal Model is possibly the most evolved and combines the benefits of both these models. The rich set of analytical tools that it comes with is described in [163].

Generally, Pearl argues in [163] that explainability and other obstacles "can be overcome using causal modeling tools, in particular, causal diagrams and their associated logic." Methods for creating such causal models are therefore of great importance. The main challenge of these methods in practice, however, is learning the model from data. Very often expert input is used in conjunction with data to build the models. Heckerman describes in [102] how it is possible to build Causal Bayesian Networks from data, under some assumptions. Causal Bayesian Networks can be extended to Bayesian Decision Networks where the decision variables and the utility (optimization) variables are explicitly identified and used for decision making. See e.g. [53] for a concise introduction to the area.

Overall, we contend that despite the progress with techniques for causality, we are still far from formalizing and being able to explain other more nuanced interpretations of causality that humans are familiar with. This is exemplified by the works on 'actual' causality, e.g. [63], which illustrates with simple examples the challenges or explainability using simple causal diagrams.

4 Remarks

We briefly discuss a few aspects related to our categorization of explanations (Section 2.3) and its relationship to MR approaches overviewed in this report.

Causality We have discussed Causal Approaches loosely as a branch of MR where explainability is investigated. Instead, *causality* can be viewed as a property or even constitute a category of explanations, see e.g. [110]. However, following e.g. [144, 189, 207], we contend that explanations may, but need not in general be causal. It is for instance acknowledged that "it does not seem possible to develop a formal connection between counterfactuals and causality."[90, p. 69] We thus maintain that the aspect of causality is orthogonal to our categorization of explanations.

Hierarchy of explanation families We see the three categories of explanations forming a hierarchy of increasing complexity in the following intuitive sense. Attributive explanations may act as components of contrastive explanations (e.g. attributions pro and con), and contrastive explanations can pave way for actionable explanations (e.g. contrasting outcomes lead to actions). This complexity also well reflects the maturity of different types of explanations, with attributive ones being the oldest and most pervasive, contrastive ones more recent and advanced, and actionable explanations arguably the most challenging and least explored.

User-centrism In addition to the three families of explanations that we proposed, we recognize the property of explanations being *user-centric*, see e.g. [40, 149, 150, 176, 207, 214]. While it clearly applies to attributive, contrastive and actionable explanations, here we also have in mind a slightly different, more general notion of user-centrism. Specifically, it pertains to approaches to explainability directly taking into account the user (human or AI, indifferently) model, preferences, intentions etc. It is not only about the explanations being e.g. accessible, comprehensible, informative, but also about the relation between the explainee and the explanation. For instance, in the human user case,

1. an attributive explanation may need to take into account the relevance of the attributions to the user, e.g. domain vs procedural knowledge in the inference from knowledge to outcome;
2. a contrastive explanation may need to take into account the context in terms of which counterfactual situations are attainable to the user, e.g. the controllable aspects of the problem domain such as resource availability;
3. an actionable explanation may need to take into account the user's preferences over decisions, e.g. costs of resources and actions.

User-centric explanations concern the system's understanding of its user, and include aspects of cooperation and adaptation. At a high level, they aim to answer the following question:

- Given a representation of the problem and of the user, e.g. their preferences or mental model, and given an object of interest (e.g. a query, a decision, a solution, a change, a desideratum), how are the user and the object related?

The challenges with attaining such explanations (and hence the current scarcity of them, at least in MR) have been brought forward in e.g. [44, 121, 194]. We hope that the developing landscape of MR explainability may soon allow (and require) an overview of the user-centric aspects of explanations.

Explanation desiderata In addition to such overarching properties as user-centrism and the categories of explanations suggested herein, explanations in AI can be studied and classified in terms of the *desiderata* (or desirable properties) they fulfil. To appreciate the evolution of desiderata for AI explainability over the previous decades we invite the reader to consult the following works: [96, 100, 110, 123, 128, 137, 138, 189, 204]. We agree that a discussion of the various desiderata would be complementary to our overview, but it is beyond the scope of this work. We believe that our categorization works at a sufficiently high-level and is adequate for the purposes of this report.

Applicability of categories We speculate that our loose categorization of explanations in AI applies to ML as well as to MR. For instance, post-hoc ML explanation techniques summarized in [13, p. 89, Figure 4], are attributive, except for local explanations, which can also be seen as contrastive. It is reasonable to expect our categories to be applicable to other XAI approaches, given that our work builds on and borrows from recent XAI overviews. However, we by no means claim that our loose categorization is exhaustive: there may obviously be types of explanations that are not covered by our three families, e.g. explanations by example that are rather popular in ML (see e.g. [149]). We leave it for future work to see how broadly our categorization applies to non-MR approaches.

5 Conclusions

In this report, we have provided a high-level conceptual overview of selected Machine Reasoning (MR) approaches to explainability. We have summarised what we believe are the most relevant MR contributions to Explainable AI (XAI), from early to modern MR research, perhaps with a stronger focus on the more recent studies. We have discussed explainability in MR branches of Classical Logic-Based Reasoning, Non-classical Logic-Based Reasoning (Logic Programming), Argumentation, Decision Theory and Planning as well as the related areas of Multi-Agent Systems, Reinforcement Learning and Causal Approaches. In particular, we have seen that MR explainability approaches are suited not only for explainable MR (i.e. explaining MR-based AI systems) but also for explainability in other fields or areas of AI, such Machine Learning (ML).

We have loosely categorized the various kinds of explanations provided by MR approaches into three families of explanations: attributive, contrastive and actionable. Attributive explanations give details as to why an AI system yields a particular output given a particular input in terms of attribution and association of the (parts of the) system and the input with the output. These type of explanations have been studied since the very early MR and continue to be relevant and widely used in modern MR as well as its approaches to XAI at large. Contrastive explanations give details as to why an AI system yields one but not another output given some input in terms of reasons for and against different outputs. This type of explanations has been advocated for in MR for a long time too and appears in

modern MR in the form of counterexamples, criticisms, counterfactuals and dialogues. Such and similar forms of contrastive explanations are being actively explored in MR and its applications to XAI. Finally, actionable explanations give details as to what can be done in order for an AI system to yield a particular output given input in terms of actions available to the system's user (human or AI, indifferently). This kind of explanations should enable interventions to the system and eventually an interactive collaboration between the user and the system so as to reach desirable outcomes. We see actionable explanations as belonging to the frontier of MR explainability where novel research approaches and directions are being proposed.

Our categorization of explanations is informed by the different types of questions about the AI system's workings which explanations seek to answer. We have indicated some of the questions addressed in the overviewed MR branches and summarized the higher-level questions. In answering the latter, an explanation provides some details and reasons about the AI system and its functioning. This pertains to what we believe are the main purposes of explanations in XAI, namely to enable the user of an AI system to both understand the system and to do something with the explanation.

The main lesson in writing this report was perhaps the (re)discovery of the evolution of XAI challenges and the wealth of MR approaches aiming to address those challenges. Importantly, we want to stress that XAI research in MR is very much active to date, as seen from our overview of modern MR explainability studies. Still, despite the advances over the years, challenges in MR explainability abound and it seems that lessons from the past hold well today too [110, p. 161]:

> If explanation provision is to become a characteristic feature of many future interfaces, then there is a special responsibility for researchers in both HCI [human-computer interaction] and AI to provide input to the debate about the nature of the explanations to be provided in future information systems. The onus on us as researchers in the area is to ensure that we profit by past research on explanation provision, identify the strengths and weaknesses in present research and build on the strengths and address the problems in the future.

We hope that this report will inform the XAI community about the progress in MR explainability and its overlap with other areas of AI, thus contributing to the bigger picture of XAI.

References

[1] Amina Adadi and Mohammed Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018. ISSN 2169-3536. doi: 10.1109/ACCESS.2018.2870052. URL https://ieeexplore.ieee.org/document/8466590/.

[2] AK Agogino, Ritchie Lee, and Dimitra Giannakopoulou. Challenges of explaining real-time planning. In *ICAPS Workshop on Explainable Planning (XAIP)*, 2019.

[3] Emanuele Albini, Antonio Rago, Pietro Baroni, and Francesca Toni. Relation-Based Counterfactual Explanations for Bayesian Network Classifiers. In Christian Bessiere, editor, *29th International Joint Conference on Artificial Intelligence*, pages 451–457, Yokohama, 2020. IJCAI. doi: 10.24963/ijcai.2020/63.

[4] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe Reinforcement Learning via Shielding. In Sheila A McIlraith and Kilian Q Weinberger, editors, *32nd AAAI Conference on Artificial Intelligence*, pages 2669–2678, New Orleans, Louisiana, 2018. AAAI Press. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17211.

[5] Leila Amgoud and Henri Prade. Using Arguments for Making and Explaining Decisions. *Artificial Intelligence*, (3-4):413–436, 2009. ISSN 00043702. doi: 10.1016/j.artint.2008.11.006.

[6] Leila Amgoud and Mathieu Serrurier. Agents that Argue and Explain Classifications. *Autonomous Agents and Multi-Agent Systems*, 16(2):187–209, 2008. ISSN 13872532. doi: 10.1007/s10458-007-9025-6.

[7] Jérôme Amilhastre, Hélène Fargier, and Pierre Marquis. Consistency Restoration and Explanations in Dynamic CSPs - Application to Configuration. *Artificial Intelligence*, 135(1-2):199–234, 2002. ISSN 00043702. doi: 10.1016/S0004-3702(01)00162-X.

[8] Ofra Amir, Finale Doshi-Velez, and David Sarne. Summarizing Agent Strategies. *Autonomous Agents and Multi-Agent Systems*, 33(5):628–644, sep 2019. ISSN 1387-2532. doi: 10.1007/s10458-019-09418-w. URL https://doi.org/10.1007/s10458-019-09418-whttp://link.springer.com/10.1007/s10458-019-09418-w.

[9] Robert Andrews, Joachim Diederich, and Alan B Tickle. Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks. *Knowledge-Based Systems*, 8(6):373–389, dec 1995. ISSN 09507051. doi: 10.1016/0950-7051(96)81920-4.

[10] Sule Anjomshoae, Davide Calvaresi, Amro Najjar, and Kary Främling. Explainable Agents and Robots: Results from a Systematic Literature Review. In Noa Agmon, Edith Elkind, Matthew E. Taylor, and Manuela Veloso, editors, *18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1078–1088, Montreal, 2019. IFAAMAS.

[11] Krzysztof R Apt and Roland N Bol. Logic Programming and Negation: A Survey. *The Journal of Logic Programming*, 19/20:9–71, 1994. ISSN 07431066. doi: 10.1016/0743-1066(94)90024-8.

[12] Abdallah Arioua, Nouredine Tamani, and Madalina Croitoru. Query Answering Explanation in Inconsistent Datalog +/- Knowledge Bases. In Qiming Chen, Abdelkader Hameurlain, Farouk Toumani, Roland Wagner, and Hendrik Decker, editors, *Database and Expert Systems Applications - 26th International Conference*, volume 9261 of *Lecture Notes in Computer Science*, pages 203–219, Valencia, 2015.

Springer. doi: 10.1007/978-3-319-22849-5_15. URL http://link.springer.com/10.1007/978-3-319-22849-5{_}15.

[13] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. *Information Fusion*, 58(December 2019):82–115, 2020. ISSN 15662535. doi: 10.1016/j.inffus.2019.12.012.

[14] K. Belahcene, C. Labreuche, N. Maudet, V. Mousseau, and W. Ouerdane. Explaining Robust Additive Utility Models by Sequences of Preference Swaps. *Theory and Decision*, 82(2):151–183, feb 2017. ISSN 0040-5833. doi: 10.1007/s11238-016-9560-1. URL http://link.springer.com/10.1007/s11238-016-9560-1.

[15] Leopoldo Bertossi. An ASP-Based Approach to Counterfactual Explanations for Classification. In *4th International Joint Conference on Rules and Reasoning*, 2020.

[16] Yves Bertot and Pierre Castéran. *Interactive Theorem Proving and Program Development - Coq'Art: The Calculus of Inductive Constructions*. Texts in Theoretical Computer Science. An EATCS Series. Springer, 2004. ISBN 978-3-642-05880-6. doi: 10.1007/978-3-662-07964-5. URL https://doi.org/10.1007/978-3-662-07964-5.

[17] Floris Bex. An Integrated Theory of Causal Stories and Evidential Arguments. *15th International Conference on Artificial Intelligence and Law*, pages 13–22, 2015. doi: 10.1145/2746090.2746094.

[18] Floris Bex and Douglas Walton. Combining Explanation and Argumentation in Dialogue. *Argument & Computation*, 7(1):55–68, 2016. doi: 10.3233/AAC-160001.

[19] Meghyn Bienvenu, Camille Bourgaux, and François Goasdoué. Computing and Explaining Query Answers over Inconsistent DL-Lite Knowledge Bases. *Journal of Artificial Intelligence Research*, 64:563–644, mar 2019. ISSN 1076-9757. doi: 10.1613/jair.1.11395. URL https://jair.org/index.php/jair/article/view/11395.

[20] Or Biran and Courtenay Cotton. Explanation and Justification in Machine Learning: A Survey. In *1st Workshop on Explainable Artificial Intelligence*, pages 8–13, 2017.

[21] Bart Bogaerts, Emilio Gamba, Jens Claes, and Tias Guns. Step-Wise Explanations of Constraint Satisfaction Problems. In *24th European Conference on Artificial Intelligence*, 2020. URL https://freuder.wordpress.com/pthg-19-the-third-workshop-on-progress-towards-the-holy-grail/.

[22] Arthur Boixel and Ulle Endriss. Automated Justification of Collective Decisions via Constraint Solving. In Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An, and Neil Yorke-Smith, editors, *19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 168–176, Auckland, 2020. IFAAMAS. doi: abs/10.5555/3398761.3398786.

[23] Andrei Bondarenko, Phan Minh Dung, Robert Kowalski, and Francesca Toni. An Abstract, Argumentation-Theoretic Approach to Default Reasoning. *Artificial Intelligence*, 93(97):63–101, 1997. doi: 10.1016/S0004-3702(97)00015-5.

[24] Richard Booth, Dov M Gabbay, Souhila Kaci, Tjitze Rienstra, and Leendert van der Torre. Abduction and Dialogical Proof in Argumentation and Logic Programming. In Torsten Schaub, Gerhard Friedrich, and Barry O'Sullivan, editors, *21st European Conference on Artificial Intelligence*, volume 263 of *Frontiers in Artificial Intelligence and Applications*, pages 117–122, Prague, 2014. IOS Press. ISBN 9781614994190. doi: 10.3233/978-1-61499-419-0-117.

[25] Léon Bottou. From Machine Learning to Machine Reasoning: An Essay. *Machine Learning*, 94(2): 133–149, 2014. ISSN 08856125. doi: 10.1007/s10994-013-5335-x.

[26] Maxime Bouton, Jesper Karlsson, Alireza Nakhaei, Kikuo Fujimura, Mykel J Kochenderfer, and Jana Tumova. Reinforcement Learning with Probabilistic Guarantees for Autonomous Driving. *CoRR*, abs/1904.0, 2019. URL http://arxiv.org/abs/1904.07189.

[27] Ronald J Brachman and Hector J Levesque. *Knowledge Representation and Reasoning*. Elsevier, 2004. ISBN 978-1-55860-932-7.

[28] Martin Brain and Marina De Vos. Answer Set Programming: A Domain in Need of Explanation - A Position Paper. In Thomas Roth-Berghofer, Stefan Schulz, David B Leake, and Daniel Bahls, editors, *3rd International and ECAI-08 Workshop on Explanation-Aware Computing*, pages 37–48, Patras, 2008.

[29] Gerhard Brewka. Preferred Subtheories: An Extended Logical Framework for Default Reasoning. In N S Sridharan, editor, *11th International Joint Conference on Artificial Intelligence*, pages 1043–1048, Detroit, 1989. Morgan Kaufmann.

[30] Gerhard Brewka, Thomas Eiter, and Mirosław Truszczyński. Answer Set Programming at a Glance. *Communications of the ACM*, 54(12):92–103, dec 2011. ISSN 0001-0782. doi: 10.1145/2043174.2043195. URL https://dl.acm.org/doi/10.1145/2043174.2043195.

[31] Cristian E. Briguez, Maximiliano C.D. Budán, Cristhian A.D. Deagustini, Ana G. Maguitman, Marcela Capobianco, and Guillermo R. Simari. Argument-based Mixed Recommenders and their Application to Movie Suggestion. *Expert Systems with Applications*, 41(14):6467–6482, oct 2014. ISSN 09574174. doi: 10.1016/j.eswa.2014.03.046. URL http://dx.doi.org/10.1016/j.eswa.2014.03.046https://linkinghub.elsevier.com/retrieve/pii/S0957417414001845.

[32] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In Prabhakaran Balakrishnan, Jaideep Srivatsava, Wai-Tat Fu, Sanda M Harabagiu, and Fei Wang, editors, *International Conference on Healthcare Informatics*, pages 160–169, Dallas, 2015. IEEE. ISBN 9781467395489. doi: 10.1109/ICHI.2015.26.

[33] Ruth M. J. Byrne. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In Sarit Kraus, editor, *28th International Joint Conference on Artificial Intelligence*, pages 6276–6282, Macao, aug 2019. IJCAI. ISBN 978-0-9992411-4-1. doi: 10.24963/ijcai.2019/876. URL https://www.ijcai.org/proceedings/2019/876.

[34] Pedro Cabalar, Jorge Fandinno, and Michael Fink. Causal Graph Justifications of Logic Programs. *Theory and Practice of Logic Programming*, 14(4-5):603–618, jul 2014. ISSN 1471-0684. doi: 10.1017/S1471068414000234. URL https://www.cambridge.org/core/product/identifier/S1471068414000234/type/journal{_}article.

[35] Roberta Calegari, Giovanni Ciatto, Jason Dellaluce, and Andrea Omicini. Interpretable Narrative Explanation for ML Predictors with LP: A Case Study for XAI. In Federico Bergenti and Stefania Monica,

editors, *20th Workshop "From Objects to Agents"*, volume 2404, pages 105–112, Parma, 2019. CEUR-WS.org.

[36] Sandra Carberry. Techniques for Plan Recognition. *User Modeling and User-Adapted Interaction*, 11 (1-2):31–48, 2001. ISSN 09241868. doi: 10.1023/A:1011118925938.

[37] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8):832, jul 2019. ISSN 2079-9292. doi: 10.3390/electronics8080832. URL https://www.mdpi.com/2079-9292/8/8/832.

[38] Michael Cashmore, Anna Collins, Benjamin Krarup, Senka Krivic, Daniele Magazzeni, and David Smith. Towards Explainable AI Planning as a Service. In *2nd International Workshop on Explainable AI Planning*, Berkeley, CA, jul 2019. URL https://openreview.net/pdf?id=rkl-Nph79V.

[39] İsmail İlkan Ceylan, Thomas Lukasiewicz, Enrico Malizia, and Andrius Vaicenavičius. Explanations for Query Answers under Existential Rules. In Sarit Kraus, editor, *28th International Joint Conference on Artificial Intelligence*, pages 1639–1646, Macao, aug 2019. IJCAI. ISBN 978-0-9992411-4-1. doi: 10.24963/ijcai.2019/227. URL https://www.ijcai.org/proceedings/2019/227.

[40] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy. In Carles Sierra, editor, *26th International Joint Conference on Artificial Intelligence*, pages 156–163, Melbourne, aug 2017. IJCAI. ISBN 9780999241103. doi: 10.24963/ijcai.2017/23. URL https://www.ijcai.org/proceedings/2017/23.

[41] Tathagata Chakraborti, Kshitij P. Fadnis, Kartik Talamadupula, Mishal Dholakia, Biplav Srivastava, Jeffrey O. Kephart, and Rachel K.E. Bellamy. Planning and Visualization for a Smart Meeting Room Assistant. *AI Communications*, 32(1):91–99, 2019. ISSN 09217126. doi: 10.3233/AIC-180609.

[42] Tathagata Chakraborti, Anagha Kulkarni, Sarath Sreedharan, David E. Smith, and Subbarao Kambhampati. Explicability? Legibility? Predictability? Transparency? Privacy? Security? The Emerging Landscape of Interpretable Agent Behavior. In J. Benton Srivastava, Nir Lipovetzky, Eva Onaindia, David E. Smith, and Siddharth Srivastava, editors, *29th International Conference on Automated Planning and Scheduling*, pages 86–96. AAAI Press, 2019. URL https://aaai.org/ojs/index.php/ICAPS/article/view/3463.

[43] Tathagata Chakraborti, Sarath Sreedharan, Sachin Grover, and Subbarao Kambhampati. Plan Explanations as Model Reconciliation. An Empirical Study. In *14th ACM/IEEE International Conference on Human-Robot Interaction*, pages 258–266. IEEE, 2019. ISBN 9781538685556. URL http://arxiv.org/abs/1802.01013.

[44] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. The Emerging Landscape of Explainable Automated Planning & Decision Making. In Christian Bessiere, editor, *29th International Joint Conference on Artificial Intelligence*, pages 4803–4811, Yokohama, 2020. IJCAI. ISBN 978-0-9992411-6-5. doi: 10.24963/ijcai.2020/669. URL https://www.ijcai.org/proceedings/2020/669.

[45] Martin Chapman, Panagiotis Balatsoukas, Mark Ashworth, Vasa Curcin, Nadin Kökciyan, Kai Essers, Isabel Sassoon, Sanjay Modgil, Simon Parsons, and Elizabeth I Sklar. Computational Argumentation-based Clinical Decision Support Demonstration. In Noa Agmon, Edith Elkind, Matthew E. Taylor, and Manuela Veloso, editors, *18th International Conference on Autonomous Agents and MultiA-*

gent Systems, pages 2345–2347, Montreal, 2019. IFAAMAS. URL http://www.ifaamas.org/Proceedings/aamas2019/pdfs/p2345.pdf.

[46] Giovanni Ciatto, Roberta Calegari, Andrea Omicini, and Davide Calvaresi. Towards XMAS: eXplainability through Multi-Agent Systems. In Claudio Savaglio, Giancarlo Fortino, Giovanni Ciatto, and Andrea Omicini, editors, *1st Workshop on Artificial Intelligence and Internet of Things*, pages 40–53, Rende, 2019. CEUR-WS.org.

[47] Gabriele Ciravegna, Francesco Giannini, Marco Gori, Marco Maggini, and Stefano Melacci. Human-Driven FOL Explanations of Deep Learning. In Christian Bessiere, editor, *29th International Joint Conference on Artificial Intelligence*, pages 2234–2240. IJCAI, 2020. ISBN 978-0-9992411-6-5. doi: 10.24963/ijcai.2020/309. URL https://www.ijcai.org/proceedings/2020/309.

[48] Gabriele Ciravegna, Francesco Giannini, Stefano Melacci, Marco Maggini, and Marco Gori. A Constraint-Based Approach to Learning and Explanation. In *34th AAAI Conference on Artificial Intelligence*, volume 34, pages 3658–3665. AAAI Press, 2020. doi: 10.1609/aaai.v34i04.5774. URL https://aaai.org/ojs/index.php/AAAI/article/view/5774.

[49] Oana Cocarascu, Kristijonas Čyras, and Francesca Toni. Explanatory Predictions with Artificial Neural Networks and Argumentation. In Dawid W. Aha, Trevor Darrell, Patrick Doherty, and Daniele Magazzeni, editors, *2nd Workshop on Explainable Artificial Intelligence*, Stockholm, 2018.

[50] Oana Cocarascu, Antonio Rago, and Francesca Toni. Extracting Dialogical Explanations for Review Aggregations with Argumentative Dialogical Agents. In Edith Elkind, Manuela Veloso, Noa Agmon, and Matthew E. Taylor, editors, *18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1261–1269, Montreal, 2019. IFAAMAS. ISBN 9781510892002.

[51] Oana Cocarascu, Andria Stylianou, Kristijonaš Cyras, and Francesca Toni. Data-Empowered Argumentation for Dialectically Explainable Predictions. In *24th European Conference on Artificial Intelligence*, Santiago de Compostela, 2020. IOS Press.

[52] Anna Collins, Daniele Magazzeni, and Simon Parsons. Towards an Argumentation-Based Approach to Explainable Planning. In Tathagata Chakraborti, Dustin Dannenhauer, Joerg Hoffmann, and Daniele Magazzeni, editors, *2nd ICAPS Workshop on Explainable Planning*, Berkeley, CA, 2019.

[53] Anthony C. Constantinou and Norman Fenton. Things to Know about Bayesian Networks: Decisions Under Uncertainty, Part 2. *Significance*, 15(2):19–23, apr 2018. ISSN 17409705. doi: 10.1111/j.1740-9713.2018.01126.x. URL http://doi.wiley.com/10.1111/j.1740-9713.2018.01126.x.

[54] Mark G. Core, H. C. Lane, Michael Van Lent, Dave Gomboc, Steve Solomon, and Milton Rosenberg. Building Explainable Artificial Intelligence Systems. In *21st National Conference on Artificial Intelligence (AAAI)*, volume 2, pages 1766–1773, Boston, 2006. AAAI Press. ISBN 1577352815. URL http://www.aaai.org/Library/AAAI/2006/aaai06-293.php.

[55] Kristijonas Čyras, Ken Satoh, and Francesca Toni. Explanation for Case-Based Reasoning via Abstract Argumentation. In *6th International Conference on Computational Models of Argument*, pages 243–254, Potsdam, 2016. IOS Press.

[56] Kristijonas Čyras, Xiuyi Fan, Claudia Schulz, and Francesca Toni. Assumption-Based Argumentation: Disputes, Explanations, Preferences. In Pietro Baroni, Dov M Gabbay, Massimiliano Giacomin, and

Leendert van der Torre, editors, *Handbook Of Formal Argumentation*, volume 1, pages 365–408. College Publications, 2018. ISBN 9781848902756.

[57] Kristijonas Čyras, David Birch, Yike Guo, Francesca Toni, Rajvinder Dulay, Sally Turvey, Daniel Greenberg, and Tharindi Hapuarachchi. Explanations by Arbitrated Argumentative Dispute. *Expert Systems with Applications*, 127:141–156, 2019. ISSN 09574174. doi: 10.1016/j.eswa.2019.03.012. URL https://linkinghub.elsevier.com/retrieve/pii/S0957417419301654.

[58] Kristijonas Čyras, Dimitrios Letsios, Ruth Misener, and Francesca Toni. Argumentation for Explainable Scheduling. In *33rd AAAI Conference on Artificial Intelligence*, volume 33, pages 2752–2759, Honolulu, Hawaii, 2019. AAAI Press. doi: 10.1609/aaai.v33i01.33012752. URL http://www.aaai.org/ojs/index.php/AAAI/article/view/4126.

[59] Kristijonas Čyras, Amin Karamlou, Myles Lee, Dimitrios Letsios, Ruth Misener, and Francesca Toni. AI-assisted Schedule Explainer for Nurse Rostering. In Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An, and Neil Yorke-Smith, editors, *19th International Conference on Autonomous Agents and MultiAgent Systems - Demo Track*, pages 2101–2103, Auckland, 2020. IFAAMAS.

[60] A. Darwiche. Model-Based Diagnosis using Structured System Descriptions. *Journal of Artificial Intelligence Research*, 8:165–222, jun 1998. ISSN 1076-9757. doi: 10.1613/jair.462. URL https://jair.org/index.php/jair/article/view/10206.

[61] Ernest Davis. Logical Formalizations of Commonsense Reasoning: A Survey. *Journal of Artificial Intelligence Research*, 59:651–723, 2017. doi: 10.1613/jair.5339.

[62] Martin Davis and Hilary Putnam. A Computing Procedure for Quantification Theory. *Journal of the ACM*, 7(3):201–215, jul 1960. ISSN 0004-5411. doi: 10.1145/321033.321034. URL http://dl.acm.org/doi/10.1145/321033.321034.

[63] Marc Denecker, Bart Bogaerts, and Joost Vennekens. Explaining Actual Causation in Terms of Possible Causal Processes. In Francesco Calimeri, Nicola Leone, and Marco Manna, editors, *Logics in Artificial Intelligence - 16th European Conference*, volume 1, pages 214–230, Rende, 2019. Springer. ISBN 9783030195694. doi: 10.1007/978-3-030-19570-0_14. URL http://dx.doi.org/10.1007/978-3-030-19570-0{_}14http://link.springer.com/10.1007/978-3-030-19570-0{_}14.

[64] Carmine Dodaro, Philip Gasteiger, Kristian Reale, Francesco Ricca, and Konstantin Schekotihin. Debugging Non-ground ASP Programs: Technique and Graphical Tools. *Theory and Practice of Logic Programming*, 19(2):290–316, 2019. ISSN 1471-0684. doi: 10.1017/S1471068418000492. URL https://www.cambridge.org/core/product/identifier/S1471068418000492/type/journal{_}article.

[65] Nicholas Downing, Thibaut Feydy, and Peter J. Stuckey. Explaining alldifferent. In Mark Reynolds and Bruce H. Thomas, editors, *25th Australasian Computer Science Conference*, volume 122, pages 115–124, Melbourne, 2012. Australian Computer Society. ISBN 978-1-921770-03-6.

[66] Phan Minh Dung. On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-person Games. *Artificial Intelligence*, 77(2):321–357, 1995. doi: 10.1016/0004-3702(94)00041-X.

[67] Phan Minh Dung, Robert Kowalski, and Francesca Toni. Dialectic Proof Procedures for Assumption-Based, Admissible Argumentation. *Artificial Intelligence*, 170(2):114–159, 2006. doi: 10.1016/j.artint.2005.07.002.

[68] Rebecca Eifler, Michael Cashmore, Jörg Hoffmann, Daniele Magazzeni, and Marcel Steinmetz. Explaining the Space of Plans through Plan-Property Dependencies. In *2nd International Workshop on Explainable AI Planning*, Berkeley, CA, 2019.

[69] Rebecca Eifler, Michael Cashmore, Jörg Hoffmann, Daniele Magazzeni, and Marcel Steinmetz. A New Approach to Plan-Space Explanation: Analyzing Plan-Property Dependencies in Oversubscription Planning. In *34th AAAI Conference on Artificial Intelligence*, volume 34, pages 9818–9826, New York, NY, 2020. AAAI Press. doi: 10.1609/aaai.v34i06.6534. URL https://aaai.org/ojs/index.php/AAAI/article/view/6534.

[70] Rebecca Eifler, Marcel Steinmetz, Álvaro Torralba, and Jörg Hoffmann. Plan-Space Explanation via Plan-Property Dependencies: Faster Algorithms & More Powerful Properties. In Christian Bessiere, editor, *29th International Joint Conference on Artificial Intelligence*, pages 4091–4097, Yokohama, 2020. IJCAI. ISBN 978-0-9992411-6-5. doi: 10.24963/ijcai.2020/566. URL https://www.ijcai.org/proceedings/2020/566.

[71] Kave Eshghi and Robert Kowalski. Abduction Compared with Negation by Failure. In Giorgio Levi and Maurizio Martelli, editors, *Logic Programming, the 6th International Conference*, number 3, pages 234–254, Lisbon, 1989. MIT Press. ISBN 0-262-62065-0.

[72] Richard Evans and Edward Grefenstette. Learning Explanatory Rules from Noisy Data. *Journal of Artificial Intelligence Research*, 61:1–64, jan 2018. ISSN 1076-9757. doi: 10.1613/jair.5714. URL https://jair.org/index.php/jair/article/view/11172.

[73] Marcelo A. Falappa, Gabriele Kern-Isberner, and Guillermo R. Simari. Explanations, Belief Revision and Defeasible Reasoning. *Artificial Intelligence*, 141(1-2):1–28, 2002. ISSN 00043702. doi: 10.1016/S0004-3702(02)00258-8.

[74] Xiuyi Fan. On Generating Explainable Plans with Assumption-Based Argumentation. In Tim Miller, Nir Oren, Yuko Sakurai, Itsuki Noda, Bastin Tony Roy Savarimuthu, and Tran Cao Son, editors, *21th International Conference on Principles and Practice of Multi-Agent Systems*, pages 344–361, Cham, 2018. Springer. ISBN 978-3-030-03098-8. doi: 10.1007/978-3-030-03098-8_21.

[75] Xiuyi Fan and Francesca Toni. A General Framework for Sound Assumption-Based Argumentation Dialogues. *Artificial Intelligence*, 216:20–54, 2014. ISSN 00043702. doi: 10.1016/j.artint.2014.06.001.

[76] Xiuyi Fan and Francesca Toni. On Computing Explanations in Argumentation. In Blai Bonet and Sven Koenig, editors, *29th AAAI Conference on Artificial Intelligence*, pages 1496–1502, Austin, Texas, 2015. AAAI Press. ISBN 978-1-57735-698-1.

[77] Xiuyi Fan and Francesca Toni. On Computing Explanations for Non-Acceptable Arguments. In Elizabeth Black, Sanjay Modgil, and Nir Oren, editors, *Theory and Applications of Formal Argumentation - 3rd International Workshop*, volume 9524 of *Lecture Notes in Computer Science*, pages 112–127, Buenos Aires, 2015. Springer. doi: 10.1007/978-3-319-28460-6_7.

[78] Jorge Fandinno and Claudia Schulz. Answering the "Why" in Answer Set Programming - A Survey of Explanation Approaches. *Theory and Practice of Logic Programming*, 19(2):114–203, 2019. ISSN 14753081. doi: 10.1017/S1471068418000534.

[79] Melvin Fitting. *First-Order Logic and Automated Theorem Proving, Second Edition*. Graduate Texts in Computer Science. Springer, 2 edition, 1996. ISBN 978-1-4612-7515-2. doi: 10.1007/978-1-4612-2360-3. URL https://doi.org/10.1007/978-1-4612-2360-3.

[80] John Fox. Cognitive Systems at the Point of Care: The CREDO Program. *Journal of Biomedical Informatics*, 68:83–95, apr 2017. ISSN 15320464. doi: 10.1016/j.jbi.2017.02.008. URL http://dx.doi.org/10.1016/j.jbi.2017.02.008https://linkinghub.elsevier.com/retrieve/pii/S1532046417300333.

[81] Maria Fox, Derek Long, and Daniele Magazzeni. Explainable Planning. In Dawid W Aha, Trevor Darrell, Michael Pazzani, Darryn Reid, Claude Sammut, and Peter Stone, editors, *1st Workshop on Explainable Artificial Intelligence*, Melbourne, 2017. URL http://arxiv.org/abs/1709.10256.

[82] Eugene C. Freuder. Explaining Ourselves: Human-Aware Constraint Reasoning. In Satinder P. Singh and Shaul Markovitch, editors, *31st AAAI Conference on Artificial Intelligence*, pages 4858–4862, San Francisco, CA, 2017. AAAI Press.

[83] Eugene C. Freuder, Chavalit Likitvivatanavong, Manuela Moretti, Francesca Rossi, and Richard J. Wallace. Computing Explanations and Implications in Preference-Based Configurators. In Barry O'Sullivan, editor, *Recent Advances in Constraints, Joint ERCIM/CologNet International Workshop on Constraint Solving and Constraint Logic Programming*, volume 2627, pages 76–92, Cork, 2002. Springer. ISBN 9783540366072. doi: 10.1007/3-540-36607-5_6. URL http://link.springer.com/10.1007/3-540-36607-5{_}6.

[84] Yosuke Fukuchi, Masahiko Osawa, Hiroshi Yamakawa, and Michita Imai. Autonomous self-explanation of behavior for interactive reinforcement learning agents. In *Proceedings of the 5th International Conference on Human Agent Interaction*, pages 97–101, 2017.

[85] M. Ganesalingam and W. T. Gowers. A fully automatic theorem prover with human-style output. *Journal of Automated Reasoning*, 58(2):253–291, 2017.

[86] Alejandro Javier García and Guillermo Ricardo Simari. Defeasible Logic Programming: DeLP-servers, Contextual Queries, and Explanations for Answers. *Argument & Computation*, 5(1):63–88, 2014. doi: 10.1080/19462166.2013.869767.

[87] Martin Gebser, Jörg Pührer, Torsten Schaub, and Hans Tompits. A Meta-Programming Technique for Debugging Answer-Set Programs. In Dieter Fox and Carla P Gomes, editors, *23rd AAAI Conference on Artificial Intelligence*, pages 448–453, Chicago, Illinois, 2008. AAAI. URL http://www.aaai.org/Library/AAAI/2008/aaai08-071.php.

[88] Hector Geffner. Causal Theories for Nonmonotonic Reasoning. In Howard E Shrobe, Thomas G Dietterich, and William R Swartout, editors, *8th National Conference on Artificial Intelligence*, pages 524–530, Boston, 1990. AAAI Press/MIT Press.

[89] Hector Geffner. Model-free, Model-based, and General Intelligence. In Jérôme Lang, editor, *27th International Joint Conference on Artificial Intelligence*, pages 10–17, Stockholm, 2018. IJCAI. URL http://arxiv.org/abs/1806.02308.

[90] Matthew L. Ginsberg. Counterfactuals. *Artificial Intelligence*, 30(1):35–79, 1986. ISSN 00043702. doi: 10.1016/0004-3702(86)90067-6.

[91] Moritz Göbelbecker, Thomas Keller, Patrick Eyerich, Michael Brenner, and Bernhard Nebel. Coming Up With Good Excuses: What to do When no Plan Can be Found. In Ronen I Brafman, Hector Geffner, Jörg Hoffmann, and Henry A Kautz, editors, *20th International Conference on Automated Planning and Scheduling*, pages 81–88, Toronto, 2010. AAAI. URL http://www.aaai.org/ocs/index.php/ICAPS/ICAPS10/paper/view/1453.

[92] Shafi Goldwasser, Silvio Micali, and Charles Rackoff. The Knowledge Complexity of Interactive Proof-systems. In Oded Goldreich, editor, *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali*, pages 203–225. ACM, 2019. doi: 10.1145/3335741.3335750. URL https://doi.org/10.1145/3335741.3335750.

[93] M. Sinan Gönül, Dilek Önkal, and Michael Lawrence. The Effects of Structural Characteristics of Explanations on Use of a DSS. *Decision Support Systems*, 42(3):1481–1493, dec 2006. ISSN 01679236. doi: 10.1016/j.dss.2005.12.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0167923605001806.

[94] Benjamin Grosof, Michael Kifer, Paul Fodor, and Janine Bloomfield. Rulelog: Highly Expressive, Yet Scalable, Semantic Rules. Technical report, Kyndi, Coherent Knowledge, Stony Brook University, 2018.

[95] Benjamin N Grosof, Michael Kifer, and Paul Fodor. Rulelog: Highly Expressive Semantic Rules with Scalable Deep Reasoning. In Nick Bassiliades, Antonis Bikakis, Stefania Costantini, Enrico Franconi, Adrian Giurca, Roman Kontchakov, Theodore Patkos, Fariba Sadri, and William Van Woensel, editors, *RuleML+RR: International Joint Conference on Rules and Reasoning*, London, 2017. CEUR-WS.org. URL http://ceur-ws.org/Vol-1875/paper18.pdf.

[96] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5):1–42, aug 2019. ISSN 03600300. doi: 10.1145/3236009. URL http://dl.acm.org/citation.cfm?doid=3271482.3236009.

[97] Mark Hall, Daniel Harborne, Richard Tomsett, Vedran Galetic, Santiago Quintana-Amate, Alistair Nottle, and Alun Preece. A Systematic Method to Understand Requirements for Explainable AI (XAI) Systems. In Tim Miller, Rosina Weber, and Daniele Magazzeni, editors, *3rd Workshop on Explainable Artificial Intelligence*, pages 21–27, Macao, 2019.

[98] Ned Hall. Two Concepts of Causation. In John Collins, Ned Hall, and Laurie Paul, editors, *Causation and Counterfactuals*, pages 225–276. MIT Press, 2004.

[99] Ned Hall. Structural Equations and Causation. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 132(1):109–136, 2007. ISSN 00318116, 15730883. URL http://www.jstor.org/stable/25471849.

[100] Lars Kai Hansen and Laura Rieger. Interpretability in Intelligent Systems – A New Concept? In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Lecture Notes in Computer Science*, volume 11700, pages 41–49. Springer, 2019. ISBN 9783030289539. doi: 10.1007/978-3-030-28954-6_3. URL http://link.springer.com/10.1007/978-3-030-28954-6{_}3.

[101] Abdelraouf Hecham, Abdallah Arioua, Gem Stapleton, and Madalina Croitoru. An Empirical Evaluation of Argumentation in Explaining Inconsistency-Tolerant Query Answering. In Alessandro Artale, Birte Glimm, and Roman Kontchakov, editors, *30th International Workshop on Description Logics*, Montpellier, 2017. CEUR-WS.org.

[102] David Heckerman. A Bayesian Approach to Learning Causal Networks. In Philippe Besnard and Steve Hanks, editors, *11th Annual Conference on Uncertainty in Artificial Intelligence*, pages 285–295, Montreal, 1995. Morgan Kaufmann. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1{&}smnu=2{&}article{_}id=444{&}proceeding{_}id=11.

[103] Christopher Hitchcock. Causal Models. In Edward N Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 202 edition, 2018. URL https://plato.stanford.edu/archives/sum2020/entries/causal-models/.

[104] Matthew Horridge, Bijan Parsia, and Ulrike Sattler. Explaining Inconsistencies in OWL Ontologies. In Lluís Godo and Andrea Pugliese, editors, *Scalable Uncertainty Management - 3rd International Conference*, volume 5785 LNAI, pages 124–137, Washington, 2009. Springer. ISBN 3642043879. doi: 10.1007/978-3-642-04388-8_11. URL http://link.springer.com/10.1007/978-3-642-04388-8{_}11.

[105] Alexey Ignatiev. Towards Trustable Explainable AI. In Christian Bessiere, editor, *29th International Joint Conference on Artificial Intelligence*, pages 5154–5158, Yokohama, 2020. IJCAI. ISBN 978-0-9992411-6-5. doi: 10.24963/ijcai.2020/726. URL https://www.ijcai.org/proceedings/2020/726.

[106] Alexey Ignatiev, Nina Narodytska, and João Marques-Silva. Abduction-Based Explanations for Machine Learning Models. In *33rd AAAI Conference on Artificial Intelligence*, volume 33, pages 1511–1519, Honolulu, Hawaii, 2019. AAAI Press. doi: 10.1609/aaai.v33i01.33011511.

[107] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. On Relating Explanations and Adversarial Examples. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence D'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *33rd Annual Conference on Neural Information Processing Systems*, pages 15857–15867, Vancouver, 2019.

[108] David Isele, Alireza Nakhaei, and Kikuo Fujimura. Safe Reinforcement Learning on Autonomous Vehicles. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1–6, Madrid, 2018. IEEE. doi: 10.1109/IROS.2018.8593420. URL https://doi.org/10.1109/IROS.2018.8593420.

[109] Nils Jansen, Bettina Könighofer, Sebastian Junges, and Roderick Bloem. Shielded Decision-Making in MDPs. *CoRR*, abs/1807.0, 2018. URL http://arxiv.org/abs/1807.06096.

[110] Hilary Johnson and Peter Johnson. Explanation Facilities and Interactive Systems. In Wayne D. Gray, William E. Hefley, and Dianne Murray, editors, *1st International Workshop on Intelligent User Interfaces*, pages 159–166, Orlando, 1993. ACM. ISBN 0897915569. doi: 10.1145/169891.169951. URL http://portal.acm.org/citation.cfm?doid=169891.169951.

[111] Ulrich Junker. QUICKXPLAIN: Preferred Explanations and Relaxations for Over-Constrained Problems. In Deborah L. McGuinness and George Ferguson, editors, *19th National Conference on Artificial Intelligence*, volume 3, pages 167–172, San Jose, CA, 2004. AAAI Press/MIT Press.

[112] Zoe Juozapaitis, Anurag Koul, Alan Fern, Martin Erwig, and Finale Doshi-Velez. Explainable reinforcement learning via reward decomposition. In *IJCAI/ECAI Workshop on Explainable Artificial Intelligence*, 2019.

[113] Antonis C Kakas, Robert Kowalski, and Francesca Toni. Abductive Logic Programming. *Journal of Logic and Computation*, 2(6):719–770, 1992. ISSN 0955-792X. doi: 10.1093/logcom/2.6.719.

[114] Aditya Kalyanpur, Bijan Parsia, Matthew Horridge, and Evren Sirin. Finding All Justifications of OWL DL Entailments. In Karl Aberer, Key-Sun Choi, Natasha Fridman Noy, Dean Allemang, Kyung-Il Lee, Lyndon J. B. Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web*, volume 4825 LNCS, pages 267–280, Busan, 2007. Springer. ISBN 3540762973. doi: 10.1007/978-3-540-76298-0_20. URL http://link.springer.com/10.1007/978-3-540-76298-0{_}20.

[115] Amin Karamlou, Kristijonas Čyras, and Francesca Toni. Complexity Results and Algorithms for Bipolar Argumentation. In Noa Agmon, Edith Elkind, Matthew E. Taylor, and Manuela Veloso, editors, *18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1713–1721, Montreal, 2019. IFAAMAS. URL http://www.ifaamas.org/Proceedings/aamas2019/pdfs/p1713.pdfhttp://dl.acm.org/citation.cfm?id=3331902.

[116] Dmitry Kazhdan, Zohreh Shams, and Pietro Liò. Marleme: A multi-agent reinforcement learning model extraction library. *arXiv preprint arXiv:2004.07928*, 2020.

[117] Omar Zia Khan, Pascal Poupart, and James P Black. Minimal sufficient explanations for factored markov decision processes. In *ICAPS*. Citeseer, 2009.

[118] Jürg Kohlas, Dritan Berzati, and Rolf Haenni. Probabilistic Argumentation Systems and Abduction. *Annals of Mathematics and Artificial Intelligence*, 34(1-3):177–195, 2002. ISSN 10122443. doi: 10.1023/A:1014482025714.

[119] Robert Kowalski. Logic Programming. In Jörg H Siekmann, editor, *Handbook of the History of Logic. Volume 9: Computational Logic*, pages 523–569. Elsevier, 2014. doi: 10.1016/B978-0-444-51624-4.50012-5. URL https://linkinghub.elsevier.com/retrieve/pii/B9780444516244500125.

[120] Benjamin Krarup, Michael Cashmore, Daniele Magazzeni, and Tim Miller. Model-Based Contrastive Explanations for Explainable Planning. In Tathagata Chakraborti, Dustin Dannenhauer, Joerg Hoffmann, and Daniele Magazzeni, editors, *ICAPS 2019 Workshop on Explainable AI Planning (XAIP)*, pages 21–29, Berkeley, 2019. URL https://strathprints.strath.ac.uk/69957/.

[121] Sarit Kraus, Amos Azaria, Jelena Fiosina, Maike Greve, Noam Hazon, Lutz Kolbe, Tim-Benjamin Lembcke, Jörg P. Müller, Sören Schleibaum, and Mark Vollrath. AI for Explaining Decisions in Multi-Agent Environments. In *34th AAAI Conference on Artificial Intelligence*, pages 13534–13538, New York, NY, 2020. AAAI Press. doi: 10.1609/aaai.v34i09.7077.

[122] Robbert Krebbers, Xavier Leroy, and Freek Wiedijk. Formal C Semantics: CompCert and the C Standard. In Gerwin Klein and Ruben Gamboa, editors, *Interactive Theorem Proving - 5th International Conference*, volume 8558 of *Lecture Notes in Computer Science*, pages 543–548, Vienna, 2014. Springer. doi: 10.1007/978-3-319-08970-6_36. URL https://doi.org/10.1007/978-3-319-08970-6{_}36.

[123] Todd Kulesza, Margaret Burnett, Weng-keen Wong, and Simone Stumpf. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In Oliver Brdiczka, Polo Chau, Giuseppe Carenini, Shimei Pan, and Per Ola Kristensson, editors, *20th International Conference on Intelligent User Interfaces*, pages 126–137, Atlanta, GA, 2015. ACM. ISBN 9781450333061. doi: 10.1145/2678025.2701399. URL http://dl.acm.org/citation.cfm?doid=2678025.2701399.

[124] Anagha Kulkarni, Satya Gautam Vadlamudi, Yantian Zha, Yu Zhang, Tathagata Chakraborti, and Subbarao Kambhampati. Explicable Planning as Minimizing Distance from Expected Behavior. In *18th International Conference on Autonomous Agents and MultiAgent Systems*, volume 4, pages 2075–2077, Montreal, 2019. IFAAMAS. ISBN 9781510892002. URL https://dl.acm.org/doi/10.5555/3306127.3332015.

[125] Christophe Labreuche. A General Framework for Explaining the Results of a Multi-Attribute Preference Model. *Artificial Intelligence*, 175(7-8):1410–1448, may 2011. ISSN 00043702. doi: 10.1016/j.artint.2010.11.008. URL http://dx.doi.org/10.1016/j.artint.2010.11.008https://linkinghub.elsevier.com/retrieve/pii/S0004370210001979.

[126] Christophe Labreuche and Simon Fossier. Explaining Multi-Criteria Decision Aiding Models with an Extended Shapley Value. In Jérôme Lang, editor, *27th International Joint Conference on Artificial Intelligence*, pages 331–339, Stockholm, jul 2018. IJCAI. ISBN 9780999241127. doi: 10.24963/ijcai.2018/46. URL https://www.ijcai.org/proceedings/2018/46.

[127] Christophe Labreuche, Nicolas Maudet, and Wassila Ouerdane. Minimal and Complete Explanations for Critical Multi-attribute Decisions. In Ronen I. Brafman, Fred S. Roberts, and Alexis Tsoukiàs, editors, *Algorithmic Decision Theory - 2nd International Conference*, pages 121–134, Piscataway, NJ, 2011. Springer. ISBN 9783642248726. doi: 10.1007/978-3-642-24873-3_10. URL http://link.springer.com/10.1007/978-3-642-24873-3{_}10.

[128] Carmen Lacave and Francisco J. Díez. A Review of Explanation Methods for Bayesian Networks. *Knowledge Engineering Review*, 17(2):107–127, 2002. ISSN 02698889. doi: 10.1017/S026988890200019X.

[129] Luís C. Lamb, Artur D'Avila Garcez, Marco Gori, Marcelo O.R. Prates, Pedro H.C. Avelar, and Moshe Y. Vardi. Graph Neural Networks Meet Neural-Symbolic Computing: A Survey and Perspective. In Christian Bessiere, editor, *29th International Joint Conference on Artificial Intelligence*, pages 4877–4884. IJCAI, 2020. ISBN 978-0-9992411-6-5. doi: 10.24963/ijcai.2020/679. URL https://www.ijcai.org/proceedings/2020/679.

[130] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. Explainable Agency for Intelligent Autonomous Systems. In Satinder P. Singh and Shaul Markovitch, editors, *31st AAAI Conference on Artificial Intelligence*, pages 4762–4764, San Francisco, CA, 2017. AAAI Press. URL https://www.aaai.org/ocs/index.php/IAAI/IAAI17/paper/view/15046/13734.

[131] Martin Lauer and Martin A Riedmiller. An Algorithm for Distributed Reinforcement Learning in Cooperative Multi-Agent Systems. In Pat Langley, editor, *17th International Conference on Machine Learning*, pages 535–542, Stanford, 2000. Morgan Kaufmann.

[132] Mark Law, Alessandra Russo, and Krysia Broda. Logic-Based Learning of Answer Set Programs. In Markus Krötzsch and Daria Stepanova, editors, *Reasoning Web. Explainable Artificial Intelligence*, volume 11810, pages 196–231, Bolzano, 2019. Springer. doi: 10.1007/978-3-030-31423-1_6. URL http://dx.doi.org/10.1007/978-3-030-31423-1{_}4http://link.springer.com/10.1007/978-3-030-31423-1{_}6.

[133] David B. Leake. Abduction, Experience, and Goals: A Model of Everyday Abductive Explanation. *Journal of Experimental and Theoretical Artificial Intelligence*, 7(4):407–428, 1995. ISSN 13623079. doi: 10.1080/09528139508953820.

[134] Freddy Lécué. On the Role of Knowledge Graphs in Explainable AI. *Semantic Web*, 11(1):41–51, jan 2020. ISSN 22104968. doi: 10.3233/SW-190374. URL https://www.medra.org/servlet/aliasResolver?alias=iospress{&}doi=10.3233/SW-190374.

[135] Xiao Li, Yao Ma, and Calin Belta. A Policy Search Method For Temporal Logic Specified Reinforcement Learning Tasks. In *2018 Annual American Control Conference*, pages 240–245, Milwaukee, 2018. IEEE. doi: 10.23919/ACC.2018.8431181. URL https://doi.org/10.23919/ACC.2018.8431181.

[136] Vladimir Lifschitz. *Answer Set Programming*. Springer, 2019. ISBN 978-3-030-24657-0. doi: 10.1007/978-3-030-24658-7. URL http://link.springer.com/10.1007/978-3-030-24658-7.

[137] Brian Y. Lim and Anind K. Dey. Toolkit to Support Intelligibility in Context-Aware Applications. In *12th ACM International Conference on Ubiquitous Computing*, pages 13–22, New York, New York, 2010. ACM Press. ISBN 9781605588438. doi: 10.1145/1864349.1864353. URL http://portal.acm.org/citation.cfm?doid=1864349.1864353.

[138] Zachary C. Lipton. The Mythos of Model Interpretability. *Communications of the ACM*, 61(10):36–43, sep 2018. ISSN 0001-0782. doi: 10.1145/3233231. URL https://dl.acm.org/doi/10.1145/3233231.

[139] Thomas Lukasiewicz, Enrico Malizia, and Cristian Molinaro. Explanations for Inconsistency-Tolerant Query Answering under Existential Rules. In *34th AAAI Conference on Artificial Intelligence*, pages 2909–2916, New York, NY, 2020. AAAI Press. doi: 10.1609/aaai.v34i03.5682. URL https://aaai.org/ojs/index.php/AAAI/article/view/5682.

[140] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems*, pages 4765–4774. Curran Associates, 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[141] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. A Grounded Interaction Protocol for Explainable Artificial Intelligence. In Edith Elkind, Manuela Veloso, Noa Agmon, and Matthew E. Taylor, editors, *18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1033–1041, Montreal, 2019. IFAAMAS. URL http://dl.acm.org/citation.cfm?id=3331801.

[142] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable Reinforcement Learning through a Causal Lens. In *34th AAAI Conference on Artificial Intelligence*, pages 2493–2500, New York, NY, 2020. AAAI Press. doi: 10.1609/aaai.v34i03.5631. URL https://aaai.org/ojs/index.php/AAAI/article/view/5631.

[143] Peter McBurney and Simon Parsons. Games that Agents Play: A Formal Framework for Dialogues Between Autonomous Agents. *Journal of Logic, Language and Information*, 11(3):315–334, 2002. ISSN 15729583. doi: 10.1023/A:1015586128739.

[144] Tim Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267:1–38, 2019. ISSN 00043702. doi: 10.1016/j.artint.2018.07.007. URL https://doi.org/10.1016/j.artint.2018.07.007.

[145] Pasquale Minervini, Matko Bošnjak, Tim Rocktäschel, Sebastian Riedel, and Edward Grefenstette. Differentiable Reasoning on Large Knowledge Bases and Natural Language. In *34th AAAI Conference on Artificial Intelligence*, pages 5182–5190, New York, NY, 2020. AAAI Press. doi: 10.1609/

aaai.v34i04.5962. URL http://arxiv.org/abs/1912.10824https://aaai.org/ojs/index.php/AAAI/article/view/5962.

[146] Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining Explanations in AI. In *2019 Conference on Fairness, Accountability, and Transparency*, pages 279–288, Atlanta, GA, 2019. ACM. ISBN 9781450361255. doi: 10.1145/3287560.3287574.

[147] Sanjay Modgil. Reasoning About Preferences in Argumentation Frameworks. *Artificial Intelligence*, 173(9-10):901–934, jun 2009. ISSN 00043702. doi: 10.1016/j.artint.2009.02.001. URL http://dx.doi.org/10.1016/j.artint.2009.02.001https://linkinghub.elsevier.com/retrieve/pii/S0004370209000162.

[148] Sanjay Modgil and Martin Caminada. Proof Theories and Algorithms for Abstract Argumentation Frameworks. In Guillermo Ricardo Simari and Iyad Rahwan, editors, *Argumentation in Artificial Intelligence*, chapter 6, pages 105–129. Springer, 2009. doi: 10.1007/978-0-387-98197-0_6.

[149] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 2020. doi: 10.1145/3387166.

[150] Johanna D. Moore and William R. Swartout. Pointing: A Way Toward Explanation Dialogue. In Howard E. Shrobe, Thomas G. Dietterich, and William R. Swartout, editors, *8th National Conference on Artificial Intelligence*, pages 457–464, Boston, 1990. AAAI Press/MIT Press. URL http://www.aaai.org/Library/AAAI/1990/aaai90-069.php.

[151] Bernard Moulin, Hengameh Irandoust, Micheline Bélanger, and G Desbordes. Explanation and Argumentation Capabilities: Towards the Creation of More Persuasive Agents. *Artificial Intelligence Review*, 17(3):169–222, 2002. ISSN 02692821. doi: 10.1023/A:1015023512975.

[152] Shane T. Mueller, Robert R. Hoffman, William Clancey, Abigail Emrey, and Gary Klein. Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI. Technical report, DARPA, 2019. URL http://arxiv.org/abs/1902.01876.

[153] Stephen Muggleton and Luc de Raedt. Inductive Logic Programming: Theory and Methods. *The Journal of Logic Programming*, 19-20(2):629–679, may 1994. doi: 10.1016/0743-1066(94)90035-3. URL https://linkinghub.elsevier.com/retrieve/pii/0743106694900353.

[154] Pradeep K Murukannaiah, Nirav Ajmeri, Catholijn M Jonker, and Munindar P Singh. New Foundations of Ethical Multiagent Systems. In Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An, and Neil Yorke-Smith, editors, *19th International Conference on Autonomous Agents and MultiAgent Systems - Blue Sky Ideas Track*, pages 1706–1710, Auckland, 2020. IFAAMAS.

[155] Robert Neches, William R Swartout, and Johanna D Moore. Explainable (and Maintainable) Expert Systems. In Aravind K. Joshi, editor, *9th International Joint Conference on Artificial Intelligence*, pages 382–389, Los Angeles, 1985. Morgan Kaufmann. URL http://ijcai.org/Proceedings/85-1/Papers/072.pdf.

[156] Ulf H. Nielsen, Jean Philippe Pellet, and André Elissee. Explanation Trees for Causal Bayesian Networks. In David A. McAllester and Petri Myllymäki, editors, *24th Conference on Uncertainty in Artificial Intelligence*, pages 427–434, Helsinki, 2008. AUAI Press. ISBN 0974903949.

[157] Andreas Niskanen and Matti Järvisalo. Smallest Explanations and Diagnoses of Rejection in Abstract Argumentation. In *17th International Conference on Principles of Knowledge Representation and Reasoning*, Rhodes, 2020.

[158] Ingrid Nunes and Dietmar Jannach. A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems. *User Modeling and User-Adapted Interaction*, 27(3-5):393–444, dec 2017. ISSN 0924-1868. doi: 10.1007/s11257-017-9195-0. URL http://link.springer.com/10.1007/s11257-017-9195-0.

[159] Ingrid Nunes, Simon Miles, Michael Luck, Simone Barbosa, and Carlos Lucena. Pattern-based Explanation for Automated Decisions. In Torsten Schaub, Gerhard Friedrich, and Barry O'Sullivan, editors, *21st European Conference on Artificial Intelligence*, volume 263, pages 669–674. IOS Press, 2014. ISBN 9781614994183. doi: 10.3233/978-1-61499-419-0-669.

[160] Barry O'Callaghan, Barry O'Sullivan, and Eugene C. Freuder. Generating Corrective Explanations for Interactive Constraint Satisfaction. In Peter van Beek, editor, *Principles and Practice of Constraint Programming*, pages 445–459, Sitges, 2005. Springer. doi: 10.1007/11564751_34. URL http://link.springer.com/10.1007/11564751{_}34.

[161] Johannes Oetsch, Jörg Pührer, and Hans Tompits. Catching the Ouroboros: On Debugging Non-ground Answer-Set Programs. *Theory and Practice of Logic Programming*, 10(4-6):513–529, jul 2010. ISSN 1471-0684. doi: 10.1017/S1471068410000256. URL https://www.cambridge.org/core/product/identifier/S1471068410000256/type/journal{_}article.

[162] Barry O'Sullivan, Alexandre Papadopoulos, Boi Faltings, and Pearl Pu. Representative Explanations for Over-Constrained Problems. In *22nd AAAI Conference on Artificial Intelligence*, pages 323–328, Vancouver, 2007. AAAI Press. ISBN 1577353234.

[163] Judea Pearl. The Seven Tools of Causal Inference, with Reflections on Machine Learning. *Communications of the ACM*, 62(3):54–60, feb 2019. ISSN 00010782. doi: 10.1145/3241036. URL http://dl.acm.org/citation.cfm?doid=3314328.3241036.

[164] Rafael Peñaloza and Barış Sertkaya. Understanding the Complexity of Axiom Pinpointing in Lightweight Description Logics. *Artificial Intelligence*, 250:80–104, sep 2017. ISSN 00043702. doi: 10.1016/j.artint.2017.06.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0004370217300711.

[165] Ramón Pino-Pérez and Carlos Uzcátegui. Preferences and Explanations. *Artificial Intelligence*, 149(1): 1–30, 2003. ISSN 00043702. doi: 10.1016/S0004-3702(03)00042-0.

[166] Enrico Pontelli, Tran Cao Son, and Omar Elkhatib. Justifications for Logic Programs under Answer Set Semantics. *Theory and Practice of Logic Programming*, 9(1):1–56, jan 2009. ISSN 1471-0684. doi: 10.1017/S1471068408003633. URL https://www.cambridge.org/core/product/identifier/S1471068408003633/type/journal{_}article.

[167] David Poole. On the Comparison of Theories: Preferring the Most Specific Explanation. In Aravind K. Joshi, editor, *9th International Joint Conference on Artificial Intelligence*, pages 144–147, Los Angeles, 1985. Morgan Kaufmann.

[168] Henry Prakken and Giovanni Sartor. Modelling Reasoning with Precedents in a Formal Dialogue Game. *Artificial Intelligence and Law*, 6(2-4):231–287, 1998. ISSN 0924-8463. doi: 10.1023/A:1008278309945.

[169] Alun Preece. Asking 'Why' in AI: Explainability of Intelligent Systems – Perspectives and Challenges. *Intelligent Systems in Accounting, Finance and Management*, 25(2):63–72, 2018. ISSN 21600074. doi: 10.1002/isaf.1422.

[170] Alessandro Previti and João Marques-Silva. Partial MUS Enumeration. In Marie DesJardins and Michael L. Littman, editors, *27th AAAI Conference on Artificial Intelligence*, pages 818–825, Bellevue, WA, 2013. AAAI Press. ISBN 9781577356158.

[171] Christopher Pulte, Jean Pichon-Pharabod, Jeehoon Kang, Sung Hwan Lee, and Chung-Kil Hur. Promising-ARM/RISC-V: A Simpler and Faster Operational Concurrency Model. In Kathryn S McKinley and Kathleen Fisher, editors, *40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, pages 1–15, Phoenix, AZ, 2019. ACM. doi: 10.1145/3314221.3314624. URL https://doi.org/10.1145/3314221.3314624.

[172] Antonio Rago, Oana Cocarascu, and Francesca Toni. Argumentation-Based Recommendations: Fantastic Explanations and How to Find Them. In Jérôme Lang, editor, *27th International Joint Conference on Artificial Intelligence*, pages 1949–1955, Stockholm, 2018. IJCAI. doi: 10.24963/ijcai.2018/269.

[173] Alex Raymond, Hatice Gunes, and Amanda Prorok. Culture-Based Explainable Human-Agent Deconfliction. In Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An, and Neil Yorke-Smith, editors, *19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1107–1115, Auckland, 2020. IFAAMAS.

[174] Raymond Reiter. A Logic for Default Reasoning. *Artificial Intelligence*, 13(1-2):81–132, 1980. ISSN 00043702. doi: 10.1016/0004-3702(80)90014-4.

[175] Raymond Reiter. A Theory of Diagnosis from First Principles. *Artificial Intelligence*, 32(1):57–95, apr 1987. ISSN 00043702. doi: 10.1016/0004-3702(87)90062-2. URL https://linkinghub.elsevier.com/retrieve/pii/0004370287900622.

[176] Mireia Ribera and Agata Lapedriza. Can We Do Better Explanations? A Proposal of User-Centered Explainable AI. In *24th Annual Meeting of the Intelligent Interfaces Community Workshops*, volume 2327, Los Angeles, USA, 2019. ACM.

[177] Avi Rosenfeld and Ariella Richardson. Explainability in Human-Agent Systems. *Autonomous Agents and Multi-Agent Systems*, pages 1–33, may 2019. ISSN 1387-2532. doi: 10.1007/s10458-019-09408-y. URL https://doi.org/10.1007/s10458-019-09408-yhttp://link.springer.com/10.1007/s10458-019-09408-y.

[178] Francesca Rossi, Peter van Beek, and Toby Walsh, editors. *Handbook of Constraint Programming*, volume 2 of *Foundations of Artificial Intelligence*. Elsevier, 2006. ISBN 978-0-444-52726-4. URL http://www.sciencedirect.com/science/bookseries/15746526/2.

[179] Chiaki Sakama. Abduction in Argumentation Frameworks. *Journal of Applied Non-Classical Logics*, 28(2-3):218–239, 2018. ISSN 19585780. doi: 10.1080/11663081.2018.1487241.

[180] Claudia Schulz and Francesca Toni. ABA-Based Answer Set Justification. *Theory and Practice of Logic Programming*, 13(Online-Supplement):4–5, 2013.

[181] Claudia Schulz and Francesca Toni. Justifying Answer Sets Using Argumentation. *Theory and Practice of Logic Programming*, 16(1):59–110, 2016. doi: 10.1017/S1471068414000702.

[182] Naziha Sendi, Nadia Abchiche-Mimouni, and Farida Zehraoui. A new Transparent Ensemble Method based on Deep learning. *Procedia Computer Science*, 159:271–280, jan 2019. ISSN 1877-0509. doi: 10.1016/J.PROCS.2019.09.182. URL https://www.sciencedirect.com/science/article/pii/S1877050919313614.

[183] Dunja Šešelja and Christian Straßer. Abstract Argumentation and Explanation Applied to Scientific Debates. *Synthese*, 190(12):2195–2217, aug 2013. ISSN 0039-7857. doi: 10.1007/s11229-011-9964-y. URL http://link.springer.com/10.1007/s11229-011-9964-y.

[184] Kostyantyn M Shchekotykhin. Interactive Query-Based Debugging of ASP Programs. In Blai Bonet and Sven Koenig, editors, *29th AAAI Conference on Artificial Intelligence*, pages 1597–1603, Austin, Texas, 2015. AAAI. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9400.

[185] Andy Shih, Arthur Choi, and Adnan Darwiche. A Symbolic Approach to Explaining Bayesian Network Classifiers. In Jérôme Lang, editor, *27th International Joint Conference on Artificial Intelligence*, pages 5103–5111, Stockholm, 2018. IJCAI. doi: 10.24963/ijcai.2018/708.

[186] Elizabeth I. Sklar and Mohammad Q. Azhar. Explanation through Argumentation. In Michita Imai, Tim Norman, Elizabeth Sklar, and Takanori Komatsu, editors, *6th International Conference on Human-Agent Interaction*, pages 277–285, Southampton, dec 2018. ACM. ISBN 9781450359535. doi: 10.1145/3284432.3284470. URL https://dl.acm.org/doi/10.1145/3284432.3284470.

[187] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In Annette N. Markham, Julia Powles, Toby Walsh, and Anne L. Washington, editors, *AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, New York, NY, 2020. ACM. ISBN 9781450371100. doi: 10.1145/3375627.3375830. URL https://dl.acm.org/doi/10.1145/3375627.3375830.

[188] Kacper Sokol and Peter Flach. Conversational Explanations of Machine Learning Predictions Through Class-Contrastive Counterfactual Statements. In Jérôme Lang, editor, *27th International Joint Conference on Artificial Intelligence*, pages 5785–5786, Stockholm, 2018. IJCAI. ISBN 9780999241127. doi: 10.24963/ijcai.2018/836.

[189] Kacper Sokol and Peter Flach. Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. In Mireille Hildebrandt, Carlos Castillo, Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna, editors, *Conference on Fairness, Accountability, and Transparency*, pages 56–67, Barcelona, jan 2020. ACM. ISBN 9781450369367. doi: 10.1145/3351095.3372870. URL http://dl.acm.org/doi/10.1145/3351095.3372870.

[190] Kacper Sokol and Peter Flach. One Explanation Does Not Fit All. *KI - Künstliche Intelligenz*, 34(2):235–250, jun 2020. ISSN 0933-1875. doi: 10.1007/s13218-020-00637-y. URL https://doi.org/10.1007/s13218-020-00637-yhttp://link.springer.com/10.1007/s13218-020-00637-y.

[191] Frode Sørmo, Jörg Cassens, and Agnar Aamodt. Explanation in Case-Based Reasoning - Perspectives and Goals. *Artificial Intelligence Review*, 24(2):109–143, 2005. ISSN 0269-2821. doi: 10.1007/s10462-005-4607-7.

[192] Mohammed H. Sqalli and Eugene C. Freuder. Inference-Based Constraint Satisfaction Supports Explanation. In William J. Clancey and Daniel S. Weld, editors, *13th National Conference on Artificial Intelligence*, volume 1, pages 318–325, Portland, 1996. AAAI Press/MIT Press. URL http://www.aaai.org/Library/AAAI/1996/aaai96-048.php.

[193] Sarath Sreedharan, Siddharth Srivastava, David Smith, and Subbarao Kambhampati. Why Can't You Do That HAL? Explaining Unsolvability of Planning Tasks. In Sarit Kraus, editor, *28th International Joint Conference on Artificial Intelligence*, pages 1422–1430, Macao, aug 2019. IJCAI. ISBN 978-0-9992411-4-1. doi: 10.24963/ijcai.2019/197. URL https://www.ijcai.org/proceedings/2019/197.

[194] Ramya Srinivasan and Ajay Chander. Explanation Perspectives from the Cognitive Sciences - A Survey. In Christian Bessiere, editor, *29th International Joint Conference on Artificial Intelligence*, pages 4812–4818, Yokohama, 2020. IJCAI. ISBN 978-0-9992411-6-5. doi: 10.24963/ijcai.2020/670. URL https://www.ijcai.org/proceedings/2020/670.

[195] Geoff Sutcliffe and Christian B Suttner. Evaluating General Purpose Automated Theorem Proving Systems. *Artificial Intelligence*, 131(1-2):39–54, 2001. doi: 10.1016/S0004-3702(01)00113-8. URL https://doi.org/10.1016/S0004-3702(01)00113-8.

[196] William R Swartout, Cécile Paris, and Johanna D Moore. Explanations in Knowledge Systems: Design for Explainable Expert Systems. *IEEE Expert*, 6(3):58–64, jun 1991. ISSN 0885-9000. doi: 10.1109/64.87686. URL http://ieeexplore.ieee.org/document/87686/.

[197] Sjoerd T. Timmer, John-Jules Ch. Meyer, Henry Prakken, Silja Renooij, and Bart Verheij. A Two-Phase Method for Extracting Explanatory Arguments from Bayesian Networks. *International Journal of Approximate Reasoning*, 80:475–494, jan 2017. ISSN 0888613X. doi: 10.1016/j.ijar.2016.09.002.

[198] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. Interpretable to Whom? A Role-Based Model for Analyzing Interpretable Machine Learning Systems. In Dawid W Aha, Trevor Darrell, Patrick Doherty, and Daniele Magazzeni, editors, *2nd Workshop on Explainable Artificial Intelligence*, Stockholm, 2018.

[199] Francesca Toni. Automated Information Management via Abductive Logic Agents. *Telematics and Informatics*, 18(1):89–104, feb 2001. ISSN 07365853. doi: 10.1016/S0736-5853(00)00020-4. URL https://linkinghub.elsevier.com/retrieve/pii/S0736585300000204.

[200] Gianluca Torta, Luca Anselma, and Daniele Theseider Dupré. Exploiting Abstractions in Cost-sensitive Abductive Problem Solving with Observations and Actions. *AI Communications*, 27(3):245–262, 2014. ISSN 09217126. doi: 10.3233/AIC-140593.

[201] Gianluca Torta, Roberto Micalizio, and Samuele Sormano. Temporal Multiagent Plan Execution: Explaining What Happened. In Davide Calvaresi, Amro Najjar, Michael Schumacher, and Kary Främling, editors, *Explainable, Transparent Autonomous Agents and Multi-Agent Systems - 1st International Workshop*, pages 167–185. Springer, 2019. ISBN 978-3-030-30390-7. doi: 10.1007/978-3-030-30391-4. URL http://link.springer.com/10.1007/978-3-030-30391-4.

[202] Carlos Viegas Damásio, Anastasia Analyti, and Grigoris Antoniou. Justifications for Logic Programming. In Pedro Cabalar and Tran Cao Son, editors, *Logic Programming and Nonmonotonic Reasoning, 12th International Conference*, pages 530–542, Coruna, 2013. Springer. doi: 10.1007/978-3-642-40564-8_53. URL http://link.springer.com/10.1007/978-3-642-40564-8{_}53.

[203] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2): 841–887, 2018. ISSN 1556-5068. doi: 10.2139/ssrn.3063289. URL https://www.ssrn.com/abstract=3063289.

[204] Douglas Walton. A New Dialectical Theory of Explanation. *Philosophical Explorations*, 7(1): 71–89, mar 2004. ISSN 1386-9795. doi: 10.1080/1386979032000186863. URL http://www.tandfonline.com/doi/abs/10.1080/1386979032000186863.

[205] Douglas Walton. A Dialogue System Specification for Explanation. *Synthese*, 182(3):349–374, oct 2011. ISSN 0039-7857. doi: 10.1007/s11229-010-9745-z. URL http://link.springer.com/10.1007/s11229-010-9745-z.

[206] Douglas Walton. A Dialogue System for Evaluating Explanations. In *Argument Evaluation and Evidence*, volume 23 of *Law, Governance and Technology Series*, pages 69–116. Springer, 2016. doi: 10.1007/978-3-319-19626-8_3. URL http://link.springer.com/10.1007/978-3-319-19626-8{_}3.

[207] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. Designing Theory-Driven User-Centric Explainable AI. In Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos, editors, *Conference on Human Factors in Computing Systems*, Glasgow, 2019. ACM. ISBN 9781450359702. doi: 10.1145/3290605.3300831. URL http://dl.acm.org/citation.cfm?doid=3290605.3300831.

[208] Felix Winter, Peter J Stuckey, and Nysret Musliu. Explaining Propagators for String Edit Distance Constraints. In Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An, and Neil Yorke-Smith, editors, *19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1676–1683, Auckland, 2020. AAAI Press.

[209] Xiaoxin Yin and Jiawei Han. CPAR: Classification Based on Predictive Association Rules. In Daniel Barbará and Chandrika Kamath, editors, *3rd SIAM International Conference on Data Mining*, pages 331–335, San Francisco, CA, 2013. SIAM. doi: 10.1137/1.9781611972733.40.

[210] Zhiwei Zeng, Xiuyi Fan, Chunyan Miao, Cyril Leung, Chin Jing Jih, and Ong Yew Soon. Context-Based and Explainable Decision Making with Argumentation. In *17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1114–1122, Stockholm, 2018. IFAAMAS. URL http://dl.acm.org/citation.cfm?id=3237383.3237862.

[211] Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. *CoRR*, abs/1911.1, 2019. URL http://arxiv.org/abs/1911.10635.

[212] Qiaoting Zhong, Xiuyi Fan, Francesca Toni, and Xudong Luo. Explaining Best Decisions via Argumentation. In Andreas Herzig and Emiliano Lorini, editors, *European Conference on Social Intelligence*, volume 1283 of *CEUR Workshop Proceedings*, pages 224–237, Barcelona, 2014. CEUR-WS.org.

[213] Qiaoting Zhong, Xiuyi Fan, Xudong Luo, and Francesca Toni. An Explainable Multi-Attribute Decision Model Based on Argumentation. *Expert Systems with Applications*, 117:42–61, 2019. ISSN 09574174. doi: 10.1016/j.eswa.2018.09.038.

[214] Yishan Zhou and David Danks. Different "Intelligibility" for Different Folks. In Annette N. Markham, Julia Powles, Toby Walsh, and Anne L. Washington, editors, *AAAI/ACM Conference on AI, Ethics, and Society*, pages 194–199, New York, NY, feb 2020. ACM. ISBN 9781450371100. doi: 10.1145/3375627.3375810. URL https://dl.acm.org/doi/10.1145/3375627.3375810.