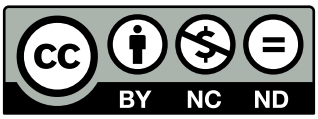
Imperial College London
Department of Computing

Non-monotonicity in Case-Based Reasoning and Explanations with Applications to Legal Reasoning

Guilherme Paulino Passos

November 2023

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Computing of Imperial College London
and the Diploma of Imperial College London



Statement of Originality

I declare that this thesis was composed by myself and that the work here presented is my own, except where otherwise stated, in which case it has been properly acknowledged.

Guilherme Paulino Passos

Abstract

In the goal of capturing patterns of human reasoning, the artificial intelligence (AI) community coined the idea of non-monotonic reasoning (NMR): reasoning that may retract conclusions given new information. This was proposed as more adequately modelling human-style reasoning, including legal reasoning. In this thesis, we propose non-monotonic reasoning analyses of subfields of AI not typically seen as such. The NMR view yields both new criteria of agreement to human-style reasoning for evaluation of models, and new models developed with such criteria in mind. In particular, we cover: i) classification models, focusing on case-based reasoning (CBR) systems, and; ii) interactive explanation models in explainable AI (XAI).

In CBR, we analyse abstract argumentation models for case-based reasoning ($AA\text{-}CBR$) as a reasoning system and prove that it does not satisfy cautious monotonicity, a property proposed in the NMR literature. We present a new alternative to $AA\text{-}CBR$ that is provably cautiously monotonic, and satisfies as well other NMR properties, such as cut and cumulativity. This alternative also results in a principled treatment of noise in "incoherent" case-bases. We present ways by which $AA\text{-}CBR$ models can be mined from data, and compare with this new cautious version of $AA\text{-}CBR$, using the COMPAS dataset of criminal recidivism. In XAI, we analyse explanations as objects subject to reasoning and present a formal model of an interactive scenario for explanation between user and system. We analyse explanations as committing to some model behaviour, suggesting a form of entailment, which, we argue, can be non-monotonic. We illustrate the approach with arbitrated dispute trees for $AA\text{-}CBR$. Thus, we argue that NMR brings considerations to a wider scope of problems in AI, including AI and law, at a time in which recent systems show impressive capacities but no guarantees on behaviour.

Acknowledgements

On November of 2018, it was hard to image what doing a PhD in Imperial College London would actually be like for the following years. While the future always holds surprises for everyone, it seems that the years from 2019 to 2023 held a larger number of them. As is the case with unexpected challenges, the support of those around us is a crucial element of how we react to them. Therefore, I would like to show my sincere appreciation for everyone that was part of the journey of this PhD, for the support of all of those was indeed crucial to me.

I will start thanking my supervisor, Prof. Francesca Toni, for all her guidance in research and in the academic world, for her consistent presence in the form of many fruitful discussions and exchanges of ideas, of showing by example how to communicate research (which I wish I had already learned a quarter of how she does it) as well for her patience and feedback of all sorts. I have learned much from her, ranging from the most interesting topics in argumentation and AI, to how to do research, to the silliest of (my) questions (such as why are there no floor drains in British floors).

I would like to also thank my examiners, Prof. Alessandra Russo and Prof. Katie Atkinson, for their thorough reading and very thoughtful feedback, which has certainly much improved this thesis. My deep gratitude to them for making the viva such a rewarding and, indeed, fun experience.

My experience in Imperial was in no small part shaped by my colleagues and friends in the Computational Logic and Argumentation (CLArg) group. I certainly have greatly improved as a researcher and as a person through their excitement and hard work in research, and their friendship. In (approximate) order of appearance, they are: Oana Cocarascu, Kristijonas Čyras, Amin Karamlou, Neema Kotonya, Piyawat Lertvittayakumjorn, Antonio Rago, Andria Stylianou, Emanuele Albini, Alex Gaskell, Fabrizio Russo, Kai Zhang, Francis Rhys Ward, Avinash Kori, Junqi (Jay) Jiang, Xiang Yin, Adam Dejl, Deniz Gorur, Nico Potyka, Hamed Ayoobi, Francesco Leofante, Gabriel Freedman, Adam Gould, and Anna Rapberger. I also extend my thanks to all my friends in the Department of Computing. Without them those years would not have been the same.

I also am greatly indebted to my family for their continued support and

love, particular to my mother, Lucia Regina Batista Paulino Passos, and to my father, Fabio Barboza Passos. Their presence, in many forms and in both past and present, has been a solid ground from which I could develop my own path in life. I also thank my sister, Juliana Paulino Passos, and my cousin, Amanda Regina Barboza Ribeiro, for their (in person and remote) companionship during this time.

I am also grateful to my friends in London. Living in a city or country is more than seeing touristic buildings daily, but who you meet and what you learn from them. My friends spread out across the world are no less deserving of acknowledgements. Many of them have been present not only at a time where even friendships in the same city became remote relationships, but from the start and to the end of this PhD journey. My thanks also include all the friendships made during research visits and conferences, definitely transforming my views of what - and how exciting and enjoyable - a work trip is.

I would like to thank the support provided by the CAPES Foundation (*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* - Brazil), through the Ph.D. Scholarship 88881.174481/2018-01. I would also like to thank both the National Institute of Informatics, in Japan, and Prof. Ken Satoh, for providing me with the opportunity to go to Tokyo to collaborate in AI and law. I sincerely hope that this work is valuable in the eyes of the public who funded it.

I thank again all those who read drafts and gave me feedback on this thesis or on papers which resulted in it. Among those, I am deeply grateful to Alexandre Augusto Abreu Almeida, Matheus de Elias Muller, Victor Luis Barroso Nascimento, Amanda Regina Barboza Ribeiro, Bárbara Barros Dumas, Thaís Pinheiro de Carvalho, Tiago Yparraguirre Viegas, Paula Gabriela de Oliveira Teixeira, Rafael Carneiro Fidalgo, not only for their feedback on text and ideas, but also for their invaluable friendship. I also thank anonymous reviewers for their feedback on drafts of papers that resulted in this thesis. Of course, any remaining errors are my solely my own.

"'Whoso draweth blood in the streets shall be severely punished.' This is a law that may serve at the same time as an example of every fault in point of extent of which a law is susceptible: want of original amplitude: want of proper discrimination: and want of residuary amplitude through improper discrimination. (...) For why must the act be confined to the streets in order to come within the censure of the law? Is it less mischievous if committed in the market-place, or in a church? (...)
This instance I chose not as an uncommon but as a simple one: for the difficulty (I might have said the impossibility) is not to find laws that are incomplete in this point, but to find such as are complete."

Jeremy Bentham

"Sed juris incertitudini mederi difficilius est."
[But to remedy the uncertainty of the law is harder.]

Gottfried Wilhelm Leibniz

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the field of Artificial Intelligence (AI), one of the main goals is building machines that improve humans' abilities to solve difficult problems and to complete hard tasks. An important aspect is behaving according to human goals and expectations; in some cases, even being involved in direct collaboration with humans. So far, in particular with the advances in data-driven methods, AI has produced groundbreaking results in many fields, such as recommender systems, machine translation, image classification, and more recently even image and text generation. However, deployment or the extent of impact is not as wide as could be imagined for possible fields of application, and one of the causes for this is the difficulty of creating systems that, in real world situations, perform well and behave according to human goals and expectations (Russell and Norvig, 2020).

A fruitful area for the application of AI is law. The legal domain presents many challenges involving the use of not only information and knowledge, but also cases, values, rules, and norms. On the one hand, law requires strong forms of reliability: it should be fair and safe, it should be not subject to manipulation, and it should be accurate, since errors can have high individual and social costs. On the other hand, the legal system cannot be so slow in acting that its decisions become untimely, useless, or such that delays effectively deny provision of justice. That is, the search for perfection in each individual decision should not be an obstacle for deciding. The usage of AI can push boundaries of this trade-off, motivating the field of research of **AI and law** since at least the 1980s (Bench-Capon et al., 2012; Bench-Capon, 2022, for

surveys). Some examples of tasks in AI and law are: finding and applying relevant legal arguments for a case presented to an attorney; drafting documents such as pleadings, appeals, and court decisions; and providing preliminary legal guidance to citizens or groups which have no means of (effectively) hiring an attorney (e.g. Branting, 2017; Greenleaf et al., 2018; Bench-Capon, 2022).

In order to reach the goals of AI in general, researchers have found useful to model forms of reasoning that are more similar to how humans do so. This has led to the development of **non-monotonic reasoning** (NMR) models: forms of reasoning which are defeasible, that may reject previous conclusions by receiving new information (e.g. Gabbay et al., 1994; Heyninck et al., 2023). One area in which reasoning is frequently seen as defeasible is law, motivating applications of NMR to legal reasoning (e.g. Schauer, 2012; Sartor, 2018). Many theoretical tools were developed for NMR, including formal semantics, algorithms, and formal properties. It is in this tradition that computational argumentation has been developed. Argumentation is a way of understanding the treatment of missing and conflicting information, or, in other words, of modelling defeasible reasoning (e.g. Dung, 1995; Baroni et al., 2018b).

In this thesis, we adopt the assumption from the NMR literature that some NMR properties are important in human reasoning and thus use NMR as the main methodology to discuss two subfields of AI, doing so with a focus on the impacts of such fields to the legal domain. One of those subfields is **case-based reasoning** (CBR). Case-based reasoning aims to capture the human style of reasoning where experience from concrete past occasions is directly used in a new situation. While it does not exclude wider world knowledge or general theories, CBR aims to capture situations in which this direct appeal to experience is a source of solutions (e.g. Richter and Weber, 2013). This style of reasoning has stronger ties to law and legal reasoning, as much of legal reasoning is done casuistically. This is especially the case in common law systems, a legal tradition originated in England, associated of the idea that the core of law is created via specific decisions in concrete cases, and via further refinement of the rules in later cases (Garner, 2009). Still, arguing from cases is not restricted to common law systems. The other major legal tradition, that of civil law systems, originated in continental Europe, is stereotypically associated with the interpretation of general rules in legislation, typically in the form of

a legal code (Schauer, 2009, §6.1). However, it has been long recognised that there is a gradually increasing importance of case law in civil law countries, in order to provide a uniform interpretation of statutes or of written constitutions (David and Brierley, 1988; Bell, 1997; Algero, 2004; Lewis, 2021).

The second subfield we discuss is **explainable AI** (XAI). This area is focused in producing explanations for AI systems, including black-box systems (such as neural networks), but also other interpretable methods (such as logic-based methods). As alignment with human goals and expectations or even direct collaboration are requirements for real AI systems, proper communication between human and machine becomes essential, particularly given the increasing complexity of tasks performed by machines. Better communication of reasons why particular decisions were taken or predictions were produced is crucial in increasing transparency, human trust, and better alignment between what the system does and the goals of its developers, among other social and legal values (e.g. Mueller et al., 2019). Those requirements are particularly sensitive in the legal domain, since in law decisions must be justified, transparent, and aligned with the legal normative system.

Thus, in this thesis we will propose models and formal analyses of CBR and XAI as NMR. We believe NMR offers useful tools for analysis that allow improvements on CBR and on XAI, and that this has particular importance in the legal domain, where transparency, trust, and expected behaviours are crucial for deployment. In order to give some concreteness to this discussion, we will use an illustrative allegory:

Example 1.1. Inspired by the hypothetical monarch Rex discussed in jurisprudence (Hart, 1994; Fuller, 1969), imagine a nation reigned by an absolute monarch, that we will call Queen Regina. As an absolute monarch, she is the ultimate authority and her decisions are unquestionable.

In particular, in this nation there is a Chancellor, who is in charge of some areas of the government, including the application of laws. This is so since Queen Regina is not to be bothered and only exceptional situations are decided by her. One of the Chancellor's duties is giving advice to other authorities in government on how to apply the law. Still, the Chancellor is always clear about how he has no power to create law: he is either following direct orders from Queen Regina, or acting on what he presumes to be her will.

It so happens that Queen Regina forbids the usage of hats in this nation, and hat-users are subject to imprisonment. The Mayor of a city in this nation is uncertain of the details of this rule, and, in advance of the local Spring Festival, decides to consult with the Chancellor, in order to properly instruct the city guards. In particular, two questions are raised:

1. Does this rule include the court jester?

2. Does this rule include the usage of special hats typically associated to the Spring Festival in the nation?

The Chancellor readily replies that, from his understanding of the orders and his personal knowledge of the Queen and her will, he believes that:

1. It does not include the court jester; **[Reply 1]**

2. It does include the Spring Festival. **[Reply 2]**

The festival comes and goes, and every situation is solved locally by the Mayor. The Queen receives no complaints nor makes any public - or private - declaration regarding hats. Months later, in winter, guards arrest a man for wearing a hat (which, it can be said, is not the Spring Festival one), and ultimately this man is taken to Queen Regina. It is then realised that the man is none other than the court jester. Providing no justification, but as expected by the Chancellor, Queen Regina decides to let the court jester go free, and no charges are pressed.

Another year comes and in the advance of the new Spring Festival, the Mayor once again discusses with the Chancellor the current legislation. Since last festival, the only case brought to trial was the one of the court jester in the winter. This time, the Chancellor asserts that, in his understanding:

3. It is allowed to wear the Spring Festival hat. **[Reply 3]**

The Chancellor justifies this opinion by saying that, given that Queen Regina had released the court jester in the previous winter, he believes it is not a problem to wear a hat during Spring Festival.

The Mayor is puzzled. As the Chancellor expected (Reply 1), Queen Regina had indeed decided that the court jester may wear hats. However, by having

this prediction confirmed, the Chancellor now states that the special Spring Festival hats are allowed (Reply 3). In other words, by confirming a previous tentative conclusion (Reply 1), a second tentative conclusion (Reply 2) is now retracted. In the literature of NMR, we would say the Chancellor has violated a property called **cautious monotonicity**, according to which adding previous conclusions to the reasoner's assumptions should not make the reasoner retract other conclusions.

The Chancellor could have stated, for instance, in Reply 2 to believe that wearing a Spring Festival hat was allowed. Alternatively, he could have suspended judgement: having the mission of giving advice, he could have stated this was a hard case and he did not know whether such hats were allowed. Yet, the Chancellor has not only changed his opinion due to the expected judgement by the Queen, but also, used such judgement as the reason for his new opinion (Reply 3), surprising the Mayor.

According to some researchers in NMR (e.g. Gabbay, 1984; Kraus et al., 1990; Makinson, 1994), the Mayor should indeed be puzzled, since the aforementioned cautious monotonicity is an important property in reasoning. This story may seem too disconnected to real situations, but it carries some important themes that will be discussed through this thesis. For instance, when Queen Regina is thought to be the entirety of a legal system, and the Chancellor a (perhaps automated) explainer of this legal system, we will be in the scenario of AI and law, where the Chancellor is a CBR system, reasoning with decided cases. Another way of seeing it is when Queen Regina is a black-box AI system, and the Chancellor is an explainer. If we agree with the Mayor, this behaviour from the Chancellor is undesirable, or even unacceptable. It is this type of problem that we will discuss in the thesis.

In the remainder of this chapter, we present context and briefly present the state of the art in NMR (Section 1.1.1), CBR (Section 1.1.2), and XAI (Section 1.1.3). We then discuss some current limitations and objectives of the thesis (Section 1.2), our contributions (Section 1.3), then thesis structure (Section 1.4), and finally our publications on which this work was based (Section 1.5).

1.1 State of the art

1.1.1 Non-monotonic reasoning

Reasoning is an activity of processing information, typically connecting premises to conclusions. Premises are information received or previously held, and conclusions are what is derived, inferred, yielded, generated from, or consequences of, the premises. Symbolically, this can be defined via a formal language, with a *consequence relation* as a main concept. In classical logic, this approach was originated with Tarski's (semantical) definition of logical consequence and of consequence operator (1930; 1936), as well as Gentzen's notion of derivability (1934). This Tarski-Gentzen consequence relation can be characterised by 3 properties: i) reflexivity – all premises are conclusions; ii) transitivity – conclusions of conclusions of the original premises are, themselves, also conclusions of the original premises; and iii) monotonicity – a superset of the premises has as consequences at least every consequence of the original premises (see e.g. Israel, 1993). This last property means that adding more information to premises can only enlarge the set of conclusions (possibly explosively, as inconsistencies in classical logic), but never reduce it.

AI researchers realised that much of human reasoning is performed non-monotonically and aimed to capture those forms of reasoning. In much of human communication and practical reasoning, conclusions are tentative and defeasible; reasoning is subject to review and change, depending on what is said or the context in which premises apply – in summary, on new information. Indeed, there is empirical evidence from psychology that at least in some environments humans do interpret information and reason non-monotonically (Byrne, 1989; Neves et al., 2004; Ragni et al., 2016; Marcos Cramer, 2021; on rule interpretation, Struchiner et al., 2020). While we do not perform empirical human experiments, we use this literature as support for the importance of NMR in human reasoning. The goal of modelling this type of commonsense reasoning resulted in the field of NMR. Having in mind particularly the monotonicity of classical logic as a limitation, many non-monotonic formalisms have been developed as alternatives to it in order to create expert systems. Some examples are circumscription, in which defeasible axioms are transformed into others that minimise their extensions; default logic, where

default rules represent defeasible assumptions and their conditions of applicability; and preference models, in which models are (partially) ordered and defeasible inference is based on minimal models; among others (overviewed by Brewka, 1991; Gabbay et al., 1994; Strasser and Antonelli, 2019). Logic programming (overviewed by Gabbay et al., 1998; O'Donnell, 1998) is closely tied to NMR, with negation as failure as a central example of a non-monotonic mechanism, being a mechanism where propositions are considered false unless they can be explicitly proved true (Clark, 1977, overviewed in, e.g., Shepherdson, 1998). NMR has also been studied in the formal linguistics literature, such as for modelling pragmatics (Rooij and Schulz, 2011).

This multitude of systems created a demand for unifying views, allowing comparison between them and an understanding of possible strengths and weaknesses of each. With this in mind, an influential line of analysis in terms of properties was developed (Makinson, 1994). Gabbay (1984) originally proposed some of them as a discussion of what is reasoning, or a possible logical system. Those properties will be presented in more detail in Section 2.1.1.

In another direction, attempts to capture many NMR systems (including negation as failure in logic programming) under a single formalism led to the developments of computational argumentation (Lin and Shoham, 1989; Kakas et al., 1992; Dung, 1995; A. C. Kakas and Toni, 1998; Prakken, 2018). This line of work is itself very diverse. Abstract argumentation (AA) (Dung, 1995) is a very general model, able to be instantiated (at least in terms of representation) as many different systems and frameworks. This is done via a set of arguments (of which internal structured is abstracted away, becoming atomic) and a binary relation of *attack* between them. Structured argumentation is an approach in which argumentation is defined in some sort of language, and argumentation occurs as a non-monotonic mechanism (Besnard et al., 2014). Examples of this are ASPIC$^+$ (Prakken, 2010; Modgil and Prakken, 2013) and assumption-based argumentation (ABA) (Bondarenko et al., 1993; Bondarenko et al., 1997; Toni, 2014). Both define arguments in terms of (chains of) deductions, but a main difference is that the former contains both defeasible and strict rules, while the latter represents all defeasibility at the level of assumptions. The structured argumentation approach has shown itself fruitful, in that argumentation provides not only a mechanism for non-monotonicity,

but also for inconsistency handling (paraconsistency). This can clearly be seen for instance in deductive argumentation (Besnard and Hunter, 2008; Arieli et al., 2021), where a system based in e.g. classical logic is enriched by an argumentation mechanism, and a knowledge base does not become explosive by having inconsistencies, but able to represent where the conflicts are and how to compute them. Generalisations of AA have also been developed and used for many other applications. Bipolar argumentation frameworks encode a notion of support, not only of attack (Cayrol and Lagasquie-Schiex, 2005; Amgoud et al., 2008), and abstract dialectical frameworks (ADF) generalise further by considering a general notion of acceptance condition functions[1] (Brewka and Woltran, 2010; Brewka et al., 2011). Instead of an attack relation or just attack and support, arguments are connected via generic links and acceptance condition functions, the latter which are Boolean functions that specify how to understand the links between arguments.

NMR properties have also been studied for argumentation, such as for ABA by Čyras and Toni (2015) and Čyras and Toni (2016), and for ASPIC$^+$-style argumentation by Dung (2014) and Dung (2016).

In particular for AI and law, computational argumentation is a topic of much interest (some overviews are Prakken and Sartor, 2015; Bench-Capon, 2020). Modelling argumentative processes as made by humans was already of interest, since this human arguing is crucial in legal practice. With legal reasoning being seen as non-monotonic and a typical example in common-sense reasoning or defeasible reasoning, connections between law and computational argumentation got deeper[2]. Legal reasoning has been a motivation for specific

[1]Strictly speaking, while ADFs subsume AAs by strictly containing them, in a sense AA is still as expressive as ADF, since Dung and Thang (2018) show that AA frameworks can be defined with support trees for ADFs as arguments, and doing so allows capturing ADF semantics based only on a simpler notion of attack (although at the level of support trees). Additionally, Brewka et al. (2011) showed that there is a polynomial-time conversion from an ADF into an AA framework by adding new argument nodes that codify the semantics of acceptance condition functions. Still, this is not necessarily a weakness of ADFs, since ADFs may be easier to interpret, as Brewka et al. (2011) argue.

[2]Interestingly, this was not without some controversy in legal philosophy. In particular, Alchourrón (1993, quoted by Maranhão (2006)) has argued against non-monotonic logics on philosophical grounds as obfuscating how a normative system is changed or affected by defeasible reasoning. His position was that changes would be made explicit with belief revision (to which Alchourrón himself made central contributions, being one of the authors of the AGM postulates (Alchourrón et al., 1985)). Later, Alchourrón (1996) seemed more favourable to it, when discussing defeasible reasoning for legal interpretation, by trying to

developments of argumentation, such as value-based argumentation (Bench-Capon, 2002; Atkinson and Bench-Capon, 2021) and rule-based argumentation (Prakken, 1997; Prakken and Sartor, 1996). Arguments have also been a central aspect of CBR systems, which we will see in the next section.

1.1.2 Case-based reasoning

In AI, case-based reasoning (CBR) is a methodology that uses particular occurrences in the past in order to solve the current issue, by finding those previous occurrences and adapting their solution to the current circumstances (Kolodner, 1992; Watson and Marir, 1994; Richter and Weber, 2013; López, 2013). While there is a multitude of possible methods for CBR, what they have in common is that a task in a new problem is solved by finding previous relevant situations and using their solutions in order to find a solution for the new problem. This is at times presented as the 4R cycle (Aamodt and Plaza, 1994): i) retrieve similar cases; ii) reuse their solutions for the new case; iii) revise the solution, evaluating it; iv) retain the case with their solution. In this it is assumed there is a casebase, containing previous cases with their respective solutions, and mechanisms for retrieving, reusing, revising, and retaining are dependent on context and instantiation.

CBR has been presented as being compatible with different forms of representation, such as tabular (feature-value) data, points in a Euclidean space, or more symbolic, logic-based representations (Richter and Weber, 2013, ch. 5; Mitchell, 1997, ch. 8). While this compatibility may be very flexible, the choice of representation is closely tied to the decisions on how to implement the CBR steps, since it affects how notions such as search, similarity, and adaptation are defined. A paradigmatic example of CBR, although in a particularly simple form, is the famous *k-nearest neighbours* (kNN) algorithm (Cover and Hart, 1967; for textbook presentations, Mitchell, 1997 and Bishop, 2007, p. 124), which solves a task (usually classification or regression) by finding the k most similar past cases and weighting their outputs in order to find the solution for

infer what were the intentions of (a real or hypothetical) legislator. This is discussed as well by Bulygin (2003) and in more length by Maranhão (2006). An interesting philosophical view is that legal rules are not necessarily defeasible, but that they are usually so is a consequence of how rules are treated and of the power ascribed to judges (Schauer, 2012). More in depth discussion in terms of legal philosophy is outside the scope of this work.

a new case. In a sense, forms of CBR can be seen as generalisations of kNN.

As the comparison with kNN suggests, and since CBR works by solving new problems from previous examples, CBR can be seen as a machine learning model (Mitchell, 1997). In comparison to other machine learning models, CBR was proposed with at least the following advantages in mind:

1. training can be easier, since typically processing of training examples can be delayed until time of classification (Mitchell, 1997);

2. solving problems by re-using past situations is a common way of thinking for humans, and indeed CBR was originally proposed with the goal of modelling this way of human reasoning[3];

3. retrieved past cases can be seen as the sources for explanations (Molnar, 2022, ch. 7). Indeed, CBR has been proposed for the goal of explanation before (Kolodner, 1992, §1.1.4), as well as for justification (Kolodner, 1992, §1.2.1).

Argumentation has been mentioned as a motivation and connection to case-based reasoning for a long time (Kolodner, 1992, esp. §1.2; Riesbeck and Schank, 1989, p. 10, §1.4). Many of those connections have been developed in AI and law, as we will see below. Still, and even if influential work in CBR is closer in time to the seminal work of Dung (1995), in computational argumentation only more recently has AA been proposed as a method for directly modelling case-based reasoning (Čyras et al., 2016a; Čyras et al., 2016b; Cocarascu et al., 2018; Cocarascu et al., 2020b). This approach, called *AA-CBR* , contrasts with other approaches that do not use AA (e.g. Verheij, 2017), or only use AA to codify semantics of, e.g., structured argumentation that in turn is used for CBR (e.g. Prakken et al., 2015a). In *AA-CBR*, past cases become arguments, attacks between arguments capture notions of specificity or priority, and the solution of the problem is selected via evaluation of the arguments. The *AA-CBR* approach is presented formally in Section 2.2.1 and it is the focus of much of our attention in this thesis especially in Chapter 3, but also in Section 5.5.

[3]Indeed, Riesbeck and Schank (1989, §1.3, p. 7) have gone so far as stating that: "Case-based reasoning is the essence of how human reasoning works".

Recently, CBR has been used as a technique for few-shot natural language question answering over knowledge bases, in a neuro-symbolic fashion (Das et al., 2020b; Das et al., 2020a; Das et al., 2022). Another neuro-symbolic style of CBR has been through the usage of a *prototype layer* in a neural network, which during training time selects prototypes from the training data and pushes encodings of inputs to be similar to at least one prototype (Li et al., 2018). This has been applied to images, such as bird classification (Chen et al., 2019) and mass lesions in mammography (Barnett et al., 2021), in order to provide interpretability to image classifications. CBR has also been implemented as a Bayesian probabilistic model, which clusters data by finding a feature subspace that is relevant for that cluster and a prototype (a data point) that represents that cluster (Kim et al., 2014).

Case-based reasoning in AI and law

AI and law is both a field in which much of CBR was developed (Bench-Capon, 2017) and one in which the connections between CBR and argumentation have flourished (Bench-Capon, 2020). One of the most influential systems, HYPO, was proposed as a way of creating argumentative dialogues with cases (Rissland and Ashley, 1987; Ashley, 1991; more formally presented by Ashley, 1989). Indeed, HYPO is seen as an important initial work in the field of CBR itself (Watson and Marir, 1994; Rissland et al., 2005) HYPO is considered seminal in the field of AI and law, by thinking of cases and arguments as main objects in legal reasoning, and representing argument moves, which motivated much further work (Bench-Capon, 2017). Some examples are the system CATO (Aleven and Ashley, 1997; Aleven, 2003a), which introduces the notion of factors[4] to the representation of cases, combinations between CBR and rule-based systems (Rissland and Skalak, 1989), and models of CBR as priority among rules, also for integrating with rule-based reasoning (Prakken and Sartor, 1998), the usage of argumentation schemes for CBR (Prakken et al., 2015b; Grabmair, 2016; Grabmair, 2017a), among many others. More recently, representation

[4]Factors and dimensions are elements for representing cases in legal CBR. In the words of Bench-Capon (2017), factors are "stereotypical patterns of fact that favour one party or the other". They are binary, being either present or absent, and contrasted with dimensions, each dimension being a continuous representation, (totally) ordered by how much it benefits one party in the case.

of case-based reasoning for law using ADFs was the theme of the work of Al-Abdulkarim (2017) (see also Al-Abdulkarim et al., 2015a). Values have also been incorporated into some CBR formalisms (Grabmair and Ashley, 2011; Al-Abdulkarim et al., 2015b), including with quantitative effects (Grabmair, 2016; Grabmair, 2017a).

Work in HYPO and CBR in AI and law has also opened the way for the formalisation of how precedents constrain future decisions (Horty, 2004; Horty and Bench-Capon, 2012; Rigoni, 2015; Horty, 2019; Bench-Capon and Atkinson, 2021; Prakken, 2021), which will be discussed later in Section 2.2.2.

1.1.3 Explainable artificial intelligence

Recent advances in AI have been very impressive, with successes in recommender systems, search, image classification, and more recently image and text generation (Russell and Norvig, 2020, §1.4). Of particular importance to the legal domain are large language models (LLMs) (Devlin et al., 2019; Brown et al., 2020), currently based on the Transformer architecture of neural networks (Vaswani et al., 2017), whether instruction-tuned LLMs (Chung et al., 2022) or not (Devlin et al., 2019). Examples of the former are OpenAI's ChatGPT (OpenAI, 2022; OpenAI, 2023) and Google's Bard (Google, 2023), both currently deployed to millions of users. Such systems are under scrutiny of both the general public (Maslej et al., 2023, §8.2) and academics, including legal scholars (Perlman, 2022) and AI and law researchers (Tan et al., 2023; Nguyen et al., 2023). Despite their impressive capacities on generating text, they bring many risks (Bender et al., 2021; Weidinger et al., 2022; Kumar et al., 2022). Important risks to users include disseminating false information or fabricating it altogether ("hallucinating") (Weidinger et al., 2022, §2.3), creating illusion of understanding (Bender and Koller, 2020), and generating unreliable explanations about its own outputs (Ye and Durrett, 2022), among many others. Indeed, as LLMs show emerging behaviour, there are currently no reliable techniques for controlling their behaviour or making them express values of their creators (Bowman, 2023). In law, current LLMs have been shown to have serious trustworthiness risks, including fabricating legal provisions and legal cases altogether, what can be highly misleading, since they are presented fluently (Tan et al., 2023).

We can trace many of those serious issues in numerous state-of-the-art AI systems, including LLMs, to them being opaque models, that is, to the fact that those models have internal representations that are hard to interpret and understand (Gunning and Aha, 2019). This means that such systems are still unable to have their decisions and behaviours *reliably* explained to humans. This opaqueness is an obstacle to human-machine collaboration, including with users, deployers, and regulators, and even with AI systems developers themselves (Gunning and Aha, 2019; Miller, 2019). This is particularly a challenge for deep learning systems, which in many scenarios have high performance, but are typically only used as black-boxes, without ways to provide their grounds for decision or set expectations of how the system will behave in the future. Still, explainability is a topic that has been of interest since early symbolic expert systems, due to simple deduction traces being obscure for a lay user. Thus, a focus of recent years is explainable artificial intelligence (XAI), which has the goal of creating methods that allow systems to produce explanations. Explanations are means to better communicate with users, in turn improving not only joint performance, but also trust and other socially relevant goals, such as fairness, transparency, and accountability (Doran et al., 2017; Adadi and Berrada, 2018; Miller, 2019; Mittelstadt et al., 2019; Guidotti et al., 2019; Mueller et al., 2019; Molnar, 2022). This resurgence of interest in the topic is reflected in academic publications, funding opportunities (DARPA, 2016; Gunning and Aha, 2019) and national and supranational strategies and guidances for AI, including from the UK (UK Department for Business, Energy & Industrial Strategy, 2022b; UK Department for Business, Energy & Industrial Strategy, 2022a; ICO and Alan Turing Institute, 2022; UK Cabinet Office, 2021), OECD (2019), and the EU (European Commission, 2020).

In the words of the UK Department for Business, Energy & Industrial Strategy (2022a) (emphasis added):

> "Achieving explainability of AI systems at a technical level remains an important research and development challenge. Presently, the logic and decision making in AI systems cannot always be meaningfully explained in an intelligible way, although in most settings this poses no substantial risk. However, in some settings the public, consumers and businesses may expect and benefit from trans-

> parency requirements that improve understanding of AI decision-making. **In some high risk circumstances, regulators may deem that decisions which cannot be explained should be prohibited entirely - for instance in a tribunal where you have a right to challenge the logic of an accusation**."

and

> "For regulatory purposes this means that the logic or intent behind the output of systems can often be extremely hard to explain, or errors and undesirable issues within the training data are replicated. This has potentially serious implications, such as when decisions are being made relating to an individual's health, wealth or longer term prospects, or when **there is an expectation that a decision should be justifiable in easily understood terms - such as legal dispute**."

There are even discussions in legal literature regarding whether explainability exists as a legal right in the GDPR (e.g. see Wachter et al. (2017) *contra*; Selbst and Powles (2017), Kaminski (2019) *pro*; Naudts et al. (2022, §3.4), for a discussion).

This topic is of great relevance particularly for high-stake decisions, scenarios where collaboration with humans is important, or highly sensitive to social, ethical, or legal values. Those criteria are valid not only for AI in healthcare (Durán, 2021), but also for AI in law as well. Indeed, AI and law is also a field that has been aware of recent XAI research, and where the notion of explanation has been central (Atkinson et al., 2020). This is so since, in law, practitioners are required at all times to justify their stances, requests, and decisions, parties and judges alike. Thus, in modelling legal reasoning or developing tool to help legal practitioners, the notion of explanation cannot be an afterthought. Still, the field does not offer centralising theories or methodologies, but offers a range of techniques not necessarily exclusive to this field, such as rule-base reasoning, CBR, and argumentation.

Argumentation is a field that contributes to XAI (for a survey, see Cocarascu et al., 2020a). It was used to provide a framework for explanations in interactive recommendations (Rago et al., 2021a; Rago et al., 2020a), as an explanation

layer for neural networks (Dejl et al., 2021; Sukpanichnant et al., 2021), optimisation (scheduling) (Čyras et al., 2020; Čyras et al., 2019b), causal models (Rago et al., 2022) and Bayesian networks (Albini et al., 2020; Timmer et al., 2015); as a methodology for decision-making from which explanations can be extracted (Zhong et al., 2019); among other work (Amgoud, 2021; Potyka et al., 2022).

More transparent machine learning models such as inductive logic programming (ILP), focused in learning logic programmes from examples (Muggleton and Raedt, 1994; Law et al., 2019), have also been used to produce explanations (Rabold et al., 2018), or studied with a focus on human interpretability (Schmid et al., 2016).

CBR has been proposed as post-hoc explainability layer for more opaque systems, including machine-learning-based ones (Nugent and Cunningham, 2005). Additionally, explanation in a general sense has always been presented as an explicit goal of CBR (Kolodner, 1992).

AA-CBR itself has been proposed on XAI grounds: as an XAI method that uses CBR implemented via argumentation (Cocarascu et al., 2018; Cocarascu et al., 2020b), allowing access to argumentation methods, such as arbitrated dispute trees, for creating explanations (Čyras et al., 2019a). Other CBR systems based on argumentation have been proposed for the same purposes (Prakken and Ratsma, 2022a).

1.2 Current limitations and objectives

We adopt the assumption from the NMR literature that at least some NMR properties are crucial for human-style reasoning. This way, NMR analyses of AI systems which aim to capture aspects of reasoning, including CBR and XAI, are important. While we are guided to this assumption by i) the importance of those properties in the NMR literature, and by ii) previous empirical studies with humans showing non-monotonic inferential behaviour, we highlight that the importance of the NMR properties discussed in the thesis for the considered fields of AI is a working hypothesis, acting as a starting point for investigation, but which could be falsified by future research (Popper, 2002). With this working hypothesis in mind, we can discuss limitations we identified in the

literature and how they correspond to the objectives of this thesis.

While much of CBR work has been in the context of knowledge-based systems (Aamodt, 1993; Watson and Marir, 1994; Richter and Weber, 2013), or even on top of NMR languages (particularly in legal CBR, including argumentation (Prakken and Sartor, 1998; Prakken et al., 2015b; Al-Abdulkarim et al., 2015a; Al-Abdulkarim, 2017)), CBR has not been itself analysed as an NMR system, as far as we know. This means that NMR languages have been used to instantiate CBR systems, but considering the CBR system itself as a reasoning system that could satisfy or not particular NMR properties is a discussion absent from the literature. This is what we see as a main gap in the literature, since, by thinking of CBR as a form of reasoning, some properties of NMR can be discussed for it. Connecting those two fields opens the way for new questions that are absent from previous work, such as: i) whether specific models have some of those NMR properties; ii) if models can be built satisfying certain properties; and iii) what is the impact of those properties in applications, among possibly others. Impact in applications can happen in many ways, such as in prediction performance (for classification systems), or effectiveness with users (e.g., for explanations). Although modelling CBR in an NMR language implies that properties of the language in general extend to this instantiation, we believe that CBR raises particular concerns, deserving a more focused discussion. This motivates our Objectives 1 and 3 below.

In the modelling of legal reasoning, many models in the literature are non-monotonic and presented as such (Rotolo and Sartor, 2018; Sartor, 2018), including the just mentioned CBR models. However properties of NMR and their relevance in legal reasoning in general are not usually discussed in depth. They can be mentioned (e.g. Rotolo and Sartor, 2018), but are not subject of much consideration. An exception is by Prakken (1997), where those properties are briefly mentioned and discussed, but are argued against, as non-essential or non-intuitive in a context of rule application. A similar discussion is made by Prakken and Modgil (2018), in the broader context of rule-based argumentation. A specific discussion on the effects of the (lack of) those NMR properties on legal reasoning properly speaking or in CBR is absent. A second exception is in deontic logic, where in some versions of it, equivalents of NMR properties are discussed. For instance, cut, the idea that, if a set of premises with one

of its conclusions added as a premise can derive a second conclusion, then the initial set of premises can derive the second conclusion by itself (See Section 2.1.1), has been discussed via the name of deontic detachment (Greenspan, 1975; Nute and Yu, 1997), and Parent (2001) is the first to discuss cautious monotonicity in (dyadic) deontic logic. Semantic characterisation of such logics satisfying some of those properties is also the subject of the work of Parent (2014). Still, not only have discussions about those properties been generally scarce in legal reasoning generally, but there is a clear limitation in that, to the best of our knowledge, those properties have not been discussed for legal CBR. This motivates our Objective 2.

In explanations, many recent approaches have been developed, but unifying theories or formal guarantees are still lacking (Doran et al., 2017; Gunning and Aha, 2019; Mueller et al., 2019; Miller, 2019; Mueller et al., 2019; Mittelstadt et al., 2019), despite efforts (Watson et al., 2021; Potyka et al., 2022). Thus, contributions are made each via a different angle or for a specific class of problems. In particular, one way of using AI artefacts is interactively: applications in which the user can enter in feedback loop interactions with the system (Teso and Kersting, 2019; Sokol and Flach, 2020; Rago et al., 2020b; Rago et al., 2021b; Mosqueira-Rey et al., 2022). As far as we know, literature on explainability does not offer a theoretical analysis of this interactive scenario. We here aim on contributing to an analysis via NMR, modelling interactions between human and machine via queries (inputs) and answers (outputs and explanations). This motivates our Objective 4. We now turn to our full list of objectives.

In this thesis, we aim to contribute by proposing NMR analyses of classes of models not typically seen as such. The NMR view yields both new criteria of agreement to human-style commonsense reasoning for evaluation of models and new models developed with such criteria in mind. Our objectives in this thesis thus are:

1. To fill a gap between CBR and NMR. Can CBR be seen as NMR? If so, what NMR properties does CBR have or should have? Can we build a CBR model that satisfies some of those properties? In particular, how do such connections hold in the context of argumentation-based CBR?

2. To understand what is the role or are the effects of NMR properties in legal reasoning, focusing on legal CBR.

3. To study the relevance of CBR in the context of data-centric AI, in particular in machine-learning-based classification tasks, and evaluate impacts of NMR properties in it.

4. To add a new theoretical understanding of interactive explanations as NMR.

1.3 Thesis contributions

In this thesis, we aim to contribute to the topic of human-style reasoning in AI in the following ways:

1. By proposing an NMR analysis of classification models, with a focus on CBR systems – to the best of our knowledge, this is the first time CBR is analysed in terms of NMR. (Objective 1; Chapter 3)

2. By analysing an existing argumentation-based model of CBR, called $AA\text{-}CBR$, in this way, proving it does not satisfy some NMR properties considered desirable in that literature, such as cautious monotonicity. (Objective 1; Chapter 3)

3. By proposing a variant of $AA\text{-}CBR$ that does satisfy many of those aforementioned properties, and that also empowers a principled treatment of noise in incoherent casebases. (Objective 1; Chapter 3)

4. By presenting the impact of NMR properties on a legal case study. (Objective 2; Chapter 3)

5. By presenting ways by which $AA\text{-}CBR$ models can be mined from data, and empirically evaluating the differences between the original $AA\text{-}CBR$ and our proposed variant on a legally relevant task (recidivism prediction). (Objectives 2 and 3; Chapter 4)

6. By presenting a formal model of an interactive scenario in explanations as (non-monotonic) reasoning, and showing how explanation methods can

be modelled in this way, such as arbitrated dispute trees for $AA\text{-}CBR$. (Objective 4; Chapter 5)

1.4 Thesis structure

In Chapter 2, we present required background. This includes the definition of NMR and the properties that are used within the thesis, as well the necessary concepts from argumentation. The formal definition of $AA\text{-}CBR$ is then presented, as well as a model of precedential constraint discussed in the AI and law literature. We also briefly present required notions of machine learning and of XAI.

In Chapter 3, we present our analysis of classifiers as NMR and apply it to $AA\text{-}CBR$. We show that, as usually defined, $AA\text{-}CBR$ is not cautiously monotonic, using a counterexample. We proceed to define a new variation of $AA\text{-}CBR$, that we call $cAA\text{-}CBR_{\succeq}$, and prove it satisfies cautious monotonicity and cut. We both give an algorithm for the new variation and present a characterisation of it in terms of a restriction of the casebase to a particular subset of cases. We then present a case study on a legal casebase of the U.S. trade secrets domain, frequently discussed in the AI and law literature (Rissland and Ashley, 1987; Brüninghaus and Ashley, 2003; Bench-Capon, 2017), and discuss the impacts of satisfying or not NMR properties in the legal CBR context.

In Chapter 4, we present an approaches for learning the notion of relevance in $AA\text{-}CBR$ using decision trees. We empirically evaluate this approach on the COMPAS dataset about criminal recidivism, and discuss the differences between using the original definition of $AA\text{-}CBR$ and our new variation $cAA\text{-}CBR_{\succeq}$. We also present preliminary results on a novel approach for learning this notion of relevance using autoencoders, a type of neural network that can be trained using only unlabelled data.

In Chapter 5, we present a formal modelling of an interactive explanation process between user and AI system. We propose modelling explanations using a notion of entailment that can be non-monotonic. This allows us to adapt properties from NMR to the context of interactive explanations. Furthermore, we analyse arbitrated dispute trees, an explanation method for argumentation,

using the proposed methods.

Finally, in Chapter 6, we conclude the thesis, summarising our contributions and presenting directions for future work.

1.5 Publications

Below is the list of publications which contain work presented in this thesis:

- Chapter 3 builds on:

    1. Guilherme Paulino-Passos and Francesca Toni (2020). 'Cautious Monotonicity in Case-Based Reasoning with Abstract Argumentation'. In: *CoRR* abs/2007.05284. 18th International Workshop On Non-Monotonic Reasoning. Informal proceedings. arXiv: 2007.05284. URL: https://arxiv.org/abs/2007.05284

    2. Guilherme Paulino-Passos and Francesca Toni (2021). 'Monotonicity and Noise-Tolerance in Case-Based Reasoning with Abstract Argumentation'. In: *Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning, KR 2021, Online event, November 3-12, 2021*. Ed. by Meghyn Bienvenu et al., pp. 508–518. ISBN: 978-1-956792-99-7. DOI: 10.24963/kr.2021/48. URL: https://doi.org/10.24963/kr.2021/48

- Chapter 4 builds on:

    3. Guilherme Paulino-Passos and Francesca Toni (2023). 'Learning Case Relevance in Case-Based Reasoning with Abstract Argumentation'. In: *Legal Knowledge and Information Systems - JURIX 2023: The Thirty-sixth Annual Conference, Maastricht, the Netherlands, 18-20 December 2023*. (Accepted, forthcoming.)

- Chapter 5 builds on:

    4. Guilherme Paulino-Passos and Francesca Toni (2022a). 'On Interactive Explanations as Non-Monotonic Reasoning'. In: *CoRR* abs/2208.00316. Workshop on Explainable Artificial Intelligence

(XAI) at IJCAI 2022. DOI: 10.48550/ARXIV.2208.00316. arXiv: 2208.00316. URL: https://doi.org/10.48550/arXiv.2208.00316

5. Guilherme Paulino-Passos and Francesca Toni (2022b). 'On Interactive Explanations as Reasoning'. XLoKR 2022: The Third Workshop on Explainable Logic-Based Knowledge Representation at KR2022. Informal proceedings. URL: https://drive.google.com/file/d/1eWsB1up6Il_-L81GaTClW06W3oKV9tEw/view?usp=sharing

6. Guilherme Paulino-Passos and Francesca Toni (2022c). 'On Monotonicity of Dispute Trees as Explanations for Case-Based Reasoning with Abstract Argumentation'. In: *1st International Workshop on Argumentation for eXplainable AI co-located with 9th International Conference on Computational Models of Argument (COMMA 2022), Cardiff, Wales, September 12, 2022*. Ed. by Kristijonas Čyras et al. Vol. 3209. CEUR Workshop Proceedings. CEUR-WS.org. URL: http://ceur-ws.org/Vol-3209/8465.pdf

Chapter 2

Background

In this chapter, we present the technical background necessary for this thesis. In Section 2.1, we first give background in non-monotonic reasoning (NMR), including description of formal properties, in Section 2.1.1, followed by argumentation, including abstract argumentation (AA) in Section 2.1.2 . This is core background necessary for Chapters 3, 4, and 5). We then turn to previous work on case-based reasoning (CBR) in Section 2.2, including a model of CBR with AA, $AA\text{-}CBR$, in Section 2.2.1, and to precedential constraint in Section 2.2.2. While $AA\text{-}CBR$ will be used in all of Chapters 3, 4, and 5 (Section 5.5), precedential constraint will be important in 3 (Section 3.6). Section 2.3 gives background on machine learning methods that are used later in Chapter 4, including decision trees (Section 2.3.1) and autoencoders (Section 2.3.2). Finally, we present some explainable AI (XAI) methods in Section 2.4 that will be important for Chapter 5.

2.1 Non-monotonic reasoning and argumentation

2.1.1 Non-monotonic reasoning

As discussed in Section 1.1.1, NMR has been proposed for modelling inferences made by humans that were not captured by classical logic. Here we will present this idea formally. We will start with an otherwise unspecified lan-

guage containing negation[1]. Later, when we are interested in applying those concepts, we will instantiate this language. For now, we will treat them as entirely abstract.

Definition 2.1. Assume a set $\mathcal{L}$ which we call a *language*. We will refer to elements a, b, c in $\mathcal{L}$ as *formulas*. Let a function $\neg : \mathcal{L} \to \mathcal{L}$ be called a *negation function*, where we will write $\neg a$ as a shorthand for $\neg(a)$.

Our interest in defining a language is in the notion of an inference relation $\vdash$. Intuitively, the focus is, starting from a body of previous information, what can be concluded, derived, inferred from it. This relation is defined from sets of formulas to formulas, so that the following are valid: $\Gamma \vdash a$, where $\Gamma \subseteq \mathcal{L}$.[2]

Definition 2.2. Assume a relation $\vdash$: $2^{\mathcal{L}} \to \mathcal{L}$ that will be called an *inference relation*.

Non-monotonicity Properties

Now, our interest here is in what properties are possessed by $\vdash$. We will first formally state the properties and then discuss their intuitive meaning (Gabbay, 1984; Makinson, 1994).

Definition 2.3. An inference relation $\vdash$ is said to satisfy:

1. *monotonicity*, iff $\Gamma \vdash a$ and $\Gamma \subseteq \Gamma'$ imply $\Gamma' \vdash a$;

2. *reflexivity*, iff $\Gamma \vdash a$ for every $a \in \Gamma$;

3. *cautious monotonicity*, iff $\Gamma \vdash a$ and $\Gamma \vdash b$ imply $\Gamma \cup \{a\} \vdash b$;

[1]Notice that this definition of negation is weaker than syntactical closure under negation. This is compatible with, but does not require, an infinite number of formulas $\neg a$, $\neg\neg a$, $\neg\neg\neg a$, and so on. One could simply have for instance $\neg(a) = b$, such that $\neg(\neg(a))$ (that is, $\neg(b)$) is simply a (syntactically). In this sense $\neg$ is a symbol in the metalanguage, not in the object language (which could still contain its own negation). This difference is not major, but simplifies languages used later, sparing this infinity by recursion in the language.

[2]Literature on NMR and on logic more generally has presented different forms of considering the notion of inference. The tradition derived from Gentzen (1969) uses the notion of sequents, where both the left-hand side and the right-hand side contain sets of formulas, meaning that at least one of the formulas in the right is a consequence of the set of formulas in the left. Differently, Tarski (1956) uses a notion of a consequence operator, which is a function from a set of formulas to another set, with the latter standing for the (possibly infinite) set of all formulas which are a consequence of the former. As more typical in the NMR literature, we will use the relation from a set to a single conclusion (Makinson, 1994).

4. *cut*, iff $\Gamma \vdash a$ and $\Gamma \cup \{a\} \vdash b$ imply $\Gamma \vdash b$;[3]

5. *cumulativity*, iff $\vdash$ is both cautiously monotonic and satisfies cut;

6. *rational monotonicity*, iff $\Gamma \vdash a$ and $\Gamma \nvdash \neg b$ imply $\Gamma \cup \{b\} \vdash a$;

7. *completeness*, iff either $\Gamma \vdash a$ or $\Gamma \vdash \neg a$.

Monotonicity is the property held by classical logic, and its negation, non-monotonicity, is what defines the field of NMR, as discussed in Chapter 1. Its intuitive meaning is that, if some set of information is sufficient for a conclusion, no extra information reverts this conclusion. Thus, conclusions are strict. Contrarily, in non-monotonic inference relations, even if some set of information is sufficient for inferring some conclusion, adding more information to the premises may result in losing this conclusion. This is what makes such conclusions defeasible.

Cautious monotonicity[4] means that, if two conclusions can be derived from some premises, then adding one of the conclusions as a premise does *not* result in losing the other conclusion. This is still compatible with this addition of information resulting in more conclusions that were previously not inferrable. This is a limited form of monotonicity, since is restricted to conclusions of some initial set of premises.

[3]The reader may wonder why this is different from the property in classical logic, particularly in proof theory, also called cut, which is: $\Gamma \vdash a$ and $\Gamma' \cup \{a\} \vdash b$ imply $\Gamma \cup \Gamma' \vdash b$ (Smullyan, 1995, p. 111). This second version, which for simplicity we can here call *stronger cut*, is not equivalent to our presentation of cut. Indeed, stronger cut is stronger than cut, which can be seen by setting Γ' to Γ in stronger cut, resulting in cut. Now, to see that it is too strong for our purposes, we can show that stronger cut and reflexivity together imply monotonicity. This can be seen in the following way: consider the non-empty sets Λ and Λ'. Now assume $\Lambda \vdash b$. We will show that $\Lambda \cup \Lambda' \vdash b$ as well, which is monotonicity (since any superset of Λ can be written this way). Take $a \in \Lambda$, and instantiate stronger cut with $\Gamma = \Lambda \cup \Lambda'$ and $\Gamma' = \Lambda$. Then we have i) $\Gamma \vdash a$ by reflexivity, since $a \in \Lambda$ and $\Lambda \subseteq \Gamma$. Additionally, $\Gamma' \vdash b$ by assumption and, since $\Gamma' \cup \{a\} = \Gamma'$, we have ii) $\Gamma' \cup \{a\} \vdash b$. Thus by (i), (ii), and stronger cut, we have $\Gamma \cup \Gamma' \vdash b$, but $\Gamma \cup \Gamma'$ is simply $(\Lambda \cup \Lambda') \cup \Lambda$, which is $\Lambda \cup \Lambda'$, completing the proof.

[4]A note on terminology: some authors call this property "*cautious monotony*", and analogously for other properties containing the word "monotonicity". Our preference for this writing is that, linguistically, "monotonicity" is a noun derived from the adjective "monotonic", while "monotony" is a noun derived from the adjective "*monotone*". Since the literature has a strong preference for "monotonic" over "monotone" (the field itself is usually called "non-monotonic reasoning", as opposed to the much less common "non-monotone reasoning"), we opt for "monotonic" and maintain consistency throughout by using "monotonicity", instead of "monotony".

Cut is the converse of cautious monotonicity. It means that if from some set of premises one can derive a conclusion, and if from that same set with this conclusion added one can derive a second conclusion, then this second conclusion was derivable in the first place only from the set of premises. This corresponds to the idea in logic that derivations are made step by step. In other words, if something is inferrable using an intermediate conclusion, then it is inferrable from the original information. Thus lemmas are allowed as a way of finding what is a conclusion from the original set of premises. Another way of seeing it is that adding intermediate conclusions does *not* result in the increase of conclusions from the initial premises.

Cumulativity is simply cautious monotonicity and cut together. This means that addition of conclusions in the premises does not decrease or increase the set of consequences, that is, does not affect it in any way. Thus any derived conclusions can be freely accumulated with the premises and this will not change the set of conclusions.

Rational monotonicity is an even stronger property. It states that if some premises imply a conclusion and does not imply the negation of some other information, then assuming as well this new information does not cause the loss of the initial conclusion. That is, what is not denied is treated as irrelevant regarding previous conclusions.

Finally, completeness simply means that any set of premises infer, for any information, either it or its negation. That is, it has a stance, or decides, all questions. This property will be important in our analysis of classifiers in Chapter 3.

2.1.2 Argumentation

Computational argumentation provides a flexible and general way of working with non-monotonic reasoning, as discussed in Section 1.1.1. It is also focused in modelling dialectical behaviours among agents, such as human argumentation. We will present the technical content necessary for this work, which is abstract argumentation (Dung, 1995).

2.1.3 Abstract argumentation

Definition 2.4 (Dung, 1995)**.** An *abstract argumentation framework (AF)* is a pair $(Args, \rightsquigarrow)$, where $Args$ is a set (of *arguments*) and $\rightsquigarrow \subseteq Args \times Args$ is a binary relation on $Args$.

For $\alpha, \beta \in Args$, if $\alpha \rightsquigarrow \beta$, then we say that α *attacks* β and that α is an *attacker of* β.

Definition 2.5 (Dung, 1995)**.** Let $E \subseteq Args$, a set of arguments, and $\alpha \in Args$, an argument. Then we say E *defends* α if for all $\beta \rightsquigarrow \alpha$ there exists $\gamma \in E$ such that $\gamma \rightsquigarrow \beta$.

Then we can turn to the only notion of semantics used in this thesis.

Definition 2.6 (Dung, 1995)**.** The *grounded extension* of $(Args, \rightsquigarrow)$ is the set $\mathbb{G} = \bigcup_{i \geqslant 0} G_i$, where G_0 is the set of all unattacked arguments, and $\forall i \geqslant 0$, G_{i+1} is the set of arguments that G_i defends.

This definition is constructive, offering an algorithm (even if naïve), for building the extension. Nofal et al. (2021) present different algorithms for computing the grounded extension.@@latex:[5]@@

Example 2.7. We can see in a small example how AA is non-monotonic, requiring first just some preliminaries. Assume given a AF $(Args, \rightsquigarrow)$. With $Args'$ a subset of $Args$, we can define the restriction of $(Args, \rightsquigarrow)$ to $Args'$ as $(Args, \rightsquigarrow)\downarrow_{Args'} = (Args', \rightsquigarrow \cap (Args' \times Args'))$ (Baroni et al., 2018a). Now we can define an inference relation with respect to $(Args, \rightsquigarrow)$ and the grounded extension as: $Args' \vdash \alpha$ iff α is in the grounded extension of $(Args, \rightsquigarrow)\downarrow_{Args'}$.

Now we can consider a simple example, where $Args = \{\alpha, \beta, \gamma\}$, and $\rightsquigarrow = \{(\gamma, \alpha), (\beta, \alpha)\}$. This is illustrated in Figure 2.1. We can see that $\{\alpha\} \vdash \alpha$, while $\{\alpha, \beta\} \nvdash \alpha$, since β is in the grounded extension of $(Args, \rightsquigarrow)\downarrow_{\{\alpha,\beta\}}$ and it attacks α, thus α is not in the grounded extension. Thus, this a non-monotonic inference relation.

[5]Semantics of abstract argumentation can also be given in terms of labellings, which are partitions of arguments into the subsets `in`, `out`, and `undecided`, corresponding to whether an argument is to be accepted, rejected, or will be considered undecided (Caminada, 2006; Caminada, 2007; Modgil and Caminada, 2009). We will not use this way of defining semantics in the thesis, but we will make references to it in minor comments.

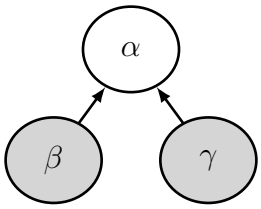


Figure 2.1: AF $(Args, \rightsquigarrow)$ from Example 2.7. Arguments represented as nodes, with grounded extension shaded.

A particular case is of *well-founded* AFs: AFs in which there is no infinite sequence $(\alpha_i)_{i\in[n]}\mathbb{N}$ of arguments such that for each i, α_i attacks α_{i+1} (Dung, 1995). In the case of finite AFs, being well-founded is equivalent to being acyclic (Baroni et al., 2018a). Important facts about the grounded extension are that, for any $(Args, \rightsquigarrow)$, the grounded extension $\mathbb{G}$ always exists and is unique and, if $(Args, \rightsquigarrow)$ is well-founded, extensions under other important semantics (e.g. preferred extensions, each of which does not attack itself, defends all of its arguments and is maximal w.r.t. set inclusion; stable extensions, which attack every argument not in the extension) are equal to $\mathbb{G}$ (Dung, 1995). Given $(Args, \rightsquigarrow)$, we will sometimes use $\alpha \in (Args, \rightsquigarrow)$ to stand for $\alpha \in Args$.

2.2 Case-based reasoning

Case-based reasoning (CBR) is a methodology that uses previous cases for solving a new problem, as discussed in Section 1.1.2. We will now present three approaches to CBR: one is via abstract argumentation (Section 2.2.1), the second is precedential constraint (Section 2.2.2), and the third is K-nearest neighbours (Section 2.2.3).

In order to illustrate each CBR model, we will present an example based on a famous hypothetical discussion in jurisprudence of vehicles in the park (Hart, 1958; Fuller, 1958; Schlag, 1999; Schauer, 2006; Schauer, 2008; Shecaira, 2015; Struchiner et al., 2020). While initially proposed to discuss the nature of law and of rule-following behaviour, our aim here is more mundane. We will simply use it as a toy example for the models below.

Example 2.8. Imagine a public park containing a simple, general rule: "No vehicles are allowed in the park". This rule was disputed in two particular

cases:

1. A person entered park cycling and was stopped and denied entry by a new park ranger. The issue was raised in front of the competent authority (say, the local judge), which ruled it in favour cyclist. It was highlighted that a bicycle is an exception to the rule.

2. An ambulance entered the park, and the competent authority once again allowed entry, for being an ambulance.

Assume there are two other situations for which someone would like to know the legality of entering the park:

3. for an electric bicycle;

4. for a bicycle ambulance;[6]

For the sake of simplicity, we will consider three features: b (bicycle), m (motorised), and a (ambulance), and two outcomes: $-$ (entry forbidden) and $+$ (entry permitted). Then the two past cases will be modelled as $\alpha = (\{b\}, +)$, and $\beta = (\{a, m\}, +)$, respectively. As for the two question cases, they can be modelled as $\gamma = (\{b, m\}, ?)$ (since electric bicycles are, indeed, motorised), and $\epsilon = (\{a, b\}, ?)$, respectively. We will call $D = \{\alpha, \beta\}$ the *casebase* or *dataset*.

2.2.1 Case-based reasoning with abstract argumentation

The approach used here uses AA for CBR, and is thus called *AA-CBR*. It was presented by Čyras et al. (2016a), and later generalised by Cocarascu et al. (2020b). The presentation here is an adaptation of the latter one. Since there are multiple presentations of *AA-CBR*, we will call this generalised presentation *AA-CBR*$_{\preceq}$. This is motivated by the fact that it is parametrised by a relation $\preceq$, as we will see below.

All incarnations of *AA-CBR*, including *AA-CBR*$_{\preceq}$, map a *dataset* D of *examples* labelled with an *outcome* and an *unlabelled example* (with unknown

[6]The reader will be forgiven for thinking this is an absurd example created by the at times idiosyncratic literature of legal philosophy or of AI, but it can be assured that not only bicycle ambulances exist, but they are also part of the London Ambulance Service, under the name of "cycle responders" (London Ambulance Service NHS Trust, 2017).

outcome) into an AF. The dataset may be understood as a *casebase*, the labelled examples as *past cases* and the unlabelled example as a *new case*: we will use these terminologies interchangeably throughout. We treat examples/cases as having characterisation, which describes them, and can also be understood as their features (Čyras et al. (2016a) originally used binary features), and outcomes are chosen from two available ones, one of which is selected up-front as the *default outcome*. Finally, we assume that the set of characterisations of (past and new) cases is equipped with a partial order $\preceq$ (whereby $\alpha \prec \beta$ holds if $\alpha \preceq \beta$ and $\alpha \neq \beta$ and is read "α is less *specific* than β") and with a relation $\not\sim$ (whereby $\alpha \not\sim \beta$ is read as "β is *irrelevant* to α"). The converse order $\succeq$ and its strict version $\succ$ are defined as usual (Davey and Priestley, 2002). Formally:

Definition 2.9 (Adapted from Cocarascu et al. (2020b))**.** Let X be a set of *characterisations*, equipped with partial order $\prec$ and binary relation $\not\sim$. Let $Y = \{\delta_o, \bar{\delta_o}\}$ be the set of (all possible) *outcomes*, with δ_o the *default outcome*. Then, a *casebase* D is a finite set such that $D \subseteq X \times Y$ (thus a *past case* $\alpha \in D$ is of the form (α_C, α_o) for $\alpha_C \in X$, $\alpha_o \in Y$) and a *new case* is of the form $(N_C, ?)$ for $N_C \in X$. We also discriminate a particular element $\delta_C \in X$ and define the *default argument* $(\delta_C, \delta_o) \in X \times Y$.

A casebase D is *coherent* iff there are no two cases $(\alpha_C, \alpha_o), (\beta_C, \beta_o) \in D$ such that $\alpha_C = \beta_C$ but $\alpha_o \neq \beta_o$, and it is *incoherent* otherwise.

For simplicity of notation, we sometimes extend the definition of $\succeq$ to $X \times Y$, by setting $(\alpha_c, \alpha_o) \succeq (\beta_c, \beta_o)$ iff $\alpha_c \succeq \beta_c$, and analogously for $\not\sim$.[7]

Definition 2.10 (Adapted from Cocarascu et al. (2020b))**.** The *AF mined from a dataset D and a new case* $(N_C, ?)$ is $(Args, \rightsquigarrow)$, in which:

- $Args = D \cup \{(\delta_C, \delta_o)\} \cup \{(N_C, ?)\}$;
- for $(\alpha_C, \alpha_o) \in D$, $(\beta_C, \beta_o) \in D \cup \{(\delta_C, \delta_o)\}$, it holds that $(\alpha_C, \alpha_o) \rightsquigarrow (\beta_C, \beta_o)$ iff
    1. $\alpha_o \neq \beta_o$,

[7] In previous work, $\succeq$ was directly given over $X \times Y$ (Cocarascu et al., 2020b). Note that, when D is coherent, our "lifted" order $\succeq$ is guaranteed to be a partial order on $X \times Y$ (and thus equivalent to the one from Cocarascu et al. (2020b)), but when D is incoherent anti-symmetry may fail for two cases with different outcomes but same characterisation, and thus $\succeq$ is merely a preorder on $X \times Y$.

2. $\alpha_C \succeq \beta_C$, and
3. $\nexists(\gamma_C, \gamma_o) \in D \cup \{(\delta_C, \delta_o)\}$ with $\alpha_C \succ \gamma_C \succ \beta_C$ and $\gamma_o = \alpha_o$;

- for $(\beta_C,\beta_o) \in D \cup \{(\delta_C, \delta_o)\}$, it holds that $(N_C, ?) \rightsquigarrow (\beta_C,\beta_o)$ iff $(N_C, ?) \not\curlyvee (\beta_C,\beta_o)$.

The *AF mined from a dataset D alone* is $(Args', \rightsquigarrow')$, with $Args' = Args \setminus \{(N_C, ?)\}$ and $\rightsquigarrow' = \rightsquigarrow \cap (Args' \times Args')$.

Note that if D is coherent, then the "equals" case in item 2 of the definition of attack will never apply. As a result, the AF mined from a coherent D (and any $(N_C, ?)$) is guaranteed to be well-founded, in the sense of Dung (1995). That is, acyclic, being finite.

Definition 2.11 (Adapted from Cocarascu et al. (2020b))**.** Let $\mathbb{G}$ be the grounded extension of the AF mined from D and $(N_C, ?)$, with default argument (δ_C, δ_o). The *outcome for N_C* is δ_o if (δ_C, δ_o) is in $\mathbb{G}$, and $\bar{\delta_o}$ otherwise.

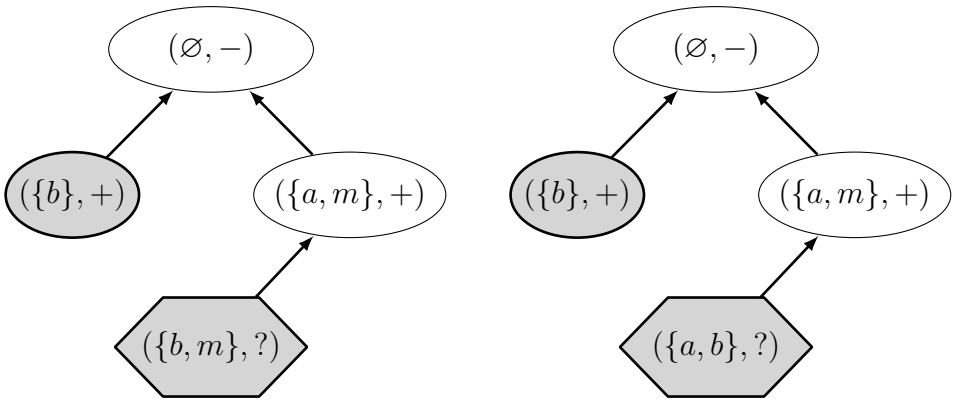
We will refer to the AF mined from D and $(N_C, ?)$, with default argument (δ_C, δ_o), as $AF_{\succeq}(D, N_C)$, and to the AF mined from D alone as $AF_{\succeq}(D)$. Also, for short, given $AF_{\succeq}(D, N_C)$, with default argument (δ_C, δ_o), we will refer to the outcome for N_C as $AA\text{-}CBR_{\succeq}(D, N_C)$.[8] Unless otherwise stated, we will assume arbitrary X, Y, D, $(N_C, ?)$, and (δ_C, δ_o) (satisfying the previously defined constraints). Finally, we will refer to $AA\text{-}CBR_{\succeq}$ instantiated with $\succeq = \supseteq$, $\not\curlyvee = \not\supseteq$, and $(\delta_C, \delta_o) = (\varnothing, -)$ as $AA\text{-}CBR_{\supseteq}$.

This approach will be considered in detail in the next chapters, particularly in Chapter 3.

Example 2.12 (continuing from Example 2.8)**.** We will model this example with $AA\text{-}CBR_{\supseteq}$, that is, using the representation of sets of features from Example 2.8 and the subset relation as the partial order. The default case is $(\varnothing, -)$, representing that, by default, a vehicle is not permitted in the park. The features represent aspects, characteristics, or reasons that could potentially change the decision of a case.

The AFs mined from D and each of the cases to be decided are represented in Figure 2.2. For both cases γ and ϵ, the result is the same: the new case

[8] In the notation we omit (δ_C, δ_o), and leave it implicit instead for readability.



(a) AF mined from D and γ.

(b) AF mined from D and ϵ.

Figure 2.2: Illustration of Example 2.12, argumentation frameworks mined from the casebase and each new case. Past cases represented as ellipses, new cases as hexagons, and arguments in the grounded extension are shaded.

attacks β, since it is irrelevant (for γ, because it is not about an ambulance; for ϵ, because it is not motorised), but not α, about the bicycle. Then α attacks the default argument, which results in the latter not being in the grounded extension, and thus the prediction being $+$, that is, the vehicle is allowed to enter. Notice how each feature is treated as a reason for a possible exception, and the new case not having the feature implies the past cases that do have it are all irrelevant.

2.2.2 Precedential constraint

A line of work in legal CBR has been on precedential constraint (Horty, 2004; Horty and Bench-Capon, 2012; Rigoni, 2015; Bench-Capon and Atkinson, 2021; Horty, 2021; Prakken, 2021; Canavotto, 2022; Van Woerkom et al., 2023). The goal is modelling formally what it means for a past decision to bind, or constraint, future decisions. This is crucial for legal reasoning with case law, due to the idea of *stare decisis*, the principle according to which previous decisions must be followed in the future, as a matter of authority (Schauer, 2014).

Here we will cover only the factor-based formulation of precedential constraint, where cases are described as binary features favouring either the plaintiff

or the defendant (Horty and Bench-Capon, 2012). More general approaches have been developed, such as for dimensions, where cases are described using multi-valued features, each called a dimension, which is ordered by how much it favours one of the parties or the other (Rigoni, 2015; Bench-Capon and Atkinson, 2021; Horty, 2021; Prakken, 2021). Bench-Capon and Atkinson (2021) offer a model about reasoning with precedents in which cases are structured into issues, and then cases not only state preferences between factors, but they may also establish how cases are associated to issues, and how case facts are mapped to factors. This is formalised with prioritised rules, in the style of Prakken and Sartor (1998), which models *a fortiori* reasoning via defeasible logic.

The basic notion of *stare decisis* is that previous decisions must be kept, in that similar cases should be treated in the same way. However, there are discussions in AI and law (literature discussed above) and also in legal philosophy itself (Horty, 2004; Raz, 2009; Schauer, 2014; Horty, 2015; Lewis, 2021; for comparative law analyses, Bell, 1997; Algero, 2004; MacCormick and Summers, 2016) on what this constraint actually entails.

In the factor-based analysis, there are two disjoint sets of features X^{Δ} and X^{Π}, being the factors that favour, respectively, the defendant and the plaintiff. We will describe the result model. Formally, each such case is of the form $(\alpha^{\Pi} \cup \alpha^{\Delta}, o)$ with $\alpha^{\Pi} \subseteq X^{\Pi}$, $\alpha^{\Delta} \subseteq X^{\Delta}$, and $o \in \{\Pi, \Delta\}$ the case outcome. For compatibility of notation with *AA-CBR*, if $\alpha = (\alpha^{\Pi} \cup \alpha^{\Delta}, o)$ is a case, we will define $\alpha_C = \alpha^{\Pi} \cup \alpha^{\Delta}$ as the characterisation of α, and $\alpha_o = o$ as the outcome of α. We will refer to the opposing outcome of o as $\bar{o}$ (that is, $\bar{\Pi} = \Delta$ and $\bar{\Delta} = \Pi$), to any subset of $X^{\Delta} \cup X^{\Pi}$ as a *characterisation*[9], and if F is a characterisation, $F^{\Pi} = F \cap X^{\Pi}$ and $F^{\Delta} = F \cap X^{\Delta}$.

Definition 2.13 (Horty, 2004)**.** Let F and G be two characterisations. Then for an outcome o, $F \leq_o G$ iff $F^o \subseteq G^o$ and $G^{\bar{o}} \subseteq F^{\bar{o}}$.

This captures the idea of factor preference: that a characterisation G has more reasons for side o than F by containing all of the favourable reasons in F but not any other unfavourable one. We will abuse notation and compare directly cases via $\leq_o$ meaning their characterisations.

[9]Originally called a *fact situation* in the literature of precedential constraint.

Definition 2.14 (Horty, 2004)**.** Let $\alpha = (\alpha^{\Delta} \cup \alpha^{\Pi}, o)$ be a case and F be a characterisation. We say α *forces* F to outcome o by iff $\alpha \leq_o F$. If CB is a set of cases (a casebase), and there is a case α which forces F to outcome o, then we say CB forces F to outcome o. If there is no characterisation G such that CB forces G to both outcomes o and $\bar{o}$, then CB is said to be *consistent*.

This is a notion of precedential constraint for the result model, in which a casebase forces a characterisation into an outcome via *a fortiori* reasoning. Consistency is defined by not forcing different results.

Example 2.15 (continuing from Example 2.8)**.** We will now model our example with precedential constraint. Compared to the representation of sets of features from Example 2.8, more background information is required: the mapping of each feature to one of the two outcomes (that is, treating each feature as a factor). We will here model denying entrance as favouring the defendant (thus $\Delta = -$), and allowing entrance as favouring the plaintiff (thus $\Pi = +$). This way, the factors in favour of allowing the vehicle to enter are $X^{\Pi} = \{a, b\}$, since being an ambulance and being a bicycle are all reasons for permitting entrance, while $X^{\Delta} = \{m\}$, since being motorised is a reason for denying entrance (say, because motors are potentially loud). Now we are ready to consider the two new cases, γ and ϵ.

For ϵ, about the bicycle ambulance, the outcome is forced to be $-$. This can be seen by noticing that $\alpha \leq_{\Pi} \epsilon$, since $\alpha_C^{\Pi} = \{b\} \subseteq \{a, b\} = \epsilon_C^{\Pi}$, while $\alpha_C^{\Delta} = \varnothing = \epsilon_C^{\Delta}$. It can also be verified that $+$ is not a forced outcome (and indeed, D is consistent). Therefore precedential constraint determines that bicycle ambulances are allowed in the park, based on the precedent that bicycles in general are allowed in the park and this is sufficient.

On the other hand, a similar analysis will show that the case γ, of the electric bicycle, is not forced to have either outcome. The bicycle case α is not sufficient, since the case γ has an additional factor in favour of the defendant, the park, namely, m. Similarly β is not sufficient, since γ is not an ambulance, and thus does not have the a factor in favour of the plaintiff. We can interpret this case as being undecided in terms of factor-based precedential constraint.

This is not intrinsically a weakness of the model, and its suitability depends on context. For instance, if γ with outcome $-$ is added to the casebase, precedential constraint would automatically constraint the case of motorised

vehicles in general to be $-$, in the absence of other factors in favour of the plaintiff. This is an intuitive result. For comparison, *AA-CBR* decides γ in an intuitive way (see Example 2.12), but can also accommodate the addition of a case $(\{m\}, +)$ to the casebase, since it has no concept of factors. This might be considered undesirable, since it can be required that being motorised can never be a reason for allowing entry. In Chapter 3 we discuss the topic of modelling factors in *AA-CBR*.

While the formal model of precedential constraint has been developed further by an active line of research (Rigoni, 2015; Bench-Capon and Atkinson, 2021; Horty, 2021; Prakken, 2021; Peters et al., 2022; Canavotto, 2022; Van Woerkom et al., 2023; Canavotto and Horty, 2023), in this thesis we restrict our scope to this basic model.

2.2.3 K-nearest neighbours

We now simultaneously turn to our last CBR model and start moving in the direction of (what is more usually considered to be) machine learning. K-nearest neighbours (kNN) is, at the same time, the prototypical example of CBR and a typical example of machine learning (as mentioned in Section 1.1.2). It will not be a focus of our investigation, but it is worth presenting, as it is a paradigmatic form of CBR (Mitchell, 1997; Bishop, 2007; Richter and Weber, 2013).

In this method, a notion of distance between cases is defined depending on application (a usual one being the Euclidean distance). Then, given a new case, the k cases in the casebase which have minimum distance are retrieved. Those cases are the *nearest neighbours*, a notion that is central in CBR in general (Richter and Weber, 2013), and which has also been referred to by other names, such as "most on-point cases" (Ashley, 1989).

After retrieval, the output of the new case is defined as a function of the output of those k cases. Typical functions are the most frequent output (mode) or the mean of this set outputs.

2.3 Machine learning

Machine learning (ML) is now widely used in AI and is responsible for many recent advances. Its core idea consists in developing a system that learns from data, that is, from examples, instead of directly implementing desired behaviour (Mitchell, 1997; Abu-Mostafa et al., 2012). In this work, we use *AA-CBR* in combination with other ML methods for a hybrid approach. Since *AA-CBR* requires a partial order $\preceq$ over the data, as well as a default case, a natural question is from where those come. While they could be hand-crafted, depending on application, an interesting use-case is when they are originated from data. For instance, it could be from the casebase itself, or perhaps from a wider dataset containing the casebase.

Of course, CBR itself (and *AA-CBR* in particular) can be seen as a type of machine learning, since the casebase is essentially a training dataset, and the end result can be usually seen as a function for classification (if the output is discrete) or regression (if continuous) (Mitchell, 1997). However, assumptions regarding the underlying data (such as assumption of an underlying probability distribution, and training set being sampled independently from it) may be different. In that context, CBR methods are sometimes called lazy learning, since most of the computation would be left for inference time, while other ML methods would be eager methods, calculating a decision function upfront (Mitchell, 1997). Depending on how CBR or kNN is applied, however, those classifications may become fuzzy. For instance, there might have a representation learning step for kNN which is carried out at training time, or a computationally costly indexing step.

2.3.1 Decision trees

One of the methods we will use later is decision trees (Breiman et al., 1984). Decision trees are a classical decision method, and learning decision trees is itself a classical and well-studied machine learning method (Bishop, 2007, Section 14.4, for a textbook presentation).

A decision tree is a binary tree where each node contains a split, that is, a test of the form $f > \tau$, for f a feature and τ a threshold, except for the leaf nodes, which contain a class. A new instance is classified by following the path

according to its evaluation on each split, and is then classified by the class of the (unique) leaf node it falls into.

One of the most used algorithms for learning decision trees is CART (Breiman et al., 1984). In a brief presentation, it consists of recursively growing a binary tree by a greedy algorithm. Starting from a single node where all the data is, the node is expanded by choosing a feature and a threshold level for creating a new split. This is chosen by exhaustively testing every feature choice for the one that minimises a loss function, such as cross-entropy or Gini index, that depend on which partition of the data is made by the split. Each new leaf then corresponds to the majority class of the data in each. Usually the tree is grown up to a maximum depth or until nodes cannot be split anymore (since every data instance in the leaves is of the same class), and at a second step, the tree is pruned, according to performance on a validation dataset.

2.3.2 Autoencoders

Autoencoders are a type of neural network which aim to reconstruct the input. They are usually seen as having two parts: an encoder e, which maps the original input to a hidden encoding, and a decoder d, which maps the hidden encoding to the original input. Usually the idea is not having perfect equality, but making $d(e(x))$ approximate x during training subject to some kind of constraint, such that the network is pushed to learn interesting properties of the data. Autoencoders are used for learning representations of data for downstream tasks by dimensionality reduction, for compression, indexing, and anomaly detection (An and Cho, 2015). We will be interested in them particularly in Chapter 4, where they will have a use for learning relevance in CBR.

A probabilistic approach is the variational autoencoder (Kingma and Welling, 2014; Rezende et al., 2014). In this approach, a probabilistic model is assumed in which the hidden encoding is sampled from a prior probability distribution and the decoder maps from the hidden encoding to a posterior probability distribution, from which the reconstructed input is sampled. During training, the encoder is used to map from the input to a probability distribution from which the hidden encoding is sampled. There are two terms in the loss function: a reconstruction term for pushing the reconstruction output to be similar to the

original input and a Kullback-Leibler divergence between the inferred values of the hidden encoding and the established prior (Goodfellow et al., 2016). Interestingly, Ghosh et al. (2020) show that even for generation a variational autoencoder can be replaced by a deterministic autoencoder if the latter is properly regularised.

2.4 Explainable AI

As presented in Section 1.1.3, explainable AI (XAI) has the goal of providing AI systems with ways to explain their decisions, that is, to communicate reasons for the decisions to humans (Ribeiro et al., 2018; Ribeiro et al., 2018; Atkinson et al., 2020; Mueller et al., 2019; Miller, 2019; Ignatiev et al., 2019; Molnar, 2022). In this section we will briefly discuss some concepts and methods which have been proposed in the literature that will be relevant for our later discussion.

One important distinction in XAI is between global versus local explanations. A global explanation is an explanation about the way the entire system works, while a local explanation is restricted to a particular decision (or perhaps a group of decisions), trying to explain it, without regard for the rest of the input space (Mueller et al., 2019, p. 66; Molnar, 2022, Section 3.3).

A second distinction is between intrinsically interpretable models and post-hoc methods (Molnar, 2022, Section 3.2). Intrinsically interpretable models, as the name suggest, are considered to be straightforward to understand, and thus can be easily understood by humans. Typical examples are decision trees with few nodes and linear models with few features. Post-hoc methods, on the other hand, are methods applied after the model is ready (for instance, after it has been trained), and might require some kind of procedure to be run on it (Ribeiro et al., 2016b).

Finally, a third distinction is between model-specific explainability methods and model-agnostic (Molnar, 2022, Section 3.2). Model-specific use mechanisms of specific models in order to build explanations, while model-agnostic use only their input-output behaviour, that is, build explanations treating the model purely as a blackbox. Some examples of model-specific explanations are gradient methods for neural networks, such as Grad-CAM (Selvaraju et al.,

2017) and SmoothGrad (Smilkov et al., 2017), among others (Molnar, 2022, Section 10.2 for a survey).

Examples of local post-hoc model-agnostic methods are LIME (Ribeiro et al., 2016a), Anchors (Ribeiro et al., 2018), and Shapley value feature attribution (Štrumbelj and Kononenko, 2010; Štrumbelj and Kononenko, 2014; Lundberg and Lee, 2017; Sundararajan and Najmi, 2020). LIME and Anchors are perturbation-based methods. This means that, in order to explain the decision for an input x, a perturbation is applied to x, that is, we define a probability distribution on inputs similar to x (and described via an interpretable representation), and we draw a sample from this distribution. Then, this sample is used in order to train an intrinsically interpretable model. In the case of LIME, this model is a (sparse) linear model. In Anchors, it is a decision rule, defined such that, if the input satisfies the body of the rule (a set of predicates in the interpretable representation), then the decision is a specific output. Anchors is trained such that those rules have high-precision (in the perturbation probability distribution). Methods based on the Shapley value, such as SHAP, on the other hand, use a method based on game theory, where, in a rough description, (interpretable) features are seen as players in a cooperative game, and the importance of each feature is evaluated by calculating a contribution value based on the outputs given each feature presence or absence (Štrumbelj and Kononenko, 2010; Štrumbelj and Kononenko, 2014; Lundberg and Lee, 2017; Sundararajan and Najmi, 2020).

2.4.1 Arbitrated dispute trees

Dispute trees (Dung et al., 2006; Dung et al., 2007) have been defined as explanations to *AA-CBR* since Čyras et al. (2016b), and further developed in the form of *arbitrated* dispute trees for an extended version of it in Čyras et al. (2019a). This last form is the one which will interest us, due to its symmetry regarding default and non-default outcomes. For compatibility with our presentation, we will redefine the main concepts, but omit results proven originally, since the adaptations are minimal. This adaptation generalises to AA in general, instead of only to *AA-CBR* models.

Definition 2.16 (Adapted from Čyras et al. (2019a))**.** Given $(Args, \rightsquigarrow)$ and

a specific topic argument δ, an **arbitrated dispute tree** (ADT) is a tree $\mathcal{DT}$ such that:

1. every node of $\mathcal{DT}$ is of the form $[\mathtt{N}{:}\alpha]$ for $N \in \{W, L\}$ and $\alpha \in \mathit{Args}$, any such node being called an N-node labelled by argument α;
2. the root of $\mathcal{DT}$ is labelled by the topic argument δ and is a W-node, if δ is in the grounded extension; and a L-node otherwise;
3. for every W-node n labelled $\alpha \in \mathit{Args}$ and for every β attacker of α in $(\mathit{Args}, \rightsquigarrow)$, there is a child m of n such that m is a L-node labelled by β;
4. for every L-node n labelled by $\alpha \in \mathit{Args}$, there is exactly one child which is a W-node labelled by some β attacker of α; and
5. there are no other nodes in $\mathcal{DT}$.

W-nodes are *winning* nodes, while L-nodes are *losing* nodes.

An example of a dispute tree is in Figure 2.3. Dispute trees have been originally defined as being possibly infinite, while Čyras et al. (2019a) assumes casebases are coherent, and one can thus prove every arbitrated dispute tree in that context is finite[10]. While we make no such assumption, we consider infinite dispute trees to be inappropriate for explaining, so we retain the original definition but explicitly define ADT explanations to limit ourselves to the finite case.

Definition 2.17 (Adapted from Čyras et al. (2019a))**.** An **arbitrated dispute tree explanation** is a finite arbitrated dispute tree.

2.5 Summary

In this chapter we have presented the technical background for the thesis, including non-monotonic reasoning, case-based reasoning, machine learning, and explainable AI. We will now move to our novel contributions, starting with the analysis of NMR properties of classification and CBR, in particular, *AA-CBR*.

[10]This not being the case, some results from the original presentation are inapplicable, such as (Čyras et al., 2019a, Prop. 3.4).

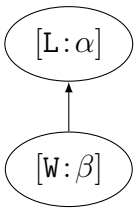


Figure 2.3: ADT with topic argument α for the AF in Example 2.7. Argument β is in the grounded extension, while α is not. Notice how γ does not label a node in this ADT, since each L-node has a single child. Besides, there is no node that is a child of the node labelled by β, since β has no attacker in $(\mathit{Args}, \rightsquigarrow)$.

Chapter 3

Cumulativity in Case-Based Reasoning with Abstract Argumentation

Abstract argumentation-based models of case-based reasoning ($AA\text{-}CBR$ in short) have been proposed (see Section 2.2.1), originally inspired by the legal domain (Athakravi et al., 2015; Čyras et al., 2016b; Čyras et al., 2016a), but also applicable as classifiers in different scenarios (Cocarascu et al., 2020b). {In particular in legal reasoning, $AA\text{-}CBR$ has been used for the task of revising legal rule-bases based on logic programming (Fungwacharakorn et al., 2021; Fungwacharakorn et al., 2022), and motivated other approaches in legal CBR (Prakken and Ratsma, 2022b).} However, the formal properties of $AA\text{-}CBR$ as a reasoning system remained largely unexplored. In this chapter, we focus on analysing the non-monotonicity properties (Section 2.1.1) of a *regular* version of $AA\text{-}CBR$ (that we call $AA\text{-}CBR_{\succeq}$), to be defined in this chapter.

Specifically, we start with some examples presenting issues with the original formulation of $AA\text{-}CBR$ (Section 3.1). We then formally define a notion of trainable classifiers as non-monotonic reasoning (with a focus on CBR systems) (Section 3.2), and use this notion to prove that $AA\text{-}CBR_{\succeq}$ is not cautiously monotonic (Section 3.4). We then define a notion of a restricted casebase consisting of all "surprising" and "sufficient" cases (in a sense to be defined) in the original casebase, and use this notion to define a novel variation of $AA\text{-}CBR_{\succeq}$ which is cautiously monotonic (Section 3.5). We show an algorithm

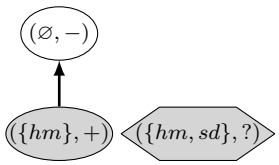
for building this restricted casebase. We also prove that this variation of $AA\text{-}CBR_{\succeq}$ satisfies the properties of cumulativity and rational monotonicity as a by-product. In addition, this variant empowers a principled treatment of noise in "incoherent" casebases, thus providing a form of incoherence-learning reasoning. Next, we illustrate $AA\text{-}CBR$ and non-monotonicity questions on a case study on the U.S. Trade Secrets domain, a legal casebase (Section 3.6), showing that the violation of the NMR properties can have legal impacts, such as manipulability. In this case study we show not only that we can encode the idea of precedential constraint in $AA\text{-}CBR$, but also that our novel variant results in a legal CBR system that is provably cumulative, which is unique in the literature, to the best of our knowledge. Finally, we discuss some limitations of our work and contextualise it further in the landscape of legal CBR (Section 3.7), before concluding the chapter.

Non-monotonicity properties have already been studied in structured argumentation, e.g. for ABA, ABA+ (Čyras and Toni, 2015; Čyras and Toni, 2016), $ASPIC^{+}$ (Dung, 2014; Dung, 2016) and logic-based argumentation (Hunter, 2010). In this chapter we study them for classification, in particular for CBR and more specifically to the model $AA\text{-}CBR_{\succeq}$, which applies abstract argumentation to CBR classification.

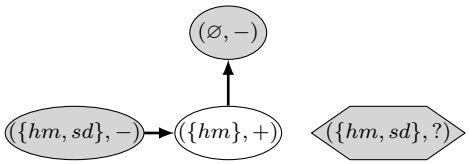
3.1 Motivating illustration

In this section we introduce a simple setting for the informal illustration of the original $AA\text{-}CBR$, its non-monotonicity and the desirability of some restrictions thereof, as well as problems raised by the presence of incoherence in the casebase when deploying $AA\text{-}CBR$. Thus, this section serves as a motivating illustration for our approach, which restricts non-monotonicity and is incoherence-tolerant.

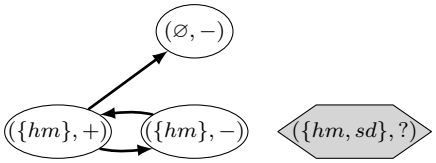
Example 3.1 (Non-monotonicity)**.** Consider a very simplified legal system built by cases and adhering, like most modern legal systems, to the principle by which, unless proven otherwise, no person is to be considered guilty of a crime. This can be represented by a "default argument" $(\varnothing, -)$, indicating that, in the absence of any information about any person, the legal system should infer a negative outcome $-$ (that the person is *not* guilty). $(\varnothing, -)$ can



(a) Initial AA framework. *AA-CBR* predicts outcome "+" for the new case.



(b) Revised framework. The added past case changes the *AA-CBR*-predicted outcome to "−".



(c) Incoherent casebase. *AA-CBR* predicts outcome "+", but the default argument is not attacked by arguments in $\mathbb{G}$.

Figure 3.1: AA frameworks when using *AA-CBR* for Examples 3.1 and 3.2. Past cases (with outcomes) and new case (with unknown outcome) are arguments. (Grounded extensions $\mathbb{G}$ are shaded.)

be understood as an argument, in the AA sense, given that it is merely what is called a relative presumption, since it is open to proof to the contrary, e.g. by proving that the person did indeed commit a crime. Let us consider here one possible crime: homicide (hm).[1] In one case, it was established that the defendant committed homicide, and he was considered guilty, represented as $(\{hm\}, +)$.

Consider now a new case $(\{hm, sd\}, ?)$, with an unknown outcome, of a defendant who committed homicide, but for which it was proven that it was in self-defence (sd). In order to predict the new case's outcome by CBR, *AA-CBR* reduces the prediction problem to that of membership of the default argument in the grounded extension (Dung, 1995) $\mathbb{G}$ of the AA framework in Figure 3.1a: given that $(\varnothing, -) \notin \mathbb{G}$, the predicted outcome is positive (i.e. guilty), disregarding sd and, indeed, no matter what other feature this case may have. Thus, up to this point, having the feature hm is a sufficient condition for predicting guilty. If, however, the courts decide that for this

[1]Of course, this is a toy example, not unlike Examples 2.8 and 2.12 earlier. Here we are neither trying to capture how a real legal decision about criminal law would effectively be made, nor are we using terminology corresponding to any specific jurisdiction. In a real legal case, many more considerations would apply, such as issues about evidence, establishing facts, consulting sources of law, and interpreting the law (Schauer, 2009; Verheij, 2020).

new case the defendant should be acquitted, the case $(\{hm, sd\}, -)$ enters in our casebase. Now, having the feature hm is no longer a sufficient condition for predicting guilty, and any case with both hm and sd will be predicted a negative outcome (i.e. that the person is innocent). This is the case for predicting the outcome of a new case with again both hm and sd, in $AA\text{-}CBR$ using the AA framework in Figure 3.1b. Thus, adding a new case to the casebase removed some conclusions which were inferred from the previous, smaller casebase, showing that $AA\text{-}CBR$ is indeed non-monotonic. This does not mean that some restrictions on non-monotonicity might not be desirable. For instance, we might expect in a legal system that, if for the current case law, two cases are to be judged in a certain way, then one of the cases happening in court and indeed being decided in that way would not affect the body of case law itself, thus the outcome for the second case would be expected to be unchanged.

The following example illustrates the challenges posed by incoherent casebases, with noisy cases, in $AA\text{-}CBR$.

Example 3.2 (Noise-intolerance)**.** Consider a different augmentation of the initial casebase in Example 3.1, resulting in the casebase $\{(\{hm\}, +), (\{hm\}, -)\}$, whereby a positive outcome and a negative outcome are both recorded for exactly the same profile for defendants. This can be deemed to be "incoherent"[2]. In $AA\text{-}CBR$, this case would have a positive outcome (guilty), since the default argument is not in the grounded extension. However, this seems unsatisfactory, since the default argument is not attacked by any argument in the grounded extension of the corresponding AA framework (see Figure 3.1c).[3] In some sense, this means that no past case serves as ground for the decision, but the outcome is guilty nonetheless. Thus, what is deciding the outcome is not a

[2]This casebase may result from a limited language for characterising cases, e.g. ignoring the possibility of indicating core differences between defendants, such as that the defendant characterised by $(\{hm\}, -)$ is a minor. We assume here that this incoherence cannot be rectified by a language variation, or simply that it comes from the data, and we cannot remove it based on the data alone. Of course, those limitations do not apply to real-life legal reasoning. In legal practice, judges and other legal decision-makers would typically *distinguish* cases, that is, find reasons for characterising two conflicting cases in different ways, in order to justify the different outcomes (Schauer, 2009; Schauer, 2014, among others).

[3]In terms of semantics via labellings (Caminada, 2006; Caminada, 2007; Modgil and Caminada, 2009), the default argument would have the `undecided` label, not the `in` label nor the `out` label.

relevant case per se, but the incoherence itself. This can be regarded as a form of intolerance to noise. Note that incoherence does not always give this form of noise intolerance in *AA-CBR*: if a past case $(\{hm, sd\}, +)$ were included, the outcome would be the same, but the default argument would be attacked by (some argument in) the grounded extension.

Our form of *AA-CBR* is guaranteed to be tolerant of noise always, independently of the casebase.

3.2 Non-monotonicity analysis of trainable classifiers

In this section we provide a generic analysis of the non-monotonicity properties of data-driven classifiers, using D, X and Y to name a generic dataset, the input set and the output set of a classifier respectively, admitting casebases, characterisations, and outcomes in CBR as special instances. Later in the chapter, we will apply this analysis to both $AA\text{-}CBR_{\succeq}$ and our new variant of it. Typically, a classifier can be understood as a function from an input set X to an output set Y. In machine learning, classifiers are obtained by training with an initial, finite $D \subseteq X \times Y$, called the training set. Thus, a learning algorithm can be seen as a function that maps from the set $2^{X\times Y}$ (which contains all possible sets of pairs $(x, y) \in X \times Y$) to a function $X \to Y$ (Abu-Mostafa et al., 2012, §1.1.1). Our interest here is to consider the variation in input-output behaviour under the change of the training set D, where the details of the separation between the learning step from the inference step are abstracted away.[4] Thus, at the risk of abusing terminology[5], we will

[4]This is also in line with the view of CBR models as lazy learners, where the computation of a model is delayed to inference time (Mitchell, 1997). In any case, this is not material to this modelling, since at which time each function is computed is not particularly important for the definitions to come.

[5]The abuse is calling all of $(2^{X\times Y} \times X) \to Y$ the classifier, instead of limiting the classifier to be a function $X \to Y$. This allows us to avoid writing our analysis at the level of a higher order function $2^{X\times Y} \to (X \to Y)$. In other words, we are calling the operation of learning a classifier on dataset D followed by the application of the learned classifier to input x as simply the (learnable) classifier, treating it as a function $(2^{X\times Y} \times X) \to Y$. This can also be seen as the operation of *uncurrying* from the literature of lambda calculus and functional programming (Michaelson, 1989, §7.14). Uncurrying is the operation of transforming a function with domain A and codomain $B \to C$ into a function with domain $A \times B$ and

characterise a trainable classifier as a two-argument function $\mathbb{C}$ that maps from a dataset $D \subseteq X \times Y$ and from a new input $x \in X$ to a prediction $y \in Y$.[6] Notice that this function is total, in line with the common assumptions that classifiers generalise beyond their training dataset to the entire input set.

Let us model directly the relationship between the dataset D and the predictions it makes via the classifier as an inference system in the following way:

Definition 3.3. Given a classifier $\mathbb{C}$: $2^{X \times Y} \times X \to Y$, let $\mathcal{L} = \mathcal{L}^+ \cup \mathcal{L}^-$ be a language consisting of atoms $\mathcal{L}^+ = X \times Y$ and negative sentences $\mathcal{L}^- = \{\neg(x, y) | (x, y) \in X \times Y\}$. Then, $\vdash_{\mathbb{C}}$ is an *inference relation* from $2^{\mathcal{L}^+}$ to $\mathcal{L}$ such that

- $D \vdash_{\mathbb{C}} (x, y)$ iff $\mathbb{C}(D, x) = y$;
- $D \vdash_{\mathbb{C}} \neg(x, y)$ iff there is a y' such that $\mathbb{C}(D, x) = y'$ and $y' \neq y$.[7]

Intuitively, $\mathcal{L}$ defines a language consisting of atoms (representing labelled examples) and their negations, and $\vdash_{\mathbb{C}}$ describes the input-output behaviour of the resulting model learned using D. Notice the language defined in Definition 3.3 can be seen as a language in the sense of Definition 2.1, by considering the negation of $\neg(x, y)$ to be simply (x, y). Of course, $\vdash_{\mathbb{C}}$ is also an inference relation in the sense of Definition 2.2. Then, we can study non-monotonicity properties of $\vdash_{\mathbb{C}}$, following definitions presented in Section 2.1.1.

Theorem 3.4 (Properties of $\vdash_{\mathbb{C}}$)**.**

1. $\vdash_{\mathbb{C}}$ *is* complete*, i.e. for every* $(x, y) \in (X \times Y)$*, either* $D \vdash_{\mathbb{C}} (x, y)$ *or* $D \vdash_{\mathbb{C}} \neg(x, y)$.

codomain C, for sets A, B and C. This is mostly a notational choice, since the definition below would be similar by replacing, for instance, the term $\mathbb{C}(D, x)$ with $\mathbb{C}(D)(x)$.

[6] Notice that this understanding relies upon the assumption that (the training of) the classifier is deterministic. Of course this is not the case for many machine learning models, e.g. artificial neural networks trained using stochastic gradient descent and randomised hyperparameter search; or even for deterministic training when random splitting between a training and a development sets is used in order to select hyperparameters for a model. This understanding is however in line with recent work using decision functions as approximations of classifiers whose output needs explaining (e.g. see Shih et al., 2019). Moreover, it works well for analysing deterministic CBR systems, including $AA\text{-}CBR_{\succeq}$.

[7] We could equivalently have defined $D \vdash_{\mathbb{C}} \neg(x, y)$ iff $\mathbb{C}(D, x) \neq y$. We have not done so as the used definition can be generalised for a scenario in which $\mathbb{C}$ is not necessarily a total function. This scenario is left for future work. A downside of the definition used is that it does not cover multi-label classification, where multiple outputs could be correct simultaneously. This is also a scenario outside our current scope.

2. $\vdash_{\mathbb{C}}$ *is* consistent*, i.e. for every* $(x, y) \in (X \times Y)$*, it does not hold that both* $D \vdash_{\mathbb{C}} (x, y)$ *and* $D \vdash_{\mathbb{C}} \neg(x, y)$*.*

3. $\vdash_{\mathbb{C}}$ *is cautiously monotonic iff it satisfies cut.*

4. $\vdash_{\mathbb{C}}$ *is cautiously monotonic iff it is cumulative.*

5. $\vdash_{\mathbb{C}}$ *is cautiously monotonic iff it satisfies rational monotonicity.*

Completeness and consistency here are immediate consequences from the fact that $\mathbb{C}$ is a total function. This is in line with how classifiers are typically considered: returning exactly one output for each input. The remaining properties, which are the non-monotonic ones, are essentially a collapse caused by completeness and consistency.

Proof. 1. By definition of $\vdash_{\mathbb{C}}$, directly from the totality of $\mathbb{C}$.

2. By definition of $\vdash_{\mathbb{C}}$, since $\mathbb{C}$ is a function.

3. Let $\vdash_{\mathbb{C}}$ be cautiously monotonic, and assume that $D \vdash_{\mathbb{C}} p$ and $D \cup \{p\} \vdash_{\mathbb{C}} q$, for $p, q \in \mathcal{L}$. We want to prove that $D \vdash_{\mathbb{C}} q$. By completeness (item 1 above), at least $D \vdash_{\mathbb{C}} q$ or $D \vdash_{\mathbb{C}} \neg q$ holds.[8] We will show the first one holds by proving that, by assuming the second case, we arrive at a contradiction. Combining the original assumption that $D \vdash_{\mathbb{C}} p$ with the new assumption that $D \vdash_{\mathbb{C}} \neg q$, by cautious monotonicity we can derive $D \cup \{p\} \vdash_{\mathbb{C}} \neg q$. But also by assumption $D \cup \{p\} \vdash_{\mathbb{C}} q$, which is absurd since $\vdash_{\mathbb{C}}$ is consistent. Therefore the $D \vdash_{\mathbb{C}} \neg q$ case is impossible and thus the first case is true, that is, $D \vdash_{\mathbb{C}} q$, as we wanted. The converse can be proven analogously.

4. Trivial from 3.

5. Since $\vdash_{\mathbb{C}}$ is complete, $D \nvdash_{\mathbb{C}} \neg p$ implies $D \vdash_{\mathbb{C}} p$, and thus rational monotonicity reduces to cautious monotonicity.

□

[8] Here we understand the negation $\neg$ in the sense of 2.1. That is, if $q \in \mathcal{L}^+$ then $\neg q$ is well defined, otherwise $q \in \mathcal{L}^-$ and thus there is $r \in \mathcal{L}^+$ such that $q = \neg r$. Then simply define $\neg q = r$.

3.3 A regular $AA\text{-}CBR$

As seen in Section 2.2.1, Cocarascu et al. (2020b) presented a model of $AA\text{-}CBR$ using a partial order between cases. In there, the concept of irrelevance $\not\sim$ between cases is defined independently from the partial order. Here, we are interested in the particular case of argumentation frameworks (AFs) where those concepts are related in the following way:

Definition 3.5. The AF mined from D alone and the AF mined from D and $(N_C, ?)$, with default argument (δ_C, δ_o), are *regular* when the following holds:

1. the irrelevance relation $\not\sim$ is defined as: $x_1 \not\sim x_2$ iff $x_1 \not\trianglerighteq x_2$, and

2. δ_C is the least (or minimum) element of X, that is, for any $x \in X$, $\delta_C \preceq x$.[9]

This restriction connects the treatment of a characterisation α_C as a new case and as a past case. In particular, the original $AA\text{-}CBR_{\trianglerighteq}$, as presented in Section 2.2.1, is regular. For the remainder of the thesis, when we refer to a mined AF, we will assume it is regular, unless otherwise stated, so that, e.g., $AF_{\succeq}(D, N_C)$ will be the *regular* AF mined from D and $(N_C, ?)$.

Regularity is necessary to guarantee some desirable behaviours, such as a relation between a new case and "most similar" (or *nearest*) cases to the new case. When these nearest cases all agree on an outcome, the prediction is necessarily this outcome. It generalises Proposition 2 of Čyras et al. (2016a) in two ways: 1) by considering the entire set of nearest cases, instead of requiring a unique nearest case; 2) doing so for regular $AA\text{-}CBR_{\succeq}$, instead of its instance $AA\text{-}CBR_{\trianglerighteq}$.

We first define the notion of nearest case, and then state the desirable behaviour in the form of a theorem.

Definition 3.6. A case $(\alpha_C, \alpha_o) \in D$ is *nearest to* N_C iff $\alpha_C \preceq N_C$ and it is maximally so, that is, there is no $(\beta_C, \beta_o) \in D$ such that $\alpha_C \prec \beta_C \preceq N_C$.

[9] Indeed this is not a strong condition, since it can be proved that if $\alpha_C \not\trianglerighteq \delta_C$ then all cases (α_C, α_o) in the casebase could be removed, as they would never change an outcome. On the other hand, assuming also the first condition in Definition 3.5, if $(\alpha_C, ?)$ is the new case and $\alpha_C \not\trianglerighteq \delta_C$, then the outcome is $\bar{\delta_o}$ necessarily.

Theorem 3.7. *If D is coherent and every nearest case to N_C is of the form (α_C, o) for some outcome $o \in Y$ (that is, all nearest cases to the new case agree on the same outcome), then AA-$CBR_{\succeq}(D, N_C) = o$ (that is, the outcome for N_C is o).*

Proof. Let $\mathbb{G}$ be the grounded extension of $AF_{\succeq}(D, N_C)$. An outline of the proof is as follows:

1. We will first prove that each argument in $\mathbb{G}$ is either $(N_C, ?)$ or of the form (β_C, o) (that is, agreeing in outcome with all nearest cases).

2. Then we will prove that if $o = \bar{\delta_o}$ (that is, o is the non-default outcome), then $(\delta_C, \delta_o) \notin \mathbb{G}$ (and thus AA-$CBR_{\succeq}(D, N_C) = \bar{\delta_o}$, as envisaged by the theorem).

3. Finally, by using the fact that $AF_{\succeq}(D, N_C)$ is well-founded (given that D is coherent), and thus $\mathbb{G}$ is also stable, we will prove that if $o = \delta_o$ (that is, o is the default outcome), then $(\delta_C, \delta_o) \in \mathbb{G}$ (and thus AA-$CBR_{\succeq}(D, N_C) = \delta_o$, as envisaged by the theorem).

We will now prove 1-3.

1. By definition $\mathbb{G} = \bigcup_{i \geqslant 0} G_i$. We prove by induction that, for every i, each argument in G_i is either $(N_C, ?)$ or of the form (β_C, o). Then, given that each element of $\mathbb{G}$ belongs to some G_i, the property holds for $\mathbb{G}$.

   (a) For the base case, consider G_0, the set of unattacked cases. Let $\beta = (\beta_C, \beta_o)$ be a case in G_0. We want to show that, if $\beta_C \neq N_C$, then $\beta_o = o$. If β is a nearest case, then by assumption we already have that $\beta_o = o$. We have then to consider the case where β is non-nearest. If $N_C \not\succeq \beta_C$, then $N_C \not\sim \beta$ (by regularity) and thus $(N_C, ?)$ attacks β, contradicting that β in unattacked. So $\beta_C \preceq N_C$. As β is not a nearest case and D is finite, there is a nearest case $\alpha = (\alpha_C, \alpha_o)$ such that $\beta_C \prec \alpha_C \preceq N_C$. By contradiction, assume $\beta_o \neq o$. Let $\Gamma = \{\gamma \in \mathit{Args} \mid \gamma = (\gamma_C, \gamma_o),\ \beta_C \prec \gamma_C \preceq \alpha_C \text{ and } \gamma_o = o\}$. Notice that Γ is non-empty, as $\alpha \in \Gamma$, using again the assumption that nearest cases have outcome o. Γ is a set of "potential attackers" of

β, but only $\preceq$-minimal arguments in Γ do actually attack β. Let ϵ be such a $\preceq$-minimal element of Γ.[10] By construction, ϵ attacks β. Thus β is attacked and not in G_0 as assumed, a contradiction. Hence, $\beta_o = o$, as required.

(b) For the inductive step, let us assume that the property holds for a generic G_i, and let us prove it for G_{i+1}. Let $\beta = (\beta_C, \beta_o) \in G_{i+1} \setminus G_i$ (if $\beta \in G_i$, the property holds by the induction hypothesis). $(N_C, ?)$ does not attack β, as otherwise β would not be defended by G_i, as G_i is conflict-free. Once again, if β is a nearest case, we are done by the assumption that nearest cases have the outcome o. Otherwise, β is not a nearest case, and thus there is a nearest case $\alpha = (\alpha_C, \alpha_o)$ such that $\beta_C \prec \alpha_C \preceq N_C$. Again, assume that $\beta_o \neq o$. Then let $\Gamma = \{\gamma \in \mathit{Args} \mid \gamma = (\gamma_C, \gamma_o),\ \beta_C \prec \gamma_C \preceq \alpha_C \text{ and } \gamma_o = o\}$, with ϵ a $\preceq$-minimal element of Γ. Then ϵ attacks β. However, as G_i defends β, there is then $\eta \in G_i$ such that η attacks ϵ. By inductive hypothesis, η is either $(N_C, ?)$ or $\eta = (\eta_C, o)$. The first option is not possible, as $\epsilon \in \Gamma$, and thus $\epsilon_C \preceq \alpha_C$, and of course $\alpha_C \preceq N_C$. Thus, $\epsilon_C \preceq N_C$ and is thus not attacked by $(N_C, ?)$ (by regularity). This means that (η_C, o) attacks $\epsilon = (\epsilon_C, \epsilon_o)$. But this is absurd as well, as $\epsilon \in \Gamma$ and thus $\epsilon_o = o = \eta_o$. Therefore, our assumption that $\beta_o \neq o$ was false, that is, $\beta_o = o$, as required.

2. If $o = \bar{\delta_o}$, the default argument (δ_C, δ_o) is not in $\mathbb{G}$, since we have just proven that all arguments in $\mathbb{G}$ other than $(N_C, ?)$ have outcome o.

3. If $o = \delta_o$, then let β be an attacker of (δ_C, δ_o), and thus of the form $\beta = (\beta_C, \bar{\delta_o})$ (again see how regularity is necessary, since otherwise $(N_C, ?)$ could be the attacker). β is not in $\mathbb{G}$ and, since $\mathbb{G}$ is also a stable extension, some argument in $\mathbb{G}$ attacks β. This is true for any attacker β of the default argument, and thus the default argument is defended by $\mathbb{G}$. As $\mathbb{G}$ contains every argument it defends, the default argument is in the grounded extension, confirming that the outcome for N_C is δ_o. □

[10] Note that ϵ is guaranteed to exist, as Γ is non-empty and finite, as otherwise we would be able to build an arbitrarily long chain of (distinct) arguments, decreasing w.r.t. $\prec$. However this would allow a chain with more elements than the cardinality of Γ, which is absurd.

Example 3.8 (continuing from Examples 3.1 and 3.2)**.** A simple example of the theorem is considering the casebase of Example 3.1, illustrated in Figure 3.1a. There is a single nearest neighbour to $(\{hm, sd\}, ?)$, which is $(\{hm\}, +)$, and the classification is thus $+$. Again in the AF in Figure 3.1b the example is simple: now with the extra past case $(\{hm, sd\}, -)$, it becomes the single nearest neighbour to $(\{hm, sd\}, ?)$, causing the outcome to be $-$. If Theorem 3.7 tells us is that if this case was instead $(\{hm, sd\}, +)$, the result would be guaranteed to be $+$. As for the AF in Figure 3.1c, $(\{hm\}, +)$ and $(\{hm\}, -)$ are two nearest neighbours to $(\{hm, sd\}, ?)$. In this situation, Theorem 3.7 does not guarantee any behaviour, since the two nearest neighbours have conflicting outcomes. A useful consideration is that in much more complicated AFs it would be unnecessary to compute the entire grounded extension in order to compute the output of new cases for which there is this agreement in outcome of nearest neighbours.

The restriction to regularity also gives us another property that will show itself useful later. This property, which we will present as a lemma, identifies a "core" subset in the casebase for the purposes of outcome prediction of a new case: this amounts to all past cases that are less (or equally) specific than the new case for which the prediction is sought. In other words, irrelevant cases in the casebase do not affect the prediction in regular AFs.

Lemma 3.9. *Let D_1 and D_2 be two datasets. Let $N_C \in X$ be a characterisation, and $D_{iN_C} = \{\alpha \in D_i \mid \alpha \preceq N_C\}$ for $i = 1, 2$. If $D_{1N_C} = D_{2N_C}$, then AA-CBR$_\succeq(D_1, N_C)$ = AA-CBR$_\succeq(D_2, N_C)$ (that is, AA-CBR$_\succeq$ predicts the same outcome for N_C given the two datasets).*

Proof. For $i = 1, 2$, let $AF_i = AF_\succeq(D_i, N_C)$ and the grounded extensions be $\mathbb{G}_i = \bigcup_{j \geqslant 0} G^i_j$. We will prove that $\forall j : G^1_j \subseteq G^2_{j+1}$ and $G^2_j \subseteq G^1_{j+1}$, and this allows us to prove that $\mathbb{G}_1 = \mathbb{G}_2$, which in turn implies the outcomes are the same. Here we consider only $G^1_j \subseteq G^2_{j+1}$, as the other case is entirely symmetric. By induction on j:

- For the base case $j = 0$:

  Let $\alpha \in G^1_0$. We want to prove that $\alpha \in G^2_1$. If $\alpha \in G^2_0$, we are done, as we always have that $G^i_j \subseteq G^i_{j+1}$. Else, we will use $\alpha \notin G^2_0$ to show

that $\alpha \in G_1^2$. Since $\alpha \in G_0^1$, α is relevant to N_C, and thus $\alpha \preceq N_C$ (by regularity). This in turn implies that $\alpha \in D_2$, because $D_{1N_C} = D_{2N_C}$.

On the other hand, as $\alpha \notin G_0^2$, there is a case $\beta \in AF_2$ such that $\beta \rightsquigarrow \alpha$. However, $\beta \notin D_1$, otherwise α would be attacked in AF_1 and thus would not be in G_0^1. But then, since $D_{2N_C} = D_{1N_C} \subseteq D_1$, this means that $\beta \notin D_{2N_C}$, thus $\beta \not\preceq N_C$. Finally, this means that $(N_C, ?) \rightsquigarrow \beta$, and thus G_0^2 defends α. Therefore, $\alpha \in G_1^2$, what we wanted to prove.

- For the induction step, from j to $j+1$:

  Again, if $G_{j+1}^1 \subseteq G_{j+1}^2$, we are done. If not, there is a $\alpha \in G_{j+1}^1 \setminus G_{j+1}^2$. Again we can check that this implies that $\alpha \in D_2$. Now, since $\alpha \in G_{j+1}^1$, then G_j^1 defends it. But now, by inductive hypothesis, $G_j^1 \subseteq G_{j+1}^2$. Therefore, G_{j+1}^2 also defends α, which implies that $\alpha \in G_{j+2}^2$, as we wanted.[11] This concludes the induction.

To conclude, if we consider $\alpha \in \mathbb{G}_1$, by definition of $\mathbb{G}_1$ there is a j such that $\alpha \in G_j^1$. But since $G_j^1 \subseteq G_{j+1}^2 \subseteq \mathbb{G}_2$, we have $\alpha \in \mathbb{G}_2$. This proves that $\mathbb{G}_1 \subseteq \mathbb{G}_2$. The converse can be proven analogously, allowing us to see that $\mathbb{G}_1 = \mathbb{G}_2$. Thus $\delta \in \mathbb{G}_1$ if and only if $\delta \in \mathbb{G}_2$, from which we conclude that $AA\text{-}CBR_{\succeq}(D_1, N_C) = AA\text{-}CBR_{\succeq}(D_2, N_C)$. □

Having stated and shown the properties of regularity in $AA\text{-}CBR_{\succeq}$ above, we are ready to discuss the NMR properties of $AA\text{-}CBR_{\succeq}$.

3.4 $AA\text{-}CBR_{\succeq}$ is not cumulative

Our first main result is about (the lack of) cautious monotonicity of the inference relation drawn from the classifier $AA\text{-}CBR_{\succeq}(D, N_C)$.

Theorem 3.10. $\vdash_{AA\text{-}CBR_{\succeq}}$ *is not cautiously monotonic.*

Proof. We show a counterexample, choosing $X = 2^{\{a,b,c,z\}}$, $Y = \{-, +\}$, and $\preceq = \supseteq$. Define $D = \{(\{a\}, +), (\{c\}, +), (\{a, b\}, -), (\{c, z\}, -)\}$ and $(\delta_C, \delta_o) = (\varnothing, -)$, and two new cases: $N_1 = \{a, b, c\}$ and $N_2 = \{a, b, c, z\}$.

[11] It can be easily verified that, if $E \subseteq Args$ defends an argument γ, and $E \subseteq E'$, then E' also defends γ.

Consider now $AA\text{-}CBR_{\succeq}(D, N_1)$ and $AA\text{-}CBR_{\succeq}(D, N_2)$. We can see in Figure 3.2a that $D \vdash_{AA\text{-}CBR_{\succeq}} (N_1, +)$ and in Figure 3.2b that $D \vdash_{AA\text{-}CBR_{\succeq}} (N_2, -)$.

Finally, let us consider $AF_{\succeq}(D \cup \{(N_1, +)\}, N_2))$ in Figure 3.2c. We can then conclude that $D \cup \{(N_1, +)\} \vdash_{AA\text{-}CBR_{\succeq}} (N_2, +)$ even though $D \vdash_{AA\text{-}CBR_{\succeq}} (N_1, +)$ and $D \vdash_{AA\text{-}CBR_{\succeq}} (N_2, -)$, as required. □

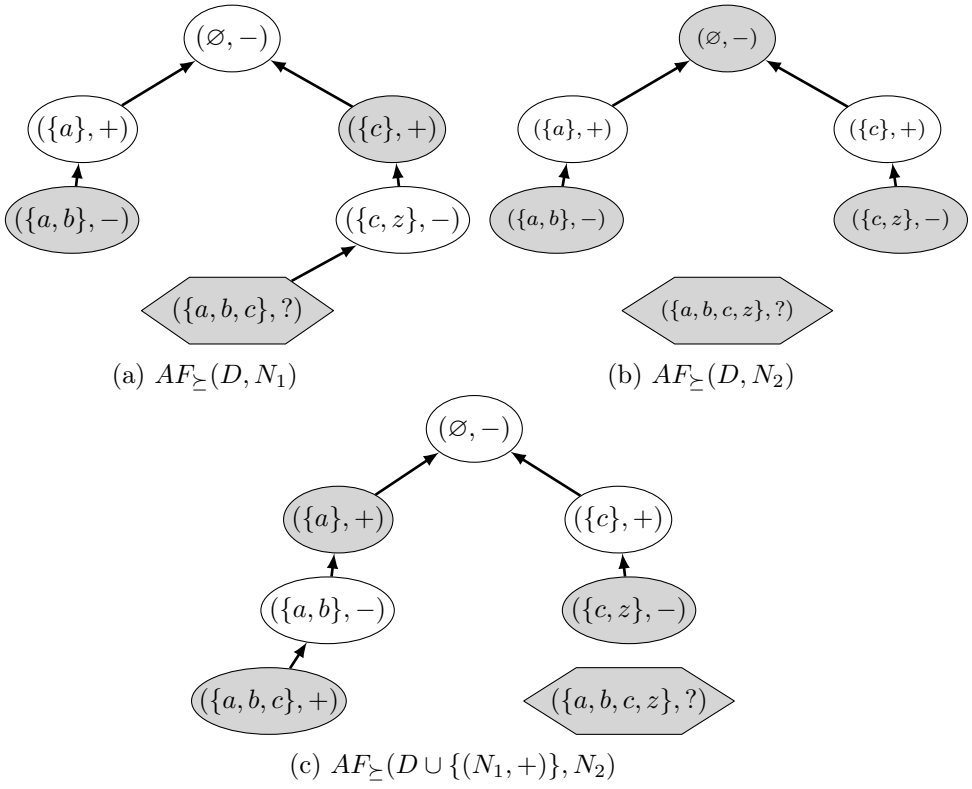


Figure 3.2: AFs for the proof of Theorem 3.10, with the grounded extension shaded.

Note that the proof of Theorem 3.10 shows that the inference relation drawn from the original $AA\text{-}CBR$ (i.e. $AA\text{-}CBR_{\supseteq}$) is also non-cautiously monotonic, given that the proof's counterexample is obtained with $AA\text{-}CBR_{\supseteq}$.

Corollary 3.11. *$AA\text{-}CBR_{\succeq}$ is not cautiously monotonic.*

The corollary is straightforward, since $AA\text{-}CBR$ is a particular case of $AA\text{-}CBR_{\succeq}$. Due to our Theorem 3.4, we can also conclude that:

Corollary 3.12. *$AA\text{-}CBR$ does not satisfy cut, cumulativity, nor rational monotonicity.*

We can also see that the counterexample amounts to an expansion of Example 3.1, as follows.

Example 3.13 (continuing from Example 3.1)**.** Assume that a different type of crime happened: public offending someone's honour, which we will call defamation (df). In one case, it was established that the defendant did publicly damage someone's honour, and was considered guilty ($(\{df\}, +)$. In a subsequent case, even if proven that the defendant did hurt someone's honour, it was established that this was done by a true allegation (td), and thus the case was dismissed, represented as $(\{df, td\}, -)$. What happens, then, if a same defendant is: 1. simultaneously proven guilty of homicide, of defamation, but shown to have committed the homicide in self-defence $((\{hm, df, sd\}, ?))$; or 2. simultaneously proven guilty of homicide, of defamation, shown to have committed the homicide in self-defence, and also shown to have committed defamation by a true allegation $((\{hm, df, sd, td\}, ?))$?

We can map these situations to our counterexample in the proof of Theorem 3.10 by setting $a = hm$, $b = sd$, $c = df$, and $z = td$. The first question is answered by the AF represented in Figure 3.2a, with outcome $+$, that is, the defendant is considered guilty. The proof of Theorem 3.10 shows that the answer to the second question in $AA\text{-}CBR_{\succeq}$ would depend on whether the case in the first question was already judged or not. If not, then the cases $(\{hm, sd\}, -)$ and $(\{df, td\}, -)$ would be the nearest cases, and the outcome would be $-$, that is, not guilty. However, if the case in the first question was already judged and incorporated into the case law, it would serve as a counterargument for $(\{hm, sd\}, -)$, and guarantee that the outcome is $+$ (guilty). Intuitively, this seems strange. This dependence on the order of cases suggests a problem with how $AA\text{-}CBR$ deals with those questions. Of course, a caveat is that our example is suggestive, since both crimes are typically unconnected, so an understanding of one should not affect the other. Still, it is strange that the case in the first question being judged first affects the answer for the case in the second question.

This example aims only to illustrate an interpretation in which $AA\text{-}CBR_{\succeq}$ operates seemingly inappropriately. Whether this behaviour is desirable in general depends on the intended application and other elements such as the relation between the characterisations and the partial order (e.g. assumptions

on interplay between features). We discuss further the issue of desirability of those properties in legal CBR in the case study (Section 3.6) and also in the discussion (Section 3.7) that follows it, regarding other legal CBR models.

3.5 A cumulative $AA\text{-}CBR_{\succeq}$

We will now present $cAA\text{-}CBR_{\succeq}$, a novel, *cumulative* incarnation of $AA\text{-}CBR$ satisfying cautious monotonicity.

Concise Dataset. Firstly, let us present some general notions, defined in terms of the $\vdash_{\mathbb{C}}$ inference relation from an arbitrary classifier $\mathbb{C}$. Intuitively, we are after a relation $\vdash'_{\mathbb{C}}$ such that if $D \vdash_{\mathbb{C}} c$ and $D \vdash_{\mathbb{C}} d$, then $D \cup \{c\} \vdash'_{\mathbb{C}} d$ (in our concrete setting, $\vdash_{\mathbb{C}} = \vdash_{AA\text{-}CBR_{\succeq}}$ and $\vdash'_{\mathbb{C}} = \vdash_{cAA\text{-}CBR_{\succeq}}$). We also want the property that, whenever D is "well-behaved" (in a sense to be made precise later), $D \vdash_{\mathbb{C}} s$ iff $D \vdash'_{\mathbb{C}} s$. In this way, given that $D \vdash'_{\mathbb{C}} c$ and $D \vdash'_{\mathbb{C}} d$, then we would conclude $D \cup \{c\} \vdash'_{\mathbb{C}} d$, making $\vdash'_{\mathbb{C}}$ a cautiously monotonic relation. We will define $\vdash'_{\mathbb{C}}$ by building a subset of the original dataset in such a way that cautious monotonicity is preserved. We start with the following notion of *includable* examples:

Definition 3.14. An example $(x, y) \in X \times Y$ is *surprising* w.r.t. D iff $D \setminus \{(x, y)\} \not\vdash_{\mathbb{C}} (x, y)$ and *sufficient* w.r.t. D iff $D \cup \{(x, y)\} \vdash_{\mathbb{C}} (x, y)$. Additionally, an example $(x, y) \in X \times Y$ is *includable* w.r.t. D iff it is both surprising and sufficient w.r.t. to D.

The definition of includable example has two parts: that the example is surprising, in the sense that, without it, the predicted outcome would be different, and that it is sufficient, in the sense that adding it makes it inferrable. While we propose the definitions above for technical reasons that will soon be explained, we can attribute a legal intuition to those concepts. A case being surprising w.r.t. a casebase means that, if it were not in the casebase, it would not be predicted as such. Legally this case is in some sense innovative, perhaps for being of first impression (that is, not bound by precedent), or for contradicting previous case law (as in the case of overrulings). As for the notion of sufficient, it means that the case being in the casebase guarantees its outcome.

This can be seen as a form of *stare decisis*, but also that the case has not been overruled by another case in the casebase. Also notice that if there is a pair of incoherent cases in a casebase, both cannot be sufficient at the same time.

We then define the notion of *concise* subsets, amounting to includable examples only:

Definition 3.15. Let $D \subseteq X \times Y$ be a dataset, $D' \subseteq D$, and let $\varphi(D') = \{(x, y) \in D \mid (x, y) \text{ is includable w.r.t. } D'\}$. Then D' is *concise w.r.t.* D whenever it is a fixed point of φ, that is, $\varphi(D') = D'$.

To illustrate in the context of *AA-CBR*, consider D from which the AF in Figure 3.2c is drawn. D is not concise w.r.t. itself, since $(\{a, b, c\}, +)$ is not includable w.r.t. D (indeed, $D \setminus \{(\{a, b, c\}, +)\} \vdash_{AA\text{-}CBR_{\geq}} (\{a, b, c\}, +)$, see Figure 3.2a). Also, $D' = D \setminus \{(\{a, b\}, -), (\{a, b, c\}, +)\}$ is not concise either (w.r.t. D), as $(\{a, b\}, -)$ is includable w.r.t. D' (the predicted outcome being $+$), but not an element of D'. The only concise subset of D here is $D'' = D \setminus \{(\{a, b, c\}, +)\}$.

Again, one can think of our motivation for considering concise subsets of a casebase as mainly technical, as we will now give a semi-formal account of how to build a cautiously monotonic system via the use of concise subsets. This account we will later formalise fully. However, one can also think in legal terms of a concise subset of a casebase as a subset of past cases in which every case is innovative, that is, brings new legal content (since the case is surprising), and also if this case were to reoccur, it would be decided in the same way (that is, the case still has value as a precedent, and has not been, e.g., overruled). At a more intuitive level, this could be seen as filtering all past cases into a subset of cases that capture current case law. However, we will not stretch this analogy further since it is not our goal here to give a formal connection between the proposed concepts and the legal value of *stare decisis*.

Let us now consider $D' \subseteq D$, for a dataset D. If D' is concise w.r.t. D, $(x, y) \in (X \times Y) \setminus D$ is an example not in D already and $D' \vdash_{\mathbb{C}} (x, y)$, then (x, y) is not includable w.r.t. D', and thus D' is still concise w.r.t. $D \cup \{(x, y)\}$. Now, *suppose that there is exactly one such concise* $D' \subseteq D$ w.r.t. D (let us refer to this subset simply as $concise(D)$). Then, it seems attractive to define $\vdash'_{\mathbb{C}}$ as: $D \vdash'_{\mathbb{C}} (x, y)$ iff $concise(D) \vdash_{\mathbb{C}} (x, y)$. Such $\vdash'_{\mathbb{C}}$ inference relation would then be cautiously monotonic if $concise(D) = concise(D \cup \{(x, y)\})$.

To see that, consider $(x', y') \in X \times Y$ such that $D \vdash'_{\mathbb{C}} (x', y')$. Then, since $concise(D) = concise(D \cup \{(x, y)\})$, for our new $\vdash'_{\mathbb{C}}$, D and $D \cup \{(x, y)\}$ would infer the exact same sentences, thus $D \cup \{(x, y)\} \vdash'_{\mathbb{C}} (x', y')$. This equality is indeed guaranteed given that $(x, y) \notin D$, thus it is not includable, and then a concise subset of D is still a concise subset of $D \cup \{(x, y)\}$ (otherwise, if $(x, y) \in D$, the equality would be trivial). Note that concision is too strong a property here: all that is needed is that a subset D' is selected such that every case in it is surprising w.r.t. D' itself. However, concision implies that as many cases are added as possible, while restricting to the ones that guarantee their outcomes.

In the remainder, we state uniqueness and give an algorithm that constructs $concise(D)$, in the case of $AA\text{-}CBR_{\succeq}$. If D is incoherent, there might be no concise subset thereof, but our method will still be useful, as we discuss later.

Uniqueness and Algorithm. We first give an important property of concise subsets in $AA\text{-}CBR_{\succeq}$:

Theorem 3.16. *For $AA\text{-}CBR_{\succeq}$, if there is a concise $D' \subseteq D$ w.r.t. D then it is unique.*

Proof. To arrive at a contradiction, let D' and D'' be distinct concise subsets of D. Let then $(x, y) \in (D' \setminus D'') \cup (D'' \setminus D')$ (the symmetric difference between D' and D'') such that (x, y) is $\preceq$-minimal in this set. It exists since the sets are different (there is at least one element in one of the sets and not in the other) and this set is finite. Then we have that:

$$\{(x', y') \in D' \mid (x', y') \prec (x, y)\} =$$
$$\{(x', y') \in D'' \mid (x', y') \prec (x, y)\},$$

otherwise (x, y) would not be minimal in the symmetrical difference.

Without loss of generality, consider $(x, y) \in D'$ (and thus $(x, y) \notin D''$). Let $\bar{y}$ be the opposite outcome to y, that is, $\bar{y} \in Y$ and $\bar{y} \neq y$. It is straightforward to check that $(x, \bar{y}) \notin D'$, otherwise D' would not be concise, since one of (x, y) and $(x, \bar{y})$ would not be includable.

Algorithm 1: *simple_add* for $AA\text{-}CBR_{\succeq}$.

Input: An $AA\text{-}CBR_{\succeq}$ framework $(Args, \rightsquigarrow)$ and a case $n = (n_C, n_o)$
Output: A new $AA\text{-}CBR_{\succeq}$ $(Args', \rightsquigarrow')$ framework

$DEF \leftarrow \{(x, y) \in AF_{\succeq}(Args, n_C) \mid$
  $(x, y) \neq (n_C, ?)$ and
  $(n_C, ?)$ defends (x, y) in $AF_{\succeq}(Args, n_C)\}$;
$Args' \leftarrow Args \cup \{n\}$;
$\rightsquigarrow' \leftarrow (\rightsquigarrow \cup \{(n, a) \mid a = (a_C, a_o), a \in DEF$, and $a_o \neq n_o\})$;
return $(Args', \rightsquigarrow')$

If we have that $(x, \bar{y}) \notin D''$, then with the equality above we conclude that

$$\{\beta \in D' \setminus \{(x, y)\} \mid \beta_C \preceq x\} = \{\beta \in D'' \setminus \{(x, y)\} \mid \beta_C \preceq x\}$$

and

$$\{\beta \in D' \cup \{(x, y)\} \mid \beta_C \preceq x\} = \{\beta \in D'' \cup \{(x, y)\} \mid \beta_C \preceq x\}.$$

Thus, by Lemma 3.9, $D' \backslash (x, y) \vdash_{AA\text{-}CBR_{\succeq}} (x, y)$ if and only if $D'' \backslash (x, y) \vdash_{AA\text{-}CBR_{\succeq}} (x, y)$. Thus (x, y) is surprising w.r.t. D if and only if it it is surprising w.r.t. D'. At the same time, also by Lemma 3.9, $D' \cup (x, y) \vdash_{AA\text{-}CBR_{\succeq}} (x, y)$ if and only if $D'' \cup (x, y) \vdash_{AA\text{-}CBR_{\succeq}} (x, y)$. This way, (x, y) is sufficient w.r.t. D if and only if it it is sufficient w.r.t. D'. Therefore, (x, y) is includable to D' if and only if it is includable to D'', which contradicts D'' being concise, since D'' would be lacking a includable case, as $(x, y) \notin D''$.

Now let us consider the situation in which $(x, \bar{y}) \in D''$. Suppose that $y = \delta_o$, that is, it is the default outcome. We can see that $D' \setminus \{(x, \bar{y})\} = D'$ and $D' \nvdash_{AA\text{-}CBR_{\succeq}} (x, \bar{y})$. Still, $D' \cup \{(x, y)\} \vdash_{AA\text{-}CBR_{\succeq}} (x, \bar{y})$, since it would create an incoherence, making the outcome be non-default. Therefore $(x, \bar{y})$ is includable to D', yet not a member of it, contradicting D' being concise. The case for $\bar{y} = \delta_o$ instead is analogous, checking that (x, y) is includable to D'' yet not a member of it. Therefore, for any possibility, the assumption of more than one concise subset of D leads to an absurdity, and thus we conclude there is at most one concise subset. □

The procedure for finding this unique $concise(D)$ (if it exists) is integrated within Algorithm 2, using in turn Algorithm 1. If $concise(D)$ does not exist, the algorithm will still return some $D' \subseteq D$ consisting only of surprising ex-

Algorithm 2: Setup/learning for $cAA\text{-}CBR_{\succeq}$.

Input: A dataset D
Output: A subset D' of D, an AF $cAA\text{-}CBR_{\succeq}(D)$

1 $unprocessed \leftarrow D$;
2 $Args \leftarrow \{(\delta_C, \delta_o)\}$;
3 $\rightsquigarrow \leftarrow \varnothing$;
4 $D' \leftarrow \varnothing$;
5 **while** $unprocessed \neq \varnothing$ **do**
6 $\quad stratum \leftarrow \{\alpha \in unprocessed \mid \alpha$ is $\preceq$-minimal in $unprocessed\}$;
7 $\quad unprocessed \leftarrow unprocessed \setminus stratum$;
8 $\quad to\_add \leftarrow \varnothing$;
9 $\quad$ **for** $\alpha \in stratum$ **do**
10 $\quad\quad (\alpha_C, \alpha_o) \leftarrow \alpha$;
11 $\quad\quad$ **if** $AA\text{-}CBR_{\succeq}(D', \alpha_C) \neq \alpha_o$ **then**
$\quad\quad\quad$ `// the outcome for` α_C `w.r.t.` $(Args, \rightsquigarrow)$ `is not` α_o
12 $\quad\quad\quad to\_add \leftarrow to\_add \cup \{\alpha\}$;
13 $\quad$ **for** $\alpha \in to\_add$ **do**
14 $\quad\quad (Args, \rightsquigarrow) \leftarrow simple\_add((Args, \rightsquigarrow), \alpha)$;
15 $\quad\quad D' \leftarrow Args \setminus \{(\delta_C, \delta_o)\}$;
16 **return** $D', (Args, \rightsquigarrow)$

amples w.r.t. D' (in fact, Algorithm 2 returns both this subset and the AF mined from it, for ease of use). The main idea behind the algorithm is simple: we start with the default argument, and progressively build the AF by adding cases from D by following the partial order $\preceq$. Before adding a past case, we test whether it is includable or not w.r.t. the dataset underpinning the current AA framework: if it is, then it is added; otherwise, it is not. More precisely, the algorithm works with strata over D, alongside $\preceq$. In the simplest setting where each stratum is a singleton, the algorithm works as follows: starting with $D_0 = \varnothing$ and the entire dataset $D = \{d_i\}_{i \in \{1,\ldots,|D|\}}$ unprocessed (lines 1-4), at each step i (loop in line 5), we obtain either $D_i = D_{i-1} \cup \{d_i\}$, if d_i is includable w.r.t. D_{i-1}, or $D_i = D_{i-1}$, otherwise (lines 9-15). Then $\hat{D} = D_{|D|} \subseteq D$ is the first output of the algorithm and the AF mined from $\hat{D}$ is the second output. Each example of the current stratum is tested for "includability" with respect to the same (current) subset D_i, and only the includable examples are added to it. However, it turns out that testing for surprise is enough for this verification, since every D_i is coherent, and thus every example is surprising

with respect to D_i. We illustrate the application of the algorithm next.

Example 3.17. Once more consider the dataset $D = \{(\{a\}, +), (\{c\}, +), (\{a, b\}, +), (\{c, z\}, +), (\{a, b, c\}, +)\}$ in Figure 3.2c, as well as the definitions used in the proof of Theorem 3.10 for X, Y, (δ_C, δ_o) and $\preceq$. Let us examine the application of Algorithm 2 to it. We start with an AA framework AF_0 consisting only of (δ_C, δ_o), that is, $D_0 = \varnothing$, $AF_0 = AF_{\succeq}(D_0) = AF_{\succeq}(\varnothing) = (\{(\varnothing, -)\}, \varnothing)$ (lines 1-4). The first stratum would be $stratum_1 = \{(\{a\}, +), (\{c\}, +)\}$ (line 6). Of course, then, we have $AA\text{-}CBR_{\succeq}(\{(\varnothing, -)\}, \{a\}) = -$, and similarly for $(\{c\}, ?)$. Thus, every argument in $stratum_1$ is includable, and is then included in the next AF, resulting in $D_1 = (\{a\}, +), (\{c\}, +)$ and $AF_1 = AF_{\succeq}(D_1)$ (lines 13-15). Now, the second stratum is $stratum_2 = \{(\{a, b\}, -), (\{c, z\}, -)\}$ (back in line 6). We can verify that $AA\text{-}CBR_{\succeq}(D_1, \{a, b\}) = +$ and $AA\text{-}CBR_{\succeq}(D_1, \{c, z\}) = +$. As a result $(\{a, b\}, -)$ and $(\{c, z\}, -)$ are both includable, and then included in the next step, that is, $D_2 = D_1 \cup \{(\{a, b\}, -), (\{c, z\}, -)\}$, and $AF_2 = AF_{\succeq}(D_2)$ (line 13). Finally, $stratum_3 = \{(\{a, b, c\}, +)\}$ (line 6). Now we verify that $(\{a, b, c\}, +)$ is *not* includable, because $AA\text{-}CBR_{\succeq}(D_2, \{a, b, c\}) = +$ (line 11). Therefore, it is *not* added to the AA framework, that is, $D_3 = D_2$ and thus $AF_3 = AF_{\succeq}(D_3) = AF_{\succeq}(D_2) = AF_2$. Now $unprocessed = \varnothing$ (breaking the loop in line 5), and the selected subset is D_3, with corresponding $AF_{\succeq}(D_3) = AF_3$, and we are done. Note that using $cAA\text{-}CBR_{\succeq}$ the counterexample in the proof of Theorem 3.10 would fail, since $(\{a, b, c\}, +)$ would not have been added to the AF.

Note that, if D is coherent, we could have defined the algorithm equivalently by looking at cases one-by-one rather than grouping them in strata. However, using strata still has the advantage of allowing for parallel testing of new cases. If D is incoherent, then using strata is necessary, otherwise the wrong case could be selected from an incoherent pair of cases.

A full complexity analysis of the algorithm is outside the scope of this work. However, note here that the algorithm refrains from building the AA framework from scratch each time a new case is considered. Still regarding Algorithm 1, note that it is easy to compute the set DEF when checking whether the next case is includable or not, thus we could optimise its implementation with the use of caching. Besides, the subset of minimal cases (that is, the

stratum) can be extracted efficiently by representing the partial order as a directed acyclic graph and traversing this graph breadth-first. Finally, the order in which the cases in the same stratum are added does not affect the outcome. Thus, each case in the same stratum can be safely tested for includability in parallel.

Definition of $cAA\text{-}CBR_{\succeq}$.

Theorem 3.18. *Let D' be the dataset returned by Algorithm 2. Then for every $\alpha \in D'$, α is surprising w.r.t. D'. Additionally, if D has a concise subset, D' is its unique concise subset. In particular, there is always a concise subset if D is coherent.*

As the proof is more involved, it is presented in Appendix A (Section A.1). We cannot generalise the existence result for any D: consider the (incoherent) counterexample when $D = \{(\{a\}, +), (\{a, b\}, -), (\{a, b\}, +)\}$, for $AA\text{-}CBR_{\supseteq}$. None of its subsets is concise. Still, our algorithm returns the subset $\{(\{a\}, +), (\{a, b\}, -)\}$, which is coherent and consists only of surprising examples.

To conclude, we can then define inference in $cAA\text{-}CBR_{\succeq}$, the classifier yielded by the strategy described until now:

Definition 3.19. Let D be a dataset and D' be the subset of D identified by Algorithm 2. Let $cAF_{\succeq}(D, N_C)$ be the AF mined from D' and $(N_C, ?)$, with default argument (δ_C, δ_o). Then, $cAA\text{-}CBR_{\succeq}(D, N_C)$ stands for the outcome for N_C, given $cAF_{\succeq}(D, N_C)$.

Thus, we directly obtain the inference relation $\vdash_{cAA\text{-}CBR_{\succeq}}$. Then, $cAA\text{-}CBR_{\succeq}$ amounts to the form of $AA\text{-}CBR$ using this inference relation. It is easy to see, using Theorem 3.16 and in line with the discussion before it, that $cAA\text{-}CBR_{\succeq}$ is cautiously monotonic:

Theorem 3.20. *$\vdash_{cAA\text{-}CBR_{\succeq}}$ is cautiously monotonic.*

Corollary 3.21. *$\vdash_{cAA\text{-}CBR_{\succeq}}$ is satisfies cut, cumulativity, and rational monotonicity.*

The corollary is straightforward from Theorem 3.4.

Incoherence. An important additional property of $cAA\text{-}CBR_{\succeq}$ is that it naturally accommodates a way to handle incoherences in the dataset. During the execution of Algorithm 2, an incoherent pair of cases would be considered at the same stratum. As every characterisation receives an outcome in a $cAA\text{-}CBR_{\succeq}$ framework, and exactly one, then if there is an incoherent pair in the dataset, one of its examples would be includable while the other would not. Therefore, only the includable example becomes an argument in the AA framework. Although an incoherent dataset may not have a concise subset, this approach finds a coherent subset which always chooses among one of the conflicting examples, using includability as the criterion for choice.[12] As an example, consider again Figure 3.1c. Following Algorithm 2, we see that in the first `while` loop both $(\{hm\}, +)$ and $(\{hm\}, -)$ are in the stratum. Since the default outcome is $-$, $(\{hm\}, +)$ is a surprising case w.r.t. $\varnothing$ and thus is added, while $(\{hm\}, -)$ is not and thus is not added, and the algorithm terminates.

Theorem 3.22. *The dataset returned by Algorithm 2 is coherent.*

Proof. Assume D is incoherent, and thus there are at least two cases with the same characterisation and different outcomes, namely, (α_C, δ_o) and $(\alpha_C, \bar{\delta_o})$. Then during the execution of Algorithm 2, (α_C, δ_o) and $(\alpha_C, \bar{\delta_o})$ occur in the same stratum, that is, they are in `stratum` in the same `while` loop. This is clear since they share the same characterisation, and (α_C, δ_o) is $\preceq$-minimal if and only if $(\alpha_C, \bar{\delta_o})$ is. At this point, there is a current AF, and the predicted outcome for α_C is either δ_o or $\bar{\delta_o}$. If the former, $(\alpha_C, \bar{\delta_o})$ is added to `to_add`, while (α_C, δ_o) is not, and if the latter, the opposite happens. Thus only one of them is indeed added to the AF, and thus the dataset underpinning the AF when Algorithm 2 terminates is coherent. □

Note that since now a coherent subset is used as basis for the inference, whenever the default case is not in the grounded extension, it will be attacked by a case which is in it.[13] Thus we have a "principled" way of dealing with incoherences, in which the includable example is always kept.

[12] Indeed, from this reasoning one can also see that *every concise subset is also coherent.*

[13] In more detail, this is so since the AF would be well-founded, and thus every argument outside the grounded extension would be attacked by it (see Dung, 1995).

Spikes. An inconvenience in $AA\text{-}CBR_{\succeq}$ is the presence of cases in the AF which do not reach the default case. While part of the AF, they do not affect whether the default case (δ_C, δ_o) is or not in the grounded extension, and thus the outcome. Formally, these cases can be defined as follows, for $AA\text{-}CBR_{\succeq}$ as well as $cAA\text{-}CBR_{\succeq}$:

Definition 3.23. Let $(Args, \rightsquigarrow) = AF_{\succeq}(D, N_C)$ or $(Args, \rightsquigarrow) = cAF_{\succeq}(D, N_C)$, and $\alpha \in D \cap Args$. Then, α is a *spike* iff there is no path in $(Args, \rightsquigarrow)$ from α to (δ_C, δ_o).

As a simple example, consider the casebase in Figure 3.2c, and add $(\{b\}, -)$ to it. It would be attacked by $(\{a, b, c\}, +)$, but it would attack no other argument. Thus, $(\{b\}, -)$ would not reach any other argument and would, then, be a spike.

Spikes are unhelpful, since their presence is entirely superfluous, that is, they can be removed with no change in outcome, for any new case.

Theorem 3.24. *Let* $(Args, \rightsquigarrow) = AF_{\succeq}(D, N_C)$ *and* $\alpha \in D \cap Args$ *be a spike. Then* $AA\text{-}CBR_{\succeq}(D \setminus \{\alpha\}, N_C) = AA\text{-}CBR_{\succeq}(D, N_C)$.

Thus, a useful step in practice is removing spikes from the AF when visualising or storing (e.g. for caching), since the AF may become significantly leaner (indeed, we do this in the case study in Section 3.6).

Instead, $cAA\text{-}CBR_{\succeq}$ shows no spikes, by construction, given that spikes are not includable, and thus are not added to $cAF_{\succeq}(D, N_C)$.

Theorem 3.25. *Let* $(Args, \rightsquigarrow) = cAF_{\succeq}(D, N_C)$. *Then, there are no spikes in* $Args$.

We prove both theorems above in Appendix A (Section A.2).

3.6 Case study

We now explore, as a case study for our $cAA\text{-}CBR_{\succeq}$ approach, the U.S. Trade Secrets domain, frequently discussed in the AI and law literature (Rissland and Ashley, 1987; Brüninghaus and Ashley, 2003; Bench-Capon, 2017). This area of law deals with misappropriation of commercially relevant information

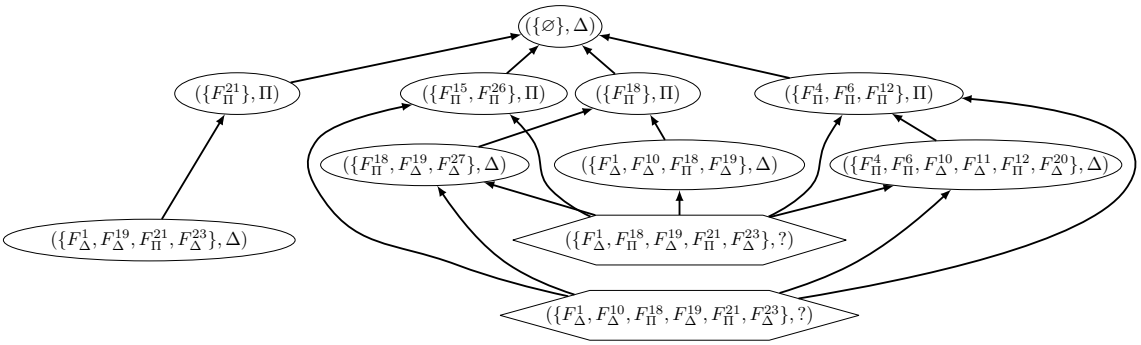


Figure 3.3: Resulting AA framework for the U.S. Trade Secrets casebase, with two new cases added. Each case in the original dataset yields possibly many arguments, each argument represented by its factors and outcomes. Some of the factors are: F^{1}_{Δ}: the plaintiff disclosed its product information in negotiations with defendant; F^{21}_{Π}: defendant obtained plaintiff's information although he knew that plaintiff's information was confidential F^{18}_{Π}: defendant's product was identical to plaintiff's. For a full list of factors with their respective descriptions, see B.1, in Appendix B.

that, allegedly, should not have been available or used by another party. The stereotypical scenario is of a company, the plaintiff, suing another, the defendant, claiming that such misappropriation happened, resulting in economic loss for the plaintiff.

In this setting, each case is represented by the factor-based analysis presented in Section 2.2.2. We will quickly recapitulate it. Cases are represented by *factors* each supporting either plaintiff or defendant, and an outcome, which may be a win for plaintiff (Π) or defendant (Δ). Formally, each such case is of the form $(F^{\Pi} \cup F^{\Delta}, o)$ where F^{Π} are the factors supporting the plaintiff, F^{Δ}, the defendant, and $o \in \{\Pi, \Delta\}$ is the case outcome. Example of pro-plaintiff factors are that the information was about a product which was unique, in the sense that only the plaintiff manufactured this product (F^{15}_{Π}), and that the defendant knew that the information was confidential (F^{21}_{Π}), while some pro-defendant factors are that the plaintiff disclosed the information in negotiations with the defendant (F^{1}_{Δ}), and that the plaintiff disclosed the information in a public forum (F^{27}_{Δ}). For this case study, we use the publicly available 32 cases considered in previous literature (Chorley and Bench-Capon, 2005; Al-Abdulkarim et al., 2015a; Al-Abdulkarim, 2017; Grabmair, 2016). The full list with their descriptions can be found in Appendix B, Table B.1.

Since factors are polarised representations, that is, they indicate a side, we

would lose information in treating them simply as features of $AA\text{-}CBR_{\supseteq}$. It is necessary to incorporate the idea that, if a case is in favour, for instance, of the plaintiff, then removing one of its pro-defendant factors should still decide the same outcome. This is the idea of a case being *constrained*, as typical in the literature of precedential constraint in AI and law (Horty and Bench-Capon, 2012; Horty, 2019; Prakken, 2021; Prakken and Ratsma, 2022b) and presented in Section 2.2.2. We accommodate this idea by changing the representation of cases. Formally, if a case is (F^{Π}, F^{Δ}, o), then it yields the following set of $AA\text{-}CBR$ cases: $\{(F^{\Pi} \cup Y, \Pi) \mid Y \subseteq F^{\Delta}\}$, if $o = \Pi$; and $\{(F^{\Delta} \cup Y, \Delta) \mid Y \subseteq F^{\Pi}\}$, if $o = \Delta$. That is, a single case becomes multiple cases w.r.t. $AA\text{-}CBR$. Even though this is not a compact representation (indeed, in the worst-case it is exponential in the number of factors), we only aim to show how $AA\text{-}CBR$ and cautious monotonicity applies in this domain, not to provide a scalable representation. This limitation might not be severe in practice since the more expensive step is learning, that is, building the AF. If most of the cases created by this multiplicity are spikes and those are removed when building the AF, a one-off task, then this may not translate into a serious problem in inference time. Alternatively, it may be possible to search for potential conflicts or interactions between cases in a way that the correctness is guaranteed without constructing the entire power set of features.[14] Still, whether those measures are sufficient to deal with the problem or whether this scalability problem is a real limitation are matters to be explored empirically and beyond our current scope.

In order to give an appropriate comparison of the resulting AF of both $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$, we remove spikes in the $AA\text{-}CBR_{\succeq}$ AF. This makes a more appropriate comparison to $cAA\text{-}CBR_{\succeq}$. It turns out that, for this casebase, the resulting AF is the same for both $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$, and shown in Figure 3.3. That is, the resulting casebase (with spikes removed)

[14] For a concrete idea of what we mean by this, assume a scenario where there is a case with many pro-plaintiff factors, a single pro-defendant factor, and the outcome is pro-defendant. In our approach, every subset of the pro-plaintiff factors would yield an extra case representation in $AA\text{-}CBR_{\succeq}$. This is the exponential complexity we are discussing. However, if none of those factors occur in other cases, then simply adding a single case containing no pro-plaintiff factor and the single pro-defendant factor would correctly capture the desired precedential constraint behaviour, assuming a fixed casebase. Whether this insight can be properly generalised is an open issue.

is already concise with respect to itself. Arguably this could be seen as suggesting that an "adequate" legal casebase is such that cautious monotonicity (and indeed, cumulativity) is not violated on it; or that an adequate representation of case law is such that the casebase is concise. However, this is too small a case study to allow for any strong conclusions. It could be, for instance, only an artefact of the manually crafted representation.

This coinciding result does not mean that using $cAA\text{-}CBR_{\succeq}$ and $AA\text{-}CBR_{\succeq}$ would be equivalent in every sense. We will now show that $AA\text{-}CBR_{\succeq}$ could be manipulated in a dynamic way due to its violation of cautious monotonicity, while $cAA\text{-}CBR_{\succeq}$ cannot. Consider the following new cases $N1_C = (\{F^1_\Delta, F^{18}_\Pi, F^{19}_\Delta, F^{21}_\Pi, F^{23}_\Delta\}, ?)$ and $N2_C = (\{F^1_\Delta, F^{10}_\Delta, F^{18}_\Pi, F^{19}_\Delta, F^{21}_\Pi, F^{23}_\Delta\}, ?)$. We can think in terms of two cases an attorney needs to argue, and would like to have a specific outcome, for instance, pro-plaintiff (Π). We will also assume that the system acts as (or models) a judge or court and respects a form of *stare decisis*, in that cases it has decided are added back to its casebase, since they must now be considered settled as well. For $N1_C$, the predicted $AA\text{-}CBR_{\succeq}$ outcome is Π. For $N2_C$, it is Δ. However, when adding $(N1_C, \Pi)$ to the casebase, the $AA\text{-}CBR_{\succeq}$ outcome of $N2_C$ then changes to Π (this can be quickly seen in Figure 3.3 by noticing that $(N1_C, \Pi)$ would then attack $(\{F^{18}_\Pi, F^{19}_\Delta, F^{27}_\Delta\}, \Delta)$). In terms of the domain, a new case, $N1_C$, which brings no different reason for change of the case law (be it distinguishing, change of social values, among others), indeed changes the system, as proved by the change in $N2_C$. This implies our attorney in consideration, with no innovation in reasons, could achieve a desired outcome in $N2_C$ by simply presenting it after $N1_C$ is judged. Thus, presenting the cases in different orders would necessarily change the results, even if no new element is introduced, such as considerations of value, policy, or change of legislation. This is not the case for $cAA\text{-}CBR_{\succeq}$. It is straightforward to check that $(N1_C, \Pi)$ is non-includable, and thus adding it to the casebase would not change the AF mined from this dataset using $cAA\text{-}CBR_{\succeq}$.

Notice this issue is not an attack only on $AA\text{-}CBR_{\succeq}$, but on any CBR system that violates cautious monotonicity. The essence of the attack is, given a CBR classifier with corresponding inference relation $\vdash$ that violates cautious monotonicity, then, by definition, there is a dataset (i.e. a casebase) D and two

cases with respective outcomes (x, y), (x', y') such that $D \vdash (x, y)$, $D \vdash (x', y')$ but (without loss of generality) $D \cup \{(x, y)\} \nvdash (x', y')$. The implication of this is that if an attorney is in a situation in which the current casebase is such that there is this pair of cases, attorney may manipulate, in the way seen above. If the outcome y' is desired, simply present x' first. Otherwise, present case x first, and the outcome is guaranteed to not be x' anymore. That is, the attack is valid.

We can also discuss an impact of the violation of cut that is particularly relevant to legal CBR. One of the original motivations of the influential system HYPO, giving its name, is reasoning with hypotheticals (Ashley, 1991; Bench-Capon, 2017). For example, HYPO employs heuristics in order to suggest possible hypotheticals that could strengthen or weaken the argument of a side.

The relation between cut and hypotheticals is that a hypothetical can be seen as a form of multi-step reasoning. In the more general scenario: if a CBR system with inference relation $\vdash$ violates cut, then there is a casebase D and two cases with respective outcomes (x, y), (x_{hyp}, y_{hyp}) such that $D \vdash (x_{hyp}, y_{hyp})$, $D \cup \{(x_{hyp}, y_{hyp})\} \vdash (x, y)$ but (without loss of generality) $D \nvdash (x, y)$. As our naming suggests, the issue here is that from the initial knowledge we can assert for some initial hypothetical x_{hyp} a (hypothetical) outcome y_{hyp}. Accepting the hypothetical with its outcome (x_{hyp}, y_{hyp}), then, would amount to adding it to the initial casebase, having thus $D \cup (x_{hyp}, y_{hyp})$. If from this addition the solution of a case is achieved $(D \cup \{(x_{hyp}, y_{hyp})\} \vdash (x, y))$, violating cut means that, without the hypothetical being confirmed, this new case x would not be given this solution y. This suggests an arbitrariness on the usage of hypotheticals: they would only imply into a solution for a new problem in case they are verified. That is, this hypothetical would only make an effect on (x, y) when it stops being a hypothetical and become an actual case that has happened and has already been judged[15].

Even worse is that, if this inference relation is complete, e.g. by following our Definition 3.3, then this actually means that x would be given a different solution, due to completeness. That is, without the hypothetical being confirmed, x would have a solution, but by its confirmation x would have

[15]The connection between hypotheticals and cut is not new, and it has raised similar considerations in the context of deontic logic, under the name of detachment, instead of cut (Greenspan, 1975).

a different solution. Indeed, this is what happens with $AA\text{-}CBR_{\succeq}$, using a simple modification of the cautious monotonicity problem above: the outcome for $N1_C$ is Π, and using $N1_C\Pi$ as a hypothetical leads $AA\text{-}CBR$ to give outcome Π to $N2_C$. However, without adding this hypothetical, outcome for $N2_C$ is Δ.

3.7 Discussion

An important element for the occurrence of incoherence in a dataset is the representation of the cases themselves. That is, an insufficiently expressive knowledge representation risks conflating otherwise distinct cases, giving rise to incoherence if they have different outcomes. We should not think this is a matter left entirely to a human user, which would model a dataset by hand. If cases are thought as being originated from previous processes, such as automatic extraction of features by a natural language processing system, as previously done with $AA\text{-}CBR$ (Cocarascu et al., 2020b), it is expected that representations could fail in this way, and thus treatment of incoherence is indeed necessary.

Our treatment of factors (features for and against) in the case study is non-scalable in general, dictated by the restrictions imposed by the structure of $AA\text{-}CBR$. Factors are subject to much research since their appearance in the work of Aleven (2003b), and making $AA\text{-}CBR$ more suitable for dealing with them is a topic left for future work, with argumentation-based treatment of them for CBR already occurring in recent research, such as in the work of Prakken and Ratsma (2022b). Another knowledge engineering element also beyond the scope of this work is background knowledge not included in the cases themselves, frequent in the legal CBR literature in the form of a (typically hand-built) domain model, enriching the factor representation (Aleven, 2003b; Ashley and Brüninghaus, 2009; Al-Abdulkarim et al., 2016; Grabmair, 2017b). On the one hand, a model that depends on formally represented background knowledge is subject to the challenge of the knowledge acquisition bottleneck, which is the challenge and high cost of extracting and representing knowledge in a format amenable to symbolic reasoning (Hitzler et al., 2020). On the other hand, being able to use background knowledge allows the reuse of high-quality

knowledge derived from, or curated by, human experts, which may be hard to extract (reliably) from data only. In the case of (regular) *AA-CBR*, it is assumed that every relevant knowledge engineering aspect is captured by the partial order, case representations, and default argument. In some scenarios, this may be sufficient for performance in an application, as is the case in the learning of the partial order that will be presented in the following Chapter 4. However, this is not the case in general, and it is a limitation of this work that we do not currently have a systematic way of incorporating background knowledge of different forms in *AA-CBR*, since knowledge engineering in general can be a challenging task. A step towards overcoming this limitation is existing work that uses *AA-CBR* for revising undesirable effects of rule-bases represented as logic programs (legal debugging), thus effectively using together CBR and rule-bases (Fungwacharakorn et al., 2021; Fungwacharakorn et al., 2022).

Notwithstanding this literature, the implications of cumulativity (and its constituent properties, cautious monotonicity and cut), or of the lack of it to legal reasoning have remained largely unexplored, particularly on CBR scenarios. We discuss in Section 3.6 an unexpected consequence of violating it, namely, manipulability of outcomes by leveraging on the order of presentation of new cases, and of violating cut, the requirement of confirmation of hypotheticals so that they can directly influence reasoning. We illustrate both violations with the case study. Of course, further exploration of the relations between CBR in law and properties of non-monotonic reasoning systems is still required and left to future work.

This does not mean those properties have not been discussed at all. Prakken (1997) and Prakken and Modgil (2018) critically discuss cautious monotonicity as well as cumulativity, but in non-monotonic reasoning more generally, not for CBR nor in a legally motivated context specifically. The discussion is based on a example in structured argumentation with default logic (respectively, with ASPIC$^+$) which violates cumulativity. This happens since, when a defeasible conclusion in the example is raised to fact, a new argument previously unavailable then makes a previous defeasible conclusion unjustified. An interesting investigation would be verifying if CBR languages modelled on top of a language such as ASPIC$^+$ (for instance Prakken et al., 2015a) would have a CBR

version of the violation of cumulativity.

In discussing casebase dynamics in the reason model of precedential constraint, Horty and Bench-Capon (2012) found out that "simply following a precedent rule can lead to a change in the law". One may be led to believe this is an affirmation that an adequate modelling of case law is not cautious monotonic. However, this is not necessarily so. They show that following a rule originated from previous cases may make a decision maker unable to distinguish a new case. That is, merely following a past rule in a case may strengthen the precedential constraint of it, but - and this is the crucial point - we can verify that it would not make a new case previously constrained to an outcome to be constrained to a different outcome. Besides, this effect is only possible if the decision maker is not constrained to an outcome in the changing case, that is, it would have been still possible to distinguish consistently, but the decision maker made the choice of following the rule.

We have not investigated the satisfaction of NMR properties on other types of classifiers (including CBR). Amgoud and Beuselinck (2022) show the incompatibility between cautious monotonicity and a principle they call strong coherence, amounting to a fuzzy rule constraining the similarity in outputs to be bounded below by the similarity in inputs. This depends on two similarity functions, at input and output level respectively, and on the idea that they in a sense correspond. This is not (at least immediately) applicable to CBR models in general, especially ones that do not use a symmetrical notion of similarity. In *AA-CBR* and in legal CBR in general, an asymmetric notion of relevance or strength is used instead. This is what captures notions of, respectively, specificity and *a fortiori* arguments. Still, it is another step to bridging the connection between CBR and NMR. As a small comment, we can show that 1-NN (that is, kNN with $k = 1$, see Section 2.2.3) is not cumulative, since it is not cautiously monotonic. For instance, in a simple example of two input points in $\mathbb{R}^n$ with different outputs, consider a third point which is a convex combination of the original two. Adding such point with its 1-NN prediction to the dataset, it can be see that the decision boundary would change, thus it is not cautiously monotonic. Exploring whether this argument can be generalised to *kNN* is, as far as we know, an open question.

3.8 Summary

We have proposed and studied regular $AA\text{-}CBR_{\succeq}$ frameworks, and proposed a new form of $AA\text{-}CBR$, denoted $cAA\text{-}CBR_{\succeq}$, which is cautiously monotonic and, as a by-product, cumulative and rationally monotonic. We also show that it results in a principled way of dealing with incoherence in casebases, something which $AA\text{-}CBR_{\succeq}$ lacks. Given that $AA\text{-}CBR_{\succeq}$ admits the original $AA\text{-}CBR_{\supseteq}$ (Čyras et al., 2016a) as an instance, we have (implicitly) also defined a cautiously monotonic version thereof.

(Some incarnations of) $AA\text{-}CBR$ have been shown successful empirically in a number of settings (Cocarascu et al., 2020b). The formal properties we have considered in this chapter do not necessarily imply better empirical results at the tasks in which $AA\text{-}CBR$ has been applied. In the next chapter, we will carry out an empirical comparison between $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$. Other issues open for future work but unexplored in this thesis are comparisons w.r.t. learnability (such as what can be said in general regarding model performance in the presence of noise; or about stability (Leen et al., 2001)), as well as a full complexity analysis of the new model. Also, we conjecture that the reduced size of the AF our method generates could possibly have advantages in terms of time and space complexity: we leave investigation of this issue to future work.

Chapter 4

Learning Relevance in Case-Based Reasoning

Important aspects for any concrete implementation of $AA\text{-}CBR$ are the representation of the data (to which we refer as characterisation, following Cocarascu et al. (2020b); see Section 2.2.1) and, crucially, the partial order $\preceq$ on the set of characterisations, which until now we have assumed given. This partial order captures the idea of relevance of past cases for a new case, as well as of specificity among past cases, in the case of regular$AA\text{-}CBR$ (Definition 3.5).

In this chapter, we will discuss ways of defining this partial order. In particular, we are interested in learning the partial order from data. This would allow one to deploy $AA\text{-}CBR$ to a wider scope of problems, instead of requiring a hand-constructed order. This carries three desiderata: 1) that the learned partial order results in a good performance; 2) that it is human interpretable, at least to a point; 3) that the methodology is as broadly applicable as possible. We here present a method for learning a partial order from tabular data using features extracted from decision trees trained on this same data (Section 4.2). We also present preliminary results on a novel approach using neural networks, in particular, autoencoders, in order to extract the partial order (Section 4.3). We show empirical results, what also allows for a comparison between $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$, our new variant (Section 4.4). The experiments are carried out on the the COMPAS dataset (Larson et al., 2016), a dataset on prediction of recidivism, that is, on prediction of whether a criminal will re-offend (Dieterich et al., 2016).

4.1 Motivation

Using *AA-CBR* as an approach requires defining a characterisation of the data and a partial order $\preceq$ on the set of characterisations. The intuition of this partial order is that when $\alpha \preceq \beta$, then β is in a sense exceptional to, or more informative than, α. The other side of this coin is that α matters to deciding about β. When two cases are not comparable in this order (that is, neither $\alpha \preceq \beta$ nor $\beta \preceq \alpha$), this means that one cannot be used for deciding the other, since they are about different issues, broadly speaking. A special characterisation is the default characterisation. Intuitively, if the default argument is rejected in deciding a case, it should not because it is irrelevant (after all, it is the default, what happens in the lack of any information), but because an exception applies that in a sense "overrules" the default. Following those intuitions is the motivation for our definition of regular *AA-CBR*$_{\succeq}$ (Definition 3.5), which connects the notion of relevance to a new case with the partial order, and ensures the default is the minimum case. The default case characterisation being the minimum in the set of characterisations imply it is relevant for every possible case.

Now the matter is which characterisation and, more importantly, which partial order should be used.[1] The original definition of *AA-CBR* is *AA-CBR*$_{\supseteq}$, in which characterisations are sets of binary features and the partial order is the subset relation (i.e. $\preceq$ is $\subseteq$, as presented in Section 2.2.1). This means that the typical situation is having no features, and each feature is a new exception. Other previous approaches have been selecting the most relevant of binary features to use in *AA-CBR* (Cocarascu et al., 2018), and, for the task of sentiment analysis in text, a characterisation based on weighted words in a sentence and a partial order defined as a greater total weight and having more features (Cocarascu et al., 2020b). We will present two novel approaches here to obtain partial orders systematically from data.

[1] We say the partial order is more important since the only aspects of the characterisation used in *AA-CBR* are equality and the partial order relation, thus there is no relevant difference between two sets of characterisations which are isomorphic with respect to the partial order.

4.2 Partial order via decision trees

Decision trees are a classic machine learning method, widely used and considered interpretable (Breiman et al., 1984; Molnar, 2022). We have reviewed them in Section 2.3.1. We will here present how we used decision trees to extract characterisations for *AA-CBR*.

Decision trees work in tabular data, regardless of whether on discrete or continuous features. In the algorithm for learning decision trees, CART, decision nodes are greedily created by choosing the feature and the split threshold which minimises some loss function (details in Section 2.3.1). Those splits result in a binary divide between examples which are below the split threshold and the ones above. Each split can then be seen as a binary feature. The approach followed is to characterise an example simply as the set of those features, that is, the set of split rules for which the example is below the threshold. With those binary features, the application of *AA-CBR* to sets, the original definition of *AA-CBR*$_{\supseteq}$ (presented in Section 2.2.1) can be applied.

We here train decision trees with pre-pruning, that is, by limiting maximum depth and maximum number of leaf nodes for the decision tree. We used the following set of values: for maximum depth, varying from 3 to 13, in a step of 2; for maximum number of leaf nodes, from 4 to 512, in geometric progression of ratio 2. Nodes are greedily created in a best-first search fashion.

Example 4.1. We now consider a simplified dataset on the task of prediction recidivism. Individuals here are described by their age and number of priors, and the goal is predicting whether or not they will re-offend within a period of time.[2]. Consider a dataset containing the following examples:

$$\alpha = ((\text{age} = 20, \text{prior\_count} = 2), \texttt{recid}),$$
$$\beta = ((\text{age} = 30, \text{prior\_count} = 1), \texttt{no\_recid}),$$
$$\gamma = ((\text{age} = 35, \text{prior\_count} = 7), \texttt{recid}),$$
$$\epsilon = ((\text{age} = 19, \text{prior\_count} = 1), \texttt{no\_recid}),$$
$$\eta = ((\text{age} = 19, \text{prior\_count} = 10), \texttt{recid})$$

[2]This is a very simplified example of the prediction of recidivism to be properly described on Section 4.4.1

Assume as well the decision tree on the left of Figure 4.1 was trained on this dataset. On the right is $AF_{\succeq}(D)$, the AF mined from the dataset, by using the partial order derived from the decision tree nodes. Each example in the training set is represented by the split tests for which it is evaluated `true`. Then they are used as a casebase for *AA-CBR*. Notice how there is no bijection between leaves in the decision tree and cases in the AF. The correspondence is one to many, since examples falling into the same leaf may correspond to different cases in the AF. For example, see how the rightmost leaf corresponds to $(\{\texttt{age\_<=\_21}\}, +)$ (via case η), to $(\{\texttt{age\_<=\_21}, \texttt{prior\_count\_<=\_3}\}, +)$ (via case α), and finally to $(\{\texttt{age\_<=\_21}, \texttt{prior\_count\_<=\_3}\}, -)$ (via case ϵ). The default outcome is $+$, reflecting that the majority of the training set has the output `recid`.

Still in Figure 4.1, we can see that $(\{\texttt{age\_<=\_21}\}, +)$ is a spike (Definition 3.23). We can also see that the pair $(\{\texttt{age\_<=\_21}, \texttt{prior\_count\_<=\_3}\}, -)$ and $(\{\texttt{age\_<=\_21}, \texttt{prior\_count\_<=\_3}\}, +)$ is incoherent. In our experiments, spikes are removed in $AA\text{-}CBR_{\succeq}$ in order to make it more comparable with $cAA\text{-}CBR_{\succeq}$, which eliminates spikes by construction. This is done by removing every argument (i.e. case) to which there is no path (in graph-theoretical sense) to the default argument.

4.3 Partial order via autoencoders

In this section, we present an approach on how to learn a partial order from autoencoders, in order to overcome the following restriction of the previous method: The decision tree approach essentially works by a form of *binning*, that is, continuous features are discretised by the split nodes, and then at the case level all the work is done at those learned splits. Even if the binning is learned according to the data, the partial order is defined as subsets according to binary features.

As discussed in Section 2.3.2, autoencoders are a type of neural network that aims on reconstructing an input. One of their main goals is learning a more convenient data representation for downstream tasks, that in some sense captures the underlying data distribution better. We here train an autoencoder and use the hidden encoding to define a partial order. Ghosh et al. (2020)

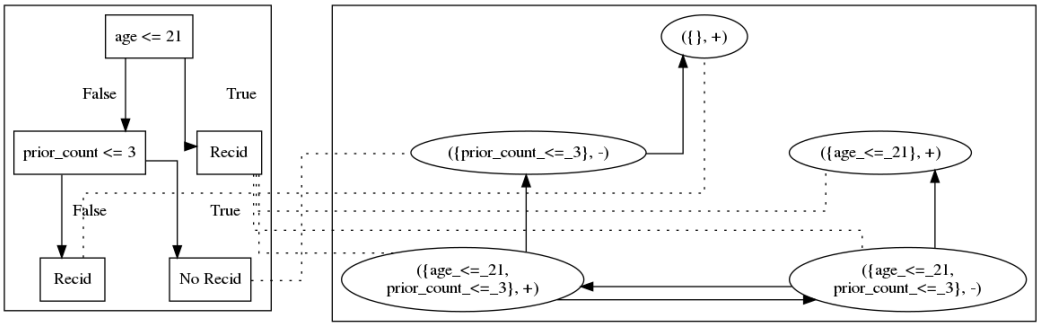


Figure 4.1: On the left, decision tree used as basis for learning of a partial order for $AF_{\succeq}(D)$. On the right, AF mined from a dataset. Dotted line indicates that a case falling in a leaf node in the decision tree corresponds to a case in the AF on the right. This correspondence is not direct, that is, this is not a transformation from the decision tree into an AF. It is intermediated by the cases in the dataset of Example 4.1. Notice that the default is +, reflecting the majority of cases in the training set having this outcome, but also there is no case $(\{\}, -)$ since the only cases with no features has the `recid` outcome. When multiple cases would have the same characterisation but different outcomes, either an incoherence is generated, or it is avoided via preprocessing. Here we illustrate an incoherence occurring.

show that properly regularised deterministic autoencoders (which from now on we will refer to simply as regularised autoencoders) can enforce a prior distribution of latent variables and generate data effectively, being thus a viable alternative to the more complex variational autoencoders. Due to this and to their simplicity, we use regularised autoencoders for producing our hidden encodings.

We will now propose a method to derive a partial order for regular $AA\text{-}CBR_{\succeq}$ from hidden encodings, in three main steps:

1. We define a partial order on a single dimension, using the intuition that a dimension corresponds to a probability distribution with a single peak (also called a unimodal distribution, with the peak itself being called the mode) (Cramer, 1946, p 179, Casella and Berger, 2002, p. 79).

2. We lift a set of partial orders to the product set, that is, from a single dimension to multiple dimensions.

3. We show how the two previous steps can be applied on the hidden encoding generated from an autoencoder.

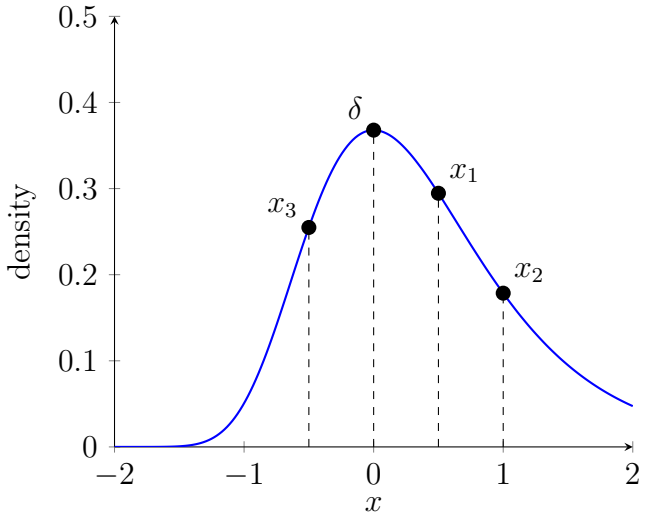


Figure 4.2: Illustration of the partial order idea for a single dimension in an example unimodal probability distribution. Here, δ is the reference point, and thus minimum point in the order, x_3 is incomparable with x_1 and x_2, while we have $x_1 \prec x_2$.

For our step 1, the question is, assuming data is described by a single dimension (that is, in the real numbers $\mathbb{R}$), what is a reasonable partial order? Different ones are possible. For instance, a straightforward partial order is simply using the usual order $\leq$ in $\mathbb{R}$. Due to our conditions for regularity, this would require the characterisation of the default argument, δ_C, to be $-\infty$ (thus extending $\mathbb{R}$ with this element). Instead of following this direction, we opt for a more probabilistic interpretation. We will use the assumption that every dimension has an underlying unimodal probability distribution. Many important probability distribution are unimodal, such as Gaussian, exponential, log-normal, Cauchy, Bernoulli, binomial, among many others (Cramer, 1946; Keilson and Gerber, 1971). Here, the mode will be thought of as a reference point, that is, as the default in that dimension.

Concretely, let $X_i = \mathbb{R}$ be this single dimension i, and $\delta^i \in X_i$ be a reference point. We can define the following partial order in X_i: for $x_1^i, x_2^i \in X_i$,

$$x_1^i \preceq_i x_2^i \text{ iff } \begin{cases} \delta^i \leq x_1^i \leq x_2^i & \text{or} \\ x_2^i \leq x_1^i \leq \delta^i \end{cases} \tag{4.1}$$

The intuition for this definition is that, in the first case, x_1^i and x_2^i are both

greater than the reference δ^i, but x_1^i is still closer to δ^i. Thus, in a sense x_2^i is "more specific": x_1^i is an exception to the reference by being greater than it, but x_2^i is even more exceptional for being even greater. The second case is analogous, but in terms of being less than the reference. In case one of the values is greater than the reference δ^i and the other one smaller than it, they are incomparable, that is, neither $x_1^i \preceq x_2^i$ nor $x_2^i \preceq x_1^i$. This can be verified to be a partial order, because from each direction of the reference it is a (total) order, with disjoint domain, and so this is the disjoint union of two orders (Davey and Priestley, 2002, p. 17). We illustrate this order in Figure 4.2.

Now we can go to step 2. Formally, assume we have a collection of sets $\{X_i\}_{i \leq n, i \in \mathbb{N}}$, for some $n \in \mathbb{N}$. Assume as well each X_i is equipped with a partial order $\preceq_i$. Let $\preceq$ be defined on the Cartesian product $X_0 \times \cdots \times X_n$ as:

$$(x_1^0, \ldots, x_1^n) \preceq (x_2^0, \ldots, x_2^n) \text{ iff } \forall i : x_1^i \preceq_i x_2^i \tag{4.2}$$

It can be shown that $\preceq$ is also a partial order on $X_0 \times \cdots \times X_n$: this is called the *product* of the partial orders (Davey and Priestley, 2002, p. 18). Returning to the probabilistic interpretation that motivated step 1, this partial order might not be entirely justified, depending on what the joint probability distribution of the variables looks like. Intuitively, using this partial order corresponds to a form of independence assumption between variables. Since our goal is simply to have an order, and independence will indeed be enforced in the next step, this is not a problem for us.

So we can now go to our final step 3. As mentioned above and discussed in Section 2.3.2, variational autoencoders can capture a data distribution enforcing a prior distribution in the hidden encodings (i.e. over the latent space). In order to do this with regularised autoencoders, it is necessary to use *ex-post density estimation* (Ghosh et al., 2020). This means that a new probability distribution is estimated on the set of encoded inputs. Using ex-post density estimation is necessary since training of regularised autoencoders does not guarantee that the latent space has a specific distribution, even if regularisation creates a bias for the autoencoder to have this prior on the hidden encoding. This technique was proposed by Ghosh et al. (2020) for generating new samples from regularised autoencoders, in a way similar to what is usu-

ally done with variational autoencoders. Our goal with the ex-post density estimation is different, since we do not aim on generating examples. We apply ex-post density estimation in order to transform the hidden encoding into a new representation that allows the application of our steps 1 and 2 above.

More specifically, we will use ex-post density estimation to estimate a multivariate Gaussian with independent variables (coordinates), each with the same variance (called an isotropic Gaussian) (Gut, 1995). We do so for the following reasons:

1. In an isotropic Gaussian, every coordinate is itself a univariate Gaussian with mean 0 (Gut, 1995). This allows the usage of our step 1 above, and in this way for each coordinate the reference value will be 0.

2. When using the definition of the lifted order in assertion 4.2 above, we will have a partial order on $\mathbb{R}^n$ such that the zero vector (also called the *origin*) is the minimum point.

3. This is a distribution that assumes that the (latent) variables are independent.

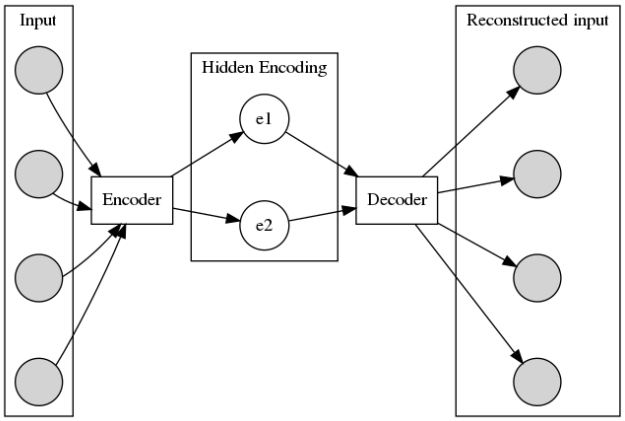
Reason 3 does not imply that latent variables are actually independent in the data (that is, in the underlying process in the world that generated the data), but the estimated model is such that the new variables are assumed independent. Intuitively, it is a "best guess" for independent variables.

In practice, we simply fit a multivariate Gaussian of the same dimension as the hidden encodings, and then we standardise the encoding values, subtracting the mean and multiplying by the inverse of the square root of the covariance matrix[3]. Formally, if X is a multivariate Gaussian distribution with mean μ and covariance matrix Σ, then there is a positive semi-definite matrix A such that $A = \Sigma^{1/2}$ and

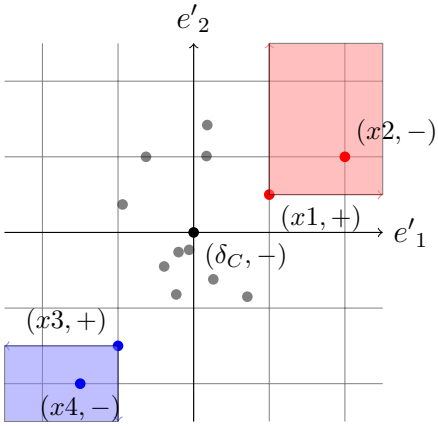
$$Z = A^{-1}(X - \mu) \tag{4.3}$$

[3]Notice this fails if the correlation matrix somehow degenerates, i.e. does not result in a positive definite matrix. In case it is positive semi-definite, this means the dimension will be reduced, effectively removing dimensions which correspond to an eigenvalue 0. This is not a problem for us unless the entire correlation matrix collapses to zero, since our goal is to learn a distribution, even of fewer dimensions. In the worst case, that of collapse, training can be reinitialised from scratch, since initialisation and optimisation are in part random.

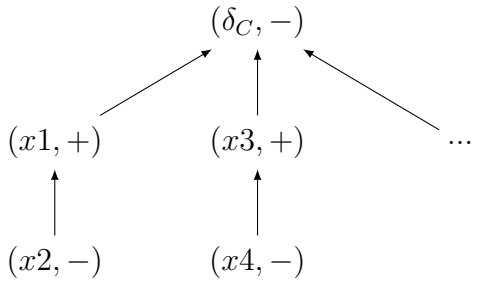
where Z is a multivariate Gaussian with mean 0 and the identity matrix as the covariance (Gut, 1995, p. 125). The formula above allows us to easily convert the encodings to a multivariate Gaussian, justifying our choice of the 0 vector as the default. We illustrate the workflow in Figure 4.3.



(a) Illustration of the autoencoder architecture. Each encoder and decoder is a feedforward neural network. We illustrate the number of dimensions in the hidden encoding as having two layers, but in experiments it is a hyperparameter.



(b) Hidden encodings in $\mathbb{R}^2$. Here we highlight an arbitrary point $x1$ (in red) and colour the region corresponding to the set of all points more specific than it, that is, of all x such that $x1 \preceq x$. Notice $x2$ is one such point. Analogously, we also show $x3$ (in blue) and colour the set of all points more specific than it. Again, $x4$ is one of those points. Notice the characterisation of the default case is in the origin. Axes are labelled as e' since we exemplify here the plane after transformation by ex-post density estimation.



(c) AF mined from the dataset, using the hidden encodings as characterisations and the corresponding partial order.

Figure 4.3: Illustration of the workflow for learning the partial order using autoencoders. First, an autoencoder is trained in the dataset (Figure 4.3a). Then, the hidden encoding is extracted an the partial order is defined over it (Figure 4.3b). Finally, with this partial order and the outcomes, we can build the AF mined from the dataset (Figure 4.3c).

4.4 Experimental results

4.4.1 Dataset

COMPAS (standing for Correctional Offender Management Profiling for Alternative Sanctions) is a proprietary prediction model for recidivism, that is developed by Northpointe Inc and widely used in the U.S. (Northpointe Inc., 2019). Its goal is evaluating the risk of recidivism, that is, of a criminal re-offending. COMPAS has been the focus of much work on algorithmic biases and (un)fairness, and on the impact of technology in the justice system since the analysis carried out by ProPublica (Larson et al., 2016). Along with this analysis, ProPublica released a dataset containing predicted scores and data of actual recidivism, and this dataset is usually referred to as the COMPAS dataset (e.g. Bao et al., 2021). We use this dataset in this chapter. Our goal is not evaluating the COMPAS algorithm or discuss its fairness(Larson et al., 2016; Barenstein, 2019; Rudin et al., 2020; Bao et al., 2021), but simply using this dataset as a way of evaluating our methodologies on learning a partial order in a legally relevant scenario. Possible errors in methodology by ProPublica may be at play (Barenstein, 2019; Bao et al., 2021) and some researchers have re-done the data processing of raw data from scratch with different conclusions (Rudin et al., 2020), but we follow much of the literature and use the dataset as is (Angelino et al., 2017; Woerkom et al., 2022). Our goal is evaluating our algorithms in a task originated from a real-world dataset related to the legal domain, but this should not be seen as results of a ready-to-deploy system or allowing clear conclusions from a criminal justice point of view (Bao et al., 2021, see discussions on bias in datasets, particularly Section 8). Indeed, the dataset itself has been criticised for being biased. For instance, a major issue is that the COMPAS dataset measures re-offence by re-arrest. This can be a very biased metric, since if policing itself is biased, then some re-offences may not result in re-arrest and some re-arrest may not actually correspond to re-offences (Bao et al., 2021). Therefore, while the dataset can be useful for evaluating model learning capacity, performance on the COMPAS dataset is not to be understood as benchmarks or fairness, as allowing broad conclusions on criminal justice, nor should results be presented as grounds for deployment without more extensive and interdisciplinary evaluations (Bao et al., 2021).

We use the two-year recidivism dataset made available by ProPublica (Larson et al., 2016). This is a tabular dataset containing 7214 entries, each corresponding to a defendant who was screened by COMPAS before trial, and given a COMPAS score on risk of recidivism. For each person, the dataset contains personal information (such as age, gender, and race), information about the current charge (such as degree of seriousness of the charge and days in which the defendant was imprisoned); criminal history and finally, whether the defendant has reoffended, and details of this new offence. In the COMPAS dataset, recidivism is defined as having been *charged* with a new crime after the COMPAS screening. Fabris et al. (2022) present more detailed documentation for COMPAS, among other datasets, in an effort to unify documentation efforts on algorithmic fairness datasets.

We apply the same filtering strategy for missing data as ProPublica[4] (Larson et al., 2016), resulting in 6172 entries. We show distributions of features of interest in Table 4.1. The original dataset also contains a feature categorising age called `age_cat`, which we omit from the table, since we already summarise the feature for age. In our trainings we use 4 different feature sets:

(A) Containing all features.

(B) Removing `age_cat`, since age is already a feature there.

(C) Removing `age_cat` and race.

(D) Removing `age_cat`, race and gender.

The reasons for those choices are as follows. We experiment removing `age_cat` for redundancy with the age feature. We also experiment removing race and gender since they are protected features in the UK (*Equality Act* 2010). Thus removing them is enforcing a kind of fairness through unawareness (Grgic-Hlaca et al., 2016; Kusner et al., 2017). This is not the only way of enforcing fairness and there are competing definitions of it (Kusner et al., 2017), but fairness through unawareness is at least a baseline understanding of fairness, and satisfies the idea in equality legislation of not discriminating individuals based on protected characteristics. Discussing what are better

[4]Larson et al. (2016) have made their code for analysis available at https://github.com/propublica/compas-analysis/blob/master/Compas%20Analysis.ipynb.

definitions of fairness for this application is the subject of other literature (Grgic-Hlaca et al., 2016; Kusner et al., 2017) and beyond our scope. In the remainder, when we refer to the dataset not specifying a feature set, we mean feature set A.

Table 4.1: Category distributions in the COMPAS dataset.

Feature	Value	Frequency (%)
Gender	Women	19.04
	Men	80.96
Race	African-american	51.4
	Caucasian	34.0
	Hispanic	8.2
	Asian	0.5
	Native American	0.1
	Other	5.5
Charge degree	Felony	64.4
	Misdemeanour	35.6
Age	18-19	0.4
	20-29	43.0
	30-39	27.9
	40-49	14.5
	50-59	10.8
	60-69	2.8
	$\geq$70	0.5
Juvenile misdemeanour count	0	94
	1	4
	$\geq$2	2
Juvenile felony count	0	97
	1	2
	$\geq$2	1
Juvenile other convictions count	0	93
	1	5
	$\geq$2	2
Prior crimes count	0	34
	1	18
	2	11
	3	8
	$\geq$4	29
Length of stay in jail for charge	0	36
	1	26
	2	7
	3	3
	$\geq$4	29
Recidivism in two years	No	54.49
	Yes	45.51

4.4.2 Results for partial order via decision trees

We now evaluate the decision tree approach (Section 4.2). We measure performance in prediction of recidivism by accuracy, and evaluate the following aspects:

(i) The impact of incoherence, and of strategies to avoid it.

(ii) The impact of hyperparameters.

(iii) How $cAA\text{-}CBR_{\succeq}$ compares to $AA\text{-}CBR_{\succeq}$, in both performance and in the resulting AFs.

(iv) How $AA\text{-}CBR$ with this partial order compares to decision trees by themselves, with respect to predictive performance.

A first issue in this approach is incoherence. Since in a decision tree multiple training examples can correspond to a leaf, it is expected that many examples, possibly with different outputs, will correspond to the same characterisation. The correspondence is not one to one from leaves to cases, since a training example will be mapped to a characterisation depending on *all* split tests, not only those on the path from the root to its leaf. Still, in properly regularised decision trees it is expected that trees will be small enough for this to be an issue.

We apply three approaches for this problem:

1. `keep`: to keep the incoherence and letting each model deal with this in their own ways

2. `removal`: to remove every incoherent pair of cases

3. `majority`: for each characterisation in the resulting transformation, count the number of training examples corresponding to each output, and select the majority output as the outcome for the (now unique) case.

We show the result of each approach in Tables 4.2 and 4.3. For this experiment we also allow the value of maximum number of leaf nodes of 2048, but later exclude it due to running time. We report results on the test set. `Keep` is a weaker strategy than the other two even for $cAA\text{-}CBR_{\succeq}$, which deals with

incoherence directly. Using `majority` seems to be the stronger choice. Since `keep` also results in slower performance, due to the many incoherent pairs, for the following experiments we will restrict ourselves to `removal` and `majority`. We highlight that although here reported in the test set, initial inspection on a validation set quickly shows how `keep` is a worse strategy. Additionally, Table 4.3 shows how even over almost all hyperparameter choices `keep` is dominated by the other strategies.

Table 4.2: Percentage accuracy of each *AA-CBR* model and each strategy for incoherence in the casebase, aggregated over hyperparameter choice (maximum depth and maximum number of leaves in the decision tree). Results on test set, feature set A.

model	strategy	min	max	mean	std dev
$AA\text{-}CBR_{\succeq}$	keep	45.57	55.35	49.13	4.26
	removal	54.43	63.93	57.27	2.23
	majority	58.21	68.09	64.12	3.95
$cAA\text{-}CBR_{\succeq}$	keep	46.98	57.83	52.32	3.04
	removal	54.43	61.88	57.42	2.40
	majority	57.45	68.09	63.94	4.20

Table 4.3: Difference in percentage accuracy between the `removal` or `keep` strategies and the `keep` strategy for incoherence, by *AA-CBR* model, aggregated over hyperparameter choice (maximum depth and maximum number of leaves in the decision tree). Aggregation is performed over the difference. Results on test set, feature set A.

model	strategy	min	max	mean	std dev
$AA\text{-}CBR_{\succeq}$	removal - keep	2.05	13.01	8.14	4.06
	majority - keep	2.92	22.52	14.99	8.15
$cAA\text{-}CBR_{\succeq}$	removal - keep	-0.11	10.42	5.10	3.53
	majority - keep	1.84	21.00	11.62	7.10

We now directly compare performance for each of the three models: $AA\text{-}CBR_{\succeq}$, $cAA\text{-}CBR_{\succeq}$, and decision trees. We do 5-fold cross validation. In each fold, hyperparameters are selected in a inner validation step (using a single split for validation set), retrained on the entire training set of that fold, and evaluated

on the test set of the fold. This is done for each model, and we report results in Table 4.4, aggregated by fold. Comparing the different feature sets, we can also see that, while the loss of mean accuracy for decision trees is small, for both $AA\text{-}CBR_{\succeq}$ and $AA\text{-}CBR_{\succeq}$ there is no difference, except by removing the gender, on feature set D, on which there is a very small increase in the mean accuracy but also in standard deviation. The variations are too small to warrant particular conclusions, but the changes imply that the protected features are indeed used by the models unless they are removed. As discussed before, the discrimination of individuals based on those features is forbidden in many countries (including the UK, via the *Equality Act* (2010)), and removing those features is at least a baseline method of enforcing non-discrimination. For a deployed system, it is crucial to evaluate in more depth the risk of discrimination not only by removing features, but evaluating feature correlation with protected features, and considering the appropriate notion of fairness for the application (Grgic-Hlaca et al., 2016; Kusner et al., 2017).

An interesting result in this experiment is that in most cases the optimal hyperparameter choice for each of $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$ resulted in both having the same classification in the test set. This suggests they may have the same decision function (that is, correspond to the same function from inputs to outputs), even if they have different inner structure. We show an example of this in Figures 4.4 and 4.5. This is an example extracted from a model trained on the dataset (feature set A). Notice how the case missing in $cAF_{\succeq}(D)$ reduces the maximum depth of the AF. One can also notice that both AFs correspond to the same decision function.

Compared to decision trees, the $AA\text{-}CBR$ approaches have a slightly lower performance, but are still comparable, being within one standard deviation from its mean.

Comparing in terms of the resulting AFs over different hyperparameter choices, $cAA\text{-}CBR_{\succeq}$ produces fewer arguments, around 60% of the number of arguments of $AA\text{-}CBR_{\succeq}$. This is very dependent on the specific configuration. The number of attacks is likewise reduced, from 40% to 60% even using the `majority` strategy. Regarding depth, an interesting effect happens regarding size: when hyperparameters that result in many cases are used (larger maximum depth and larger maximum number of leaves), $cAA\text{-}CBR_{\succeq}$ tends to

Table 4.4: Performance measured in percentage accuracy, for each approach, using 5-fold cross validation and hyperparameter search by internal validation split. Reported by feature set.

	Feature set A		Feature set B		Feature set C		Feature set D	
	mean	std dev	mean	std dev	mean	std dev	mean	std dev
Decision tree	67.60	1.31	67.60	1.31	67.48	1.56	67.00	1.15
$AA\text{-}CBR_{\succeq}$	66.32	1.20	66.32	1.20	66.32	1.20	66.41	1.31
$cAA\text{-}CBR_{\succeq}$	66.32	1.20	66.32	1.20	66.32	1.20	66.41	1.31

produce AFs that are not as deep. However, for hyperparameters that result in fewer cases, and indeed in the optimal hyperparameter choices chosen over validation (resulting in Table 4.4), $cAA\text{-}CBR_{\succeq}$ tends to show deeper AFs, since the decision function between both approaches results in the same, and the removed nodes act merely as "shortcuts". Evaluating how this affects interpretability of those models is beyond our current scope and left for future work.

In the appendix we provide more complete tables showing comparing the values for accuracy (Table C.1), number of nodes (Table C.2), number of attacks (Table C.3), maximum depth of the AF (Table C.4), and average depth of the AF (Table C.5). An interesting result that can be seen is how, even using the `removal` or `majority` strategies, for AFs with more than two nodes $cAA\text{-}CBR_{\succeq}$ has 50 to 60% of the number of nodes in $AA\text{-}CBR_{\succeq}$ (Table C.2). As for attacks, $cAA\text{-}CBR_{\succeq}$ has only 43 to 61% of the number for $AA\text{-}CBR_{\succeq}$, and even less (as low as 28 or 33%) for the `keep` strategy. In summary, $cAA\text{-}CBR_{\succeq}$ results in considerably smaller AFs, and the effect is more noticeable in larger AFs. This is an advantage for $cAA\text{-}CBR_{\succeq}$ since at smaller cost of memory space, it can achieve the same performance as $AA\text{-}CBR_{\succeq}$. A formal analysis of space complexity is outside our scope here.

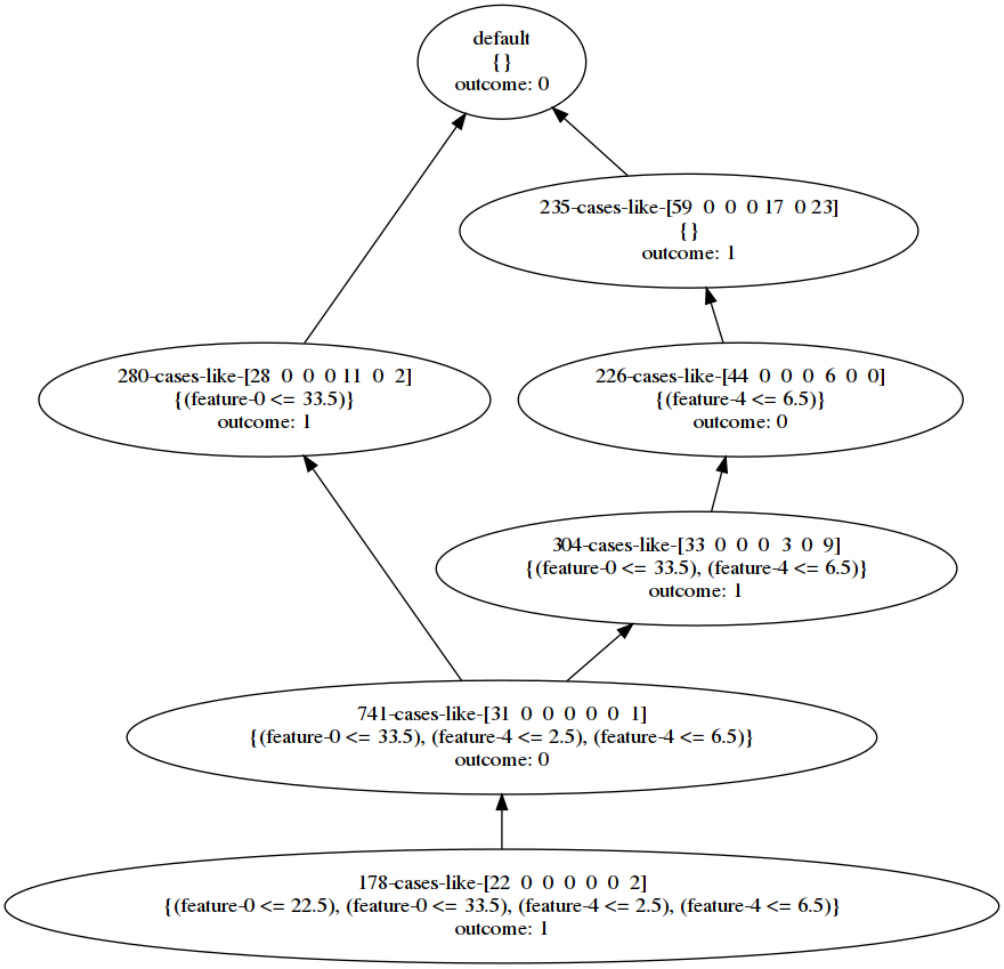


Figure 4.4: $AF_{\geq}(D)$, the AF mined using $AA\text{-}CBR_{\geq}$ from the training set (single split) using feature set D and the partial order from decision trees, using maximum depth 3 and maximum number of leaf nodes 8, removing spikes, and using the `majority` strategy for incoherence. Each non-default case is indicated as:

N-cases-like-`example`
`charac`
outcome: O

where N is how many examples in the training set correspond to that case (that is, have the same characterisation in terms of the learned features), `example` is the (original) feature values of one such example in order to represent the case,`charac` is the characterisation of the case, in terms of sets of features, each corresponding to a split, and O is the outcome of that case.
Features are as follows: feature0 = age, feature1 = juvenile felony count, feature2 = juvenile misdemeanour count, feature3 = juvenile other count, feature4 = priors crimes count, feature5 = charge degree, feature6 = length of stay in jail.

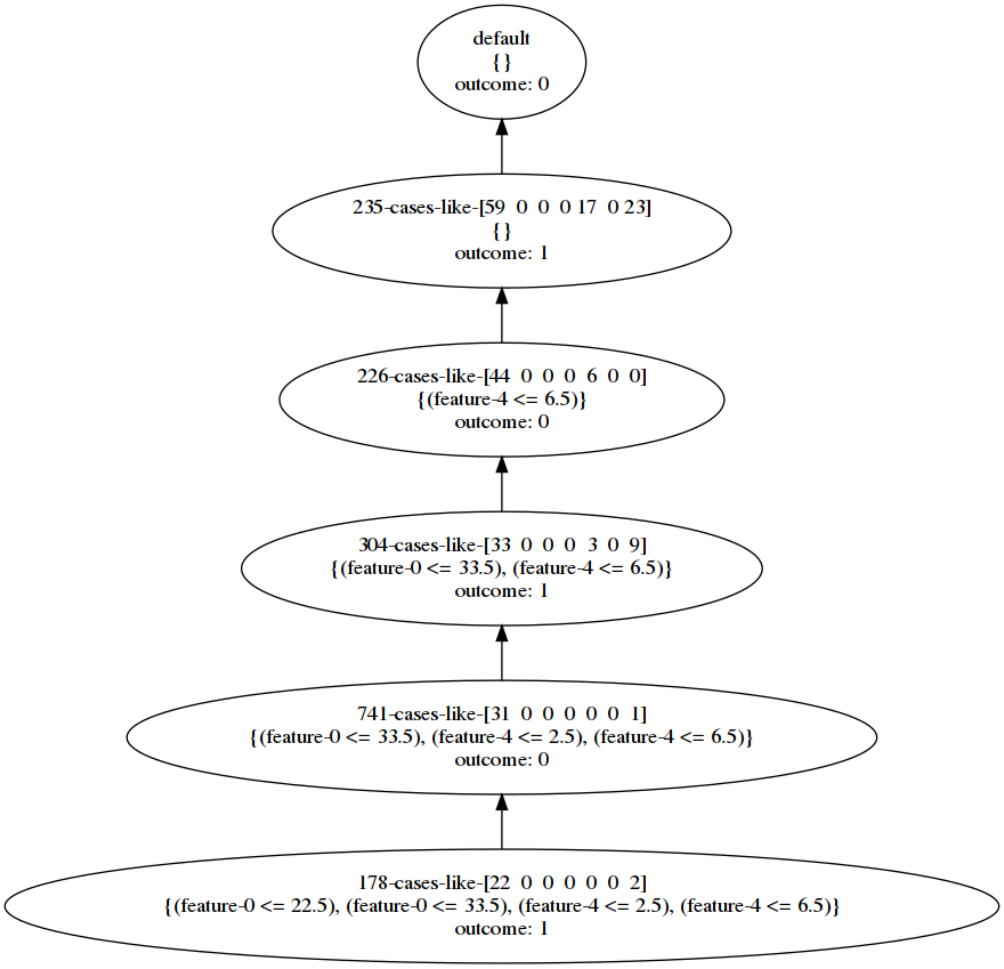


Figure 4.5: $cAF_{\succeq}(D)$, the AF mined using $cAA\text{-}CBR_{\succeq}$ from the training set (single split) using feature set D and the partial order from decision trees, using maximum depth 3 and maximum number of leaf nodes 8, removing spikes, and using the `majority` strategy for incoherence. See caption of Figure 4.4 for details.

4.4.3 Results for partial order via autoencoders

We now present our results for the autoencoder approach for learning the partial order (Section 4.3). We evaluated performance using hyperparameter search on a validation set, for each of $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$. Hyperparameters used are in Table 4.5. We use rectified linear units (ReLU) (Nair and Hinton, 2010; Glorot et al., 2011) as the hidden activation function, and Adam as the optimiser (Kingma and Ba, 2015).

We report results in Tables 4.6 and 4.7. Accuracy results show that $AA\text{-}CBR_{\succeq}$ is below the baseline of always predicting the majority class, while $cAA\text{-}CBR_{\succeq}$ is above, although barely. As expected, $cAA\text{-}CBR_{\succeq}$ produces a smaller AF, with fewer nodes and attacks. Again, this is an advantage in terms of space, that is, memory complexity. Interestingly, in this case $cAA\text{-}CBR_{\succeq}$ depth values are smaller than for $AA\text{-}CBR_{\succeq}$, differently from what happened in the decision tree approach. Since the learned representation is very noisy, the difference in performance might be caused by a higher overfit of $AA\text{-}CBR_{\succeq}$ by including every case in the AF. This is suggested by the higher depth of $AA\text{-}CBR_{\succeq}$, since it implies the extra cases are not merely acting as "shortcuts" to the default case, as in the decision-tree-based AFs (e.g. see Figure 4.4). Those are preliminary results on the methodology, and further analyses are required to diagnose this difference in performance.

We leave as future work to verify if current limitations in performance can be overcome with improvement of the autoencoder performance (that is, by other architecture and training choices that reduce reconstruction error). A second possibility is to also use output labels in the neural network training, so that the autoencoder architecture is informed by the final task that we aim to produce. Initial trials with this approach show that naïve ways of multi-task training result in autoencoder collapse (that is, all training examples are mapped to the same hidden encoding). Further exploring this direction is still an open avenue for future work.

4.5 Discussion

An interesting consideration is that, for (regular) $AA\text{-}CBR_{\succeq}$, the definition of the partial order defines much of the 4R CBR cycle (Section 1.1.2): retrieval

Table 4.5: Hyperparameter options for the autoencoder approach experiment. Only number of dimensions of hidden encoding and of hidden layers are tuned. Best combination chosen by grid search on validation set.

number of encoder layers	3
number of decoder layers	3
number of dimensions of hidden encoding	$\{2, 3, 5, 10\}$
number of dimensions of hidden layers	$\{5, 7, 10, 12, 15\}$
learning rate	0.001
epochs	40

Table 4.6: Results for the experiment in learning the partial order via autoencoders. For comparison with the models below, the baseline approach of always predicting the majority class (`no recid`) results in an accuracy of 54.43%.

				depth		
	accuracy (%)	# nodes	# attacks	max	mean	std dev
$AA\text{-}CBR_{\succeq}$	53.19	2505	6391	6	2.03	0.92
$cAA\text{-}CBR_{\succeq}$	54.86	1840	3635	5	1.63	0.64

is determined by relevance, which in turn is determined by the partial order. Reuse of solutions depend on the structure of the AF mined from the data, which again depends on the partial order. Assuming no (human) revision of the solution, then retaining is also dependent on the partial order, since retaining is nothing more than adding into the mined AF, thus a function of the partial order, as just mentioned.

There are different approaches in the literature for connecting CBR with machine learning methods with the goal of applying CBR. We will briefly mention some recent approaches to this issue in AI and law. Mumford et al. (2022) proposes a neural-network-based method for ascribing factors from natural language facts. Those factors come from a hierarchy of factors specified

Table 4.7: Hyperparameter values found in hyperparameter optimisation for each model.

	encoding_dim	hidden_size
$AA\text{-}CBR_{\succeq}$	5	10
$cAA\text{-}CBR_{\succeq}$	5	10

with ADFs (Al-Abdulkarim et al., 2015b). Grabmair (2016; 2017b) proposes a method using argument schemes with issues and quantitative value effects. Factors have effects on values with weights learned via an iterative training procedure, in which argumentation-based CBR is employed. Although the details are involved, the notion of relevance in this work is based on the argument schemes (since they define argument moves that cite precedents), which in turn depend on factors, issues, and values. Woerkom et al. (2022) also work with the COMPAS dataset having CBR in mind. They apply the result model of precedential constraint (See Section 2.2.2), although for dimensions, instead of for factors. Dimensions are inferred from the data by determining a direction depending on the coefficients of a logistic regression. In terms of learning a notion of relevance in CBR, their methods essentially considers that the notion of relevance is the precedential constraint rule, and their learning is giving a total order on each feature, transforming them into dimensions. In the resulting model, they show that only 8% of the dataset is consistent, and that this is caused by outliers for which opposite outcomes would be expected. This is an interesting direction, and as mentioned in Chapter 3, better understanding the interplay between *AA-CBR* and precedential constraint is a path for future work.

Outside AI and law, other combinations of CBR and machine learning have been proposed for adding interpretability to models, in particular to neural networks. For instance, some ideas have been clustering data and finding prototypes (Kim et al., 2014), as well as neuro-symbolic methods which add a prototype layer in a neural network, so inference is done by calculating notions of relevance to past cases and prediction based on their outputs (Li et al., 2018; Chen et al., 2019; Barnett et al., 2021). Those can be seen as methods using neural networks for defining relevance in CBR. Additionally, recent work by by Monath et al. (2021) proposes learning a DAG structure of data using nearest neighbours. This could have connections to the DAG structure of a (coherent) *AA-CBR* mined AF, to be explored in future work.

Finally, an interesting potential advantage of the autoencoder approach is allowing *AA-CBR* to be deployed even on unstructured data, such as images and text. We have focused on COMPAS and thus have not explored this direction, which we leave to future work. Future work could also explore

the relationship between the effectiveness of the autoencoder and the final performance, as well as ways of informing the autoencoder about the outputs of the final task, in order to make this approach feasible. Another open problem is how to make partial orders learned from an autoencoder more interpretable, perhaps by the comparison of examples along the learned partial order or generating interpolations between them.

4.6 Summary

In this chapter we presented approaches to learn the partial order for $AA\text{-}CBR$ from data on the case of COMPAS, a tabular dataset, on the task of predicting recidivism. We show that binary splits of decision trees learned using CART can be used as features for $AA\text{-}CBR$ and allow its instantiation as $AA\text{-}CBR_{\supseteq}$. This methodology is empirically successful, having performance comparable to decision trees. This satisfies our first desideratum from the beginning of the chapter. As the learned characterisation is of binary features consisting of split tests, they are also interpretable. However, our third desideratum is limited since decision trees are not well-suited to unstructured data. At the cost of interpretability, we propose a novel approach for learning the partial order based on hidden encodings derived from an autoencoder, a neural network architecture based on reconstructing the input. We show preliminary results on this methodology.

Here we also empirically compared $AA\text{-}CBR_{\succeq}$, the original presentation, with $cAA\text{-}CBR_{\succeq}$, our novel cautiously monotonic version. We concluded that in the presence of incoherence (either by noise in the data or due to the learning of the partial order), $cAA\text{-}CBR_{\succeq}$ presents a strong advantage. This can be offset by the use of strategies to remove incoherence. When simply removing incoherent pairs, in our experiments $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$ present comparable performance, controlling by hyperparameter choice. When the application domain allows a more intuitive way of dealing with incoherence, the difference might disappear entirely or even $AA\text{-}CBR_{\succeq}$ may present a (in our experiments, slight) advantage. This is seen in our experiments in the strategy of selecting the majority outcome. Still, $cAA\text{-}CBR_{\succeq}$ results in considerably smaller AFs, compared to $AA\text{-}CBR_{\succeq}$, and cumulativity is enforced in this way.

As opposed to controlling by hyperparameter choice, we also evaluate the difference between $AA\text{-}CBR_{\succeq}$ and $AA\text{-}CBR$ in a proper methodology of model selection on a validation set. Both alternatives present not only identical performance, but identical decision functions, albeit with differences in inner structure. Using $cAA\text{-}CBR_{\succeq}$ results in AFs with fewer nodes, but at least in the optimal models $cAA\text{-}CBR_{\succeq}$ can present greater maximum depth, interestingly. This could have impacts in terms of explainability. On the one hand, fewer nodes mean the AF on inspection is easier to comprehend. On the other hand, increasing maximum depth could lead to more complicated arbitrated dispute trees (Section 2.4.1), for instance. Still, what differences will actually occur can be application-dependent, and thus our conclusions might not immediately generalise, requiring more experiments on other datasets. This is an avenue for future research. Impacts on memory and speed of using $cAA\text{-}CBR_{\succeq}$ and $AA\text{-}CBR_{\succeq}$ were not directly measured, but we expected that $cAA\text{-}CBR_{\succeq}$ would present advantages due to its smaller AFs.

Chapter 5

Interactive Explanations as Non-Monotonic Reasoning

Until now, we have discussed properties of non-monotonic reasoning (NMR) for classification. Another topic in this thesis is whether we can also use NMR to understand explainability. This is the topic to which we now turn. In particular, we consider a scenario of interactive exchange between a system, which tries to explain itself, and a user, who tries to understand the system.

Recent work shows issues of consistency with explanations, with methods generating local explanations that seem reasonable instance-wise, but that are inconsistent across instances (Camburu et al., 2020; Merrer and Trédan, 2020). This suggests not only that instance-wise explanations can be unreliable, but mainly that, when interacting with a system via multiple inputs, a user may actually lose confidence in the system. To better analyse this issue, in this chapter we treat explanations as objects that can be subject to reasoning and present a formal model of the interactive scenario between user and system, via sequences of inputs and outputs. We argue that explanations can be thought of as committing to some model behaviour (even if only *prima facie*). This in turn can be modelled in terms of entailment, which, we argue, should be thought of as non-monotonic. This allows: 1) to solve some considered inconsistencies in explanations; and 2) to consider properties from the non-monotonic reasoning (NMR) literature presented in Section 2.1.1 and discuss their desirability, gaining more insight on the interactive explanation scenario. We also consider the case of arbitrated dispute trees (ADTs) as explanations

for *AA-CBR* using this methodology, providing a method for readapting ADT explanations to other cases, and showing an interpretation of this method as entailment.

We begin this chapter motivating further the problem (Section 5.1). This is followed by a description of the scenario we are modelling (Section 5.2) and, in turn, for the formal modelling of this scenario (Section 5.3). Next, we present the NMR properties adapted to this modelling (Section 5.4). We then apply this modelling to arbitrated disputed trees as explanations for *AA-CBR* (Section 5.5).

5.1 Motivation

A growing body of research is dedicated to providing (post-hoc) explanations to AI models, in particular machine learning systems (Nugent and Cunningham, 2005; Štrumbelj and Kononenko, 2014; Ribeiro et al., 2016a; Čyras et al., 2016b; Li et al., 2018; Wachter et al., 2018; Mueller et al., 2019; Teso and Kersting, 2019; Ignatiev et al., 2019; Gunning, 2019; Kenny and Keane, 2019; Miller, 2019; Atkinson et al., 2020; Kenny and Keane, 2021; Zhou et al., 2022; Raczyński et al., 2023; Gyevnar et al., 2023; Gniewkowski et al., 2023). This literature includes many discussions on what explanations are and should be, including important aspects according to social sciences (Miller, 2019) and to law (Gyevnar et al., 2023), and lists of desiderata (Langer et al., 2021). However, there is no standardised formal definition of explanation, and of what *formal* properties explanations should satisfy, with different explanation methods having different goals, often only implicitly stated.

Issues of consistency have been raised in recent work. Camburu et al. (2020) show how inconsistent textual explanations can be generated for a natural language inference task. Merrer and Trédan (2020) argue that local explanations for a single prediction are always subject to manipulability by hiding how a protected feature was used, but present, as a method of detecting this manipulation, that an auditor queries the system multiple times and detects whether there is an inconsistency in the returned explanations. Thus, lack of consistency may not only cause a user to wrongly expect some system behaviour, but may also cause loss of user's confidence in the system overall, or imply that the

explanations are not faithful, or even malicious. Therefore it is increasingly important to discuss the meaning of consistency in the case of local explanations, an issue that can occur when querying a system multiple times, with different inputs. While the idea of consistency may first evoke classical logic, a well-established tradition in logic-based AI argues that in many applications the form of reasoning more natural for humans is not classical logic, but different models of reasoning: NMR. Indeed, a line of work in psychology, cognitive science and AI empirically evaluates inferences humans make in a task with conditional arguments, challenging a classical logical interpretation of them (Byrne, 1989; Neves et al., 2004; Stenning and Lambalgen, 2005; Klauer et al., 2007; Fugard et al., 2011; Ragni et al., 2016; Ragni et al., 2017; Spiegel et al., 2019; Struchiner et al., 2020; Marcos Cramer, 2021). Understanding explanations as presenting conditional arguments for expected outputs given some inputs would, then, lead one to model explanations as NMR.

Inspired by this line of work, we propose an analysis of reasoning with explanations for an interactive scenario. We think of explanations as objects upon which one can reason. Thus we model explanations abstractly and define which inferences we can make from them, while also considering reasoning properties they may satisfy.

5.2 The interactive explanation process

Consider the scenario of a user or auditor evaluating the behaviour of a model. They may ask for the output for a specific input, and ask as well for an explanation. While much of this literature treats this as a one-off task, a single explanation may be insufficient. The user could keep exploring model behaviour by querying for more outputs (with the same or different inputs) and explanations of the model. We thus consider an interactive process (as overviewed in Fig. 5.1), where the user gives an input, querying for an output and explanation thereof. The system returns them, and then the user may then query again. We see the AI system as including a *classifier*[1] and an *explainer*, with both considered black-boxes by our conceptual model.

[1]This choice is dictated by the focus, in the XAI literature, on explaining outputs of classifiers, as well as for simplicity. However, our conceptual understanding of the interactive explanation process could in principle be applied to different types of model.

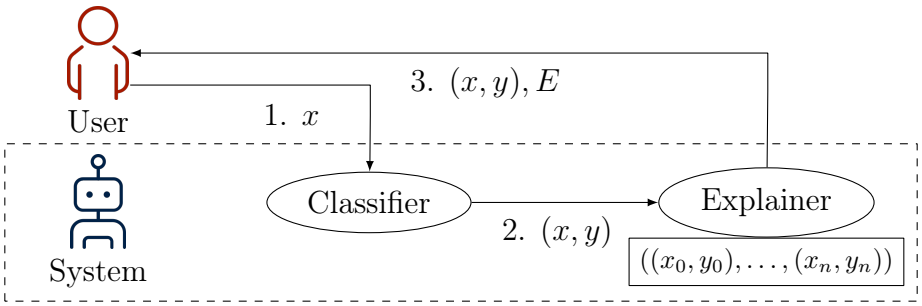


Figure 5.1: Overview of the *interactive explanation* process. 1. The user queries the system with an input x, which goes to the classifier. 2. The classifier produces output y, and sends the pair (x, y) to the explainer. 3. The explainer produces an explanation E for (x, y), using information from the history of inputs/outputs $((x_0, y_0), \ldots, (x_n, y_n))$, and sends (x, y) and E back to the user, who may then stop or further query the system.

We assume that the explanation method (although not the classifier) can keep track of the *history* of inputs it received and their outputs. This allows supporting two scenarios: i) of an explainer that tries to improve its explanations by knowing what it explained before; and ii) of a malicious explainer that, trying to manipulate or mislead the user and to avoid being detected, keeps track of what was explained before. Histories give a snapshot of the process by finite sequences of inputs and their outputs. Note that we assume the explainer to be a function, thus ignoring randomness of the explanation method itself. This assumption implies that no information about the computed explanations needs to be stored in the history, as awareness of earlier inputs and outputs already gives all information need to know what explanations were also previously returned.[2]

Example 5.1. For concreteness, let us consider a scenario where inputs are drawn from $X = \mathbb{R}^2$, outputs from $Y = \{0, 1\}$, and explanations ($\mathcal{E}$) are decision sets (Lakkaraju et al., 2016), that is, sets of rules, each of the form $s \to y = c$, where s is an itemset (a conjunction of predicates of the form

[2]Our assumption here that the explainer is a deterministic function simplifies our modelling, especially Fig. 5.2 (discussed later). In case it is not deterministic, such as a method that includes sampling, past explanations should be included directly in the history. We leave this extension of our model to future study. In any case, this assumption is a typical assumption in machine learning (e.g. Abu-Mostafa et al., 2012). Notice that here being deterministic is a matter only of inference, not of training, since here we consider the classifier fixed. This is a more restricted issue than the one in Section 3.2.

(feature, operator, value), such as $f > 1$), and $c \in Y$.[3]

A possible scenario of interaction is as follows. The user queries for $x_0 = (5, 0)$, corresponding to feature assignments $f = 5, g = 0$; the classifier returns $y_0 = 1$, and the explanation given is $E_0 = \{f > 0 \to y = 1\}$. The user decides to investigate further and tries $x_1 = (20, 5)$. Even though this is covered by the previous rule, which would commit to the output $y_1 = 1$, suppose the output is $y_1 = 0$ and the explanation is $E_1 = \{f > 10 \wedge g > 3 \to y = 0\}$. How should the user interpret this? Is this an inconsistency in the explanatory process?

5.3 Formal modelling

We assume an input set X and an output set Y as well as a set of possible explanations $\mathcal{E}$, that we keep abstract. Regarding notation, for a set S, we denote the set of finite sequences of element of S as $Seq(S)$, i.e., $Seq(S) = \bigcup_{i \in \mathbb{N}} S^i$ for S^i a sequence of i elements of S. Given $n \in \mathbb{N}$, we use the notation $[n] = \{m \in \mathbb{N} \mid m \leq n\}$. Thus a sequence $(s_0, s_1, \ldots, s_n) \in Seq(S)$ can be written as $(s_i)_{i \in [n]}$.

Definition 5.2. A *classifier* is a function $\mathbb{C} : X \to Y$. An *explanation method* is a function $\mathbb{E} : Seq(X \times Y) \to Seq(\mathcal{E})$, that is, mapping from a sequence of input-output pairs $(x_i, y_i)_{i \in [n]}$ to a sequence of explanations $(E_i)_{i \in [n]}$ of the same length.

We consider the system to be composed of a classifier and an explanation method. Notice that the explainer uses information on the entire sequence of inputs-outputs, as we discussed in Section 5.2. We can think of this sequence of input-output pairs as a *history*, so that, when explaining the most recent input, the entire past of input-outputs is taken into account. We consider that at each time step $t \in \mathbb{N}$, the user queries for an input $x_t \in X$, which receives a classification $\mathbb{C}(x_t) = y_t$ and an explanation E_t. In this way, the explainer provides an explanation motivated by a specific input-output, while considering the history $((x_i, y_i))_{i \in [t-1]}$.

[3]However, as opposed to the original definition in (Lakkaraju et al., 2016), we assume no default class (so that if no rule applies, nothing is expected) and no tie-breaking function, that is, multiple classes may be predicted.

An important particular case is when the explanation method is defined in terms of explanations that originating from a single input-output pair:

Definition 5.3. An *individual explanation method* is a function $\mathbb{E}_\bullet : X \times Y \to \mathcal{E}$, mapping from a single example (x, y) to an explanation E. The *explanation method lifted from* $\mathbb{E}_\bullet$ is $\mathbb{E}(((x_i, y_i))_{i\in[n]}) = (\mathbb{E}_\bullet(x_i, y_i))_{i\in[n]}$.

In this case, the explainer function $\mathbb{E}$ is defined as applying $\mathbb{E}_\bullet$ to each element of the sequence. This way, the history is disregarded when explanations are computed.

The reader may notice that this view of the interactive explanation process does not enforce that previously exhibited explanations are kept, that is, that E_i is unchanged for all $i \in [t]$ when x_{t+1} is queried. If instead we want past explanations to be unretractable, in the sense that the system cannot replace the explanation of any previous query, then we can require the following property:

Definition 5.4 (Interaction-stability)**.** An explainer $\mathbb{E}$ is said to be *interaction-stable* whenever, for every sequence of input-output pairs $(x_i, y_i)_{i\in[n]}$ and for every $m < n$, if $(E_i)_{i\in[n]} = \mathbb{E}((x_i, y_i)_{i\in[n]})$ and $(E'_i)_{i\in[m]} = \mathbb{E}((x_i, y_i)_{i\in[m]})$ then $E_i = E'_i$ for any $i \in [m]$.

That is, an interaction-stable explainer will always keep the explanation E_i associated to the pair (x_i, y_i), even as the interaction moves on. Explainers which are not interaction-stable may not be unreasonable. Assume that, at a new instance, a new explanation would be inconsistent with the previous ones. The explainer might be such that previous explanations are "corrected" in order to avoid the inconsistency. In a sense, this would make this explainer more analogous to a belief revision system, instead of solving the issue as NMR with the inference relation later. Still, this may or may not be appropriate depending on what kind of user interface is available, since the user would need to be aware that previous explanations were adjusted. Thus, interaction-stable explainers are arguably simpler in a way, since the system would not retract a previous explanation.

We now turn to the modelling of inference. We assume that a sequence of explanations $(E_i)_{i\in[n]}$ "commits" to some model behaviour. We model this in the following way:

Definition 5.5. An *entailment relation* $\models$ is a relation between $Seq(\mathcal{E})$ and $X \times Y$, that is, $\models \subseteq Seq(\mathcal{E}) \times (X \times Y)$. We say a sequence of explanations *entails* a pair (x, y) iff $(E_i)_{i\in[n]} \models (x, y)$.

Intuitively, $(E_i)_{i\in[n]} \models (x, y)$ means that $(E_i)_{i\in[n]}$ "commits" to the outcome y, given the input x. We will abuse notation and define $E \models (x, y)$ to mean $(E) \models (x, y)$ (for (E) the sequence with just one explanation, E).

This entailment relation we keep abstract and application-dependent. What is important is that it captures how a user would interpret the explanation or plausible inferences from it, including as regards the input-output being explained. One example is explanations as sufficient reasons: any sufficient reason is exactly a rule that guarantees the output for a part of the input space, including the given input.

Example 5.1 (continuing from p. 128)**.** Assume the entailment relation $\models$ from sequences of explanations $(E_i)_{i\in[n]}$ is to be interpreted as classification from the union $\bigcup_{i\in[n]} E_i$ of all explanations, corresponding to a naive interpretation of decision sets without tie-breaks. Then, x_1 is covered by both E_0 and E_1, for different classifications, which the user may take to be an inconsistency, since no input can have two different outputs. A different intended interpretation of the sequence of rules is possible. Define a rule to be more specific than another whenever every input x that satisfies the itemset of the first rule also satisfies the itemset of the second rule. Now assume that only most specific rules are applicable. For this example, E_1 is more specific than E_0, solving the inconsistency issue. This interpretation is *non-monotonic* since, while $(E_0) \models (x_1, 0)$, we have $(E_0, E_1) \models (x_1, 1)$.

From this core notion of $\models$, relating explanations to input-output, we can derive two "homogeneous" notions of "entailment", that is, from sequences of elements of a set to elements of the same set. One is at input-output level, and the other at explanation level. For the former, we say $(x_i, y_i)_{i\in[n]} \vdash (x, y)$ iff $\mathbb{E}((x_i, y_i)_{i\in[n]}) \models (x, y)$. For the latter, $(E_i)_{i\in[n]} \vdash_e E$ iff $\forall (x, y) \in X \times Y$, if $E \models (x, y)$ then $(E_i)_{i\in[n]} \models (x, y)$. This we summarise in Fig. 5.2.

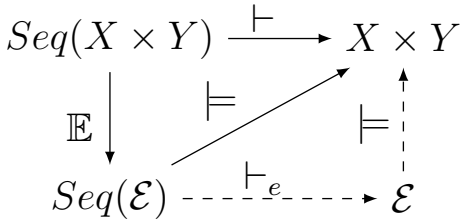


Figure 5.2: Diagram illustrating the relations between sequences of input-output pairs, sequences of explanations, single pairs and single explanations. This diagram does not always commute (indicated by dashed lines) since for a sequence of explanations $\vdash$-entailing an input-output pair, there may not exist an (individual) explanation $\vdash_e$-entailed by this sequence which $\models$-entails this pair.

5.4 Consistency and non-monotonicity

We now turn to properties of explainers, expressed by properties of *consistency* and *non-monotonicity* for the relations associated with explainers. Notice that, since the notion of entailment here is on sequences, the definitions from Section 2.1.1 are not readily available, and require adaptation. We provide our adaptations here.

Definition 5.6 (Consistency)**.** A sequence of explanations $(E_i)_{i\in[n]}$ is said to be *consistent* iff there does not exist $x \in X$, $y, y' \in Y$, with $y \neq y'$, such that $(E_i)_{i\in[n]} \models (x, y)$ and $(E_i)_{i\in[n]} \models (x, y')$.

An entailment relation $\models$ is said to be *consistent* iff every sequence of explanations is consistent.

A relation $\vdash$ is said to be *consistent* iff there does not exist $x \in X$, $y, y' \in Y$, with $y \neq y'$, and $((x_i, y_i))_{i\in[n]}$ such that $((x_i, y_i))_{i\in[n]} \vdash (x, y)$ and $((x_i, y_i))_{i\in[n]} \vdash (x, y')$.

Since the relations $\vdash$ and $\vdash_e$, derived from the base notion of $\models$, are "homogeneous", we can define properties borrowed from the literature on NMR (Makinson, 1994; Kraus et al., 1990), in a manner that would not be possible for the relation $\models$. The difference from the definitions here from the ones presented in Section 2.1.1 is that we generalise them to sequences, instead of using sets.

Our hypothesis is that if those properties are important in (non-monotonic, defeasible) reasoning, and if the (interactive) explanation process can be seen

as form of (non-monotonic) reasoning, then satisfying (or not) these properties may play a role on how users interact with explanations and what expectations they may (legitimately) have from them. Indeed, conditional reasoning has been explicitly evaluated in humans via questions motivated by such properties and, preliminary, support has been found for some properties, such as cautious monotonicity (Neves et al., 2004), although the general picture is less clear in evaluations comparing human inferences with different non-monotonic logic formalisms (Ragni et al., 2016; Ragni et al., 2017). Investigations with formal argumentation in mind have been surveyed by Cerutti et al. (2021), although results from this literature are not yet conclusive. In any case, what seems to be clear is that monotonic logics are not psychologically plausible models of conditional reasoning in humans in general (Byrne, 1989; Stenning and Lambalgen, 2005; Fugard et al., 2011; Spiegel et al., 2019; Espino and Byrne, 2020; Ragni et al., 2020; Cerutti et al., 2021; Marcos Cramer, 2021).

Definition 5.7 (Non-monotonicity and reflexivity)**.** The relation $\models$ is said to be *non-monotonic* iff there are $(E_i)_{i\in[n]}$, E_{n+1} and (x, y) such that $(E_i)_{i\in[n]} \models (x, y)$ and $(E_i)_{i\in[n+1]} \not\models (x, y)$.

Given a set S, a relation $\vdash'$ from $Seq(S)$ to S, and $s, s_i \in S$, for $i \in \mathbb{N}$, the relation $\vdash'$ is said to satisfy:

- *non-monotonicity* iff there are $(s_i)_{i\in[n]}, s_{n+1}, s$ such that $(s_i)_{i\in[n]} \vdash' s$ and $(s_i)_{i\in[n+1]} \not\vdash' s$;
- *reflexivity* iff for every $(s_i)_{i\in[n]}$ and i, $(s_i)_{i\in[n]} \vdash' s_i$;
- *cautious monotonicity* iff for every $(s_i)_{i\in[n]}$, s_{n+1}, and s, if $(s_i)_{i\in[n]} \vdash' s_{n+1}$ and $(s_i)_{i\in[n]} \vdash' s$, then $(s_i)_{i\in[n+1]} \vdash' s$;
- *cut* iff for every $(s_i)_{i\in[n]}, s_{n+1}, s$, if $(s_i)_{i\in[n+1]} \vdash' s$ and $(s_i)_{i\in[n]} \vdash' s_{n+1}$, then $(s_i)_{i\in[n]} \vdash' s$.

In the definition above, notice that S and $\vdash'$ are defined generally in order to include $(x_i, y_i)_{i\in[n]}$ and $\vdash$, as well as $(E_i)_{i\in[n]}$ and $\vdash_e$.

We can now make some general considerations regarding those properties. Reflexivity asserts that whatever is a premise should also be entailed. We would expect this to hold for the relation $\vdash$ for many explainers, since violating

this means that, after some interactions, an input-output pair could become unexplained by the current understanding of the sequence of explanations. That is, the interactive process fails for some asked input and returned output, which seems a failure of explanation. On the other hand, for $\vdash_e$, a failure of reflexivity would not be surprising. Indeed, Example 5.1 shows such a failure, where $(E_0, E_1) \nvdash_e E_0$.

Regarding cautious monotonicity, intuitively it captures the idea that if an expected behaviour is confirmed and made explicit, then other previously expected behaviours are still expected, i.e., a break of expectation is only caused by explicitly observing an unexpected behaviour. For $\vdash$ this means that, when querying for an input which is classified as the (then) expected output, the resulting sequence of explanations, regardless of what it is, does not entail any less than what was entailed before. Cut, on the other hand, is in a way the converse: the resulting sequence of explanations does not entail any more than what was already entailed. As for $\vdash_e$, what cautious monotonicity means is that, if a sequence of explanations entails a single explanation, and this is appended to the sequence, then no entailment is lost. Again, cut means that no conclusion is gained.

Cut fails in some probabilistic logics (Makinson, 1994), and indeed one can think of further explanations as detailing the previous ones, even if already entailed, so that more cases are covered as more is explained. Cautious monotonicity also has a pragmatical motivation of reducing updates in beliefs (Kraus et al., 1990, p. 12), which arguably could be relevant for the plausibility to, or comfort of, a user in an interactive explanation process. Which concrete methods satisfy which of those properties and what is the impact on user experience are open questions. We here do not empirically investigate the severity of violating or satisfying those properties. This is a direction left for future work.

Example 5.8. As a variation of Example 5.1, assume the specificity interpretation and that for the input $x_1 = (20, 5)$, the output was now instead $y_1 = 1$, with second explanation now $E_1' = \{f > 10 \wedge g \leq 0 \rightarrow y = 0, f > 10 \wedge g > 0 \rightarrow y = 1\}$. It is the case that $(E_0) \models (x_1, 1)$, and so $((x_0, y_0)) \vdash (x_1, 1)$. Now consider a third input $x_2 = (20, -10)$. While $(E_0) \models (x_2, 0)$ (thus $((x_0, y_0)) \vdash (x_2, 0)$), we now have that $(E_0, E_1') \models (x_2, 1)$, since both rules in E_1' are more

specific than the one in E_0. Therefore we have that $((x_0, y_0), (x_1, 1)) \vdash (x_2, 0)$, even if $((x_0, y_0)) \vdash (x_1, 1)$ and $((x_0, y_0)) \vdash (x_2, 0)$, violating cautious monotonicity.

5.4.1 Specificity for non-monotonic explanations

We will generalise here how the idea of specificity used in Examples 5.1 and 5.8 can be defined for general entailment relations $\models$, or even to characterise one, in an abstract way.

Definition 5.9. A sequence of explanations $(E_i)_{i\in[n]}$ is said to **cover** an input $x \in X$ iff there is $y \in Y$ such that $(E_i)_{i\in[n]} \models (x, y)$.

A sequence of explanations $(E_i)_{i\in[n]}$ is said to be **more specific than** another $(E'_i)_{i\in[m]}$ iff for every input $x \in X$ that $(E_i)_{i\in[n]}$ covers is also covered by $(E'_i)_{i\in[m]}$.

In the definitions above, we can use single explanations as a sequence of explanations, following our mentioned abuse of notation. We can now define what respecting specificity means for an entailment relation $\models$.

Definition 5.10. An entailment relation $\models$ *respects specificity* iff it is consistent and if E_{n+1} is more specific than $(E_i)_{i\in[n]}$ and $E_{n+1} \models (x, y)$, then $(E_i)_{i\in[n+1]} \models (x, y)$.

A sceptical definition of $\models$ can be defined by extending the entailment relation from single explanations to sequences recursively via specificity. This exemplifies how one could define entailment in sequences from entailment in single explanations, based only on the abstract concept of entailment presented in this formalism, independent of instantiation. The definition below captures such entailment relations, presenting the inductive step:

Definition 5.11. An entailment relation $\models$ is *most sceptically specific* iff it is consistent, it respects specificity, and, if E_{n+1} is not more specific than $(E_i)_{i\in[n]}$, for every $x \in X$, $y, y' \in Y$: a) if $E_{n+1} \models (x, y)$ and $(E_i)_{i\in[n]} \models (x, y')$, with $y \neq y'$, then there is no $y'' \in Y$ such that $(E_i)_{i\in[n+1]} \models (x, y'')$; else b) if $E_{n+1} \models (x, y)$ or $(E_i)_{i\in[n]} \models (x, y)$, then $(E_i)_{i\in[n+1]} \models (x, y)$.

This does not imply our framework is restricted to such entailment relations, as it can capture other types of explanation. This is just an example of how the entailment notion of explanation from single explanations to a sequence of explanations, using specificity, a common notion in NMR (Poole, 1985; Horty, 1994; Dung and Son, 2001; Stolzenburg et al., 2003; Wirth and Stolzenburg, 2016, among many), and sceptical inference (when disagreeing, entail neither) is used for consistency (on relative consistency preservation for sceptical reasoning, see Section 2.2 of Makinson, 1994). As mentioned, an example of a most sceptically specific entailment relation is the one used used in Example 5.1.

5.5 Arbitrated dispute trees as explanations to *AA-CBR*

Now we instantiate our previous model of interactive explanations for explanations with *AA-CBR*. Dispute trees have been proposed as explanations for argumentation methods, in particular for classification with *AA-CBR*, a model based on abstract argumentation (AA) for case-based reasoning (CBR) (Čyras et al., 2016b). In the more recent version, *arbitrated* dispute trees (ADTs, here presented in Section 2.4.1) are shown as trees reflecting whether an argument is winning or losing, depending on the new case being classified (Čyras et al., 2019a).

Here we show how an ADT for classifying a new case can possibly be reused for other new cases. This shows that an ADT can be seen as "committing" to outcomes for other possible cases, instead of simply the case to be explained (and previous cases occurring in the ADT itself). Indeed, this reuse yields another admissible dispute tree, and is therefore a faithful explanation by construction.

In this section we first show an example of counterfactual explanation in *AA-CBR*, illustrating concepts defined previously in this chapter. Then we instantiate the model with a novel method of readapting an arbitrated dispute tree that explains a prediction. This method is not always applicable but, when it is, it results in a provably faithful explanation. When modelling this as reasoning, we show this will result in a form of monotonic inference.

5.5.1 An example of counterfactual explanation in $AA\text{-}CBR$

We will work with a simple example of credit application.

Example 5.12. Applicants are described by two features: age and income. System developers decided from background knowledge and previous data that the default case of an applicant is of $\{\texttt{age}:30; \texttt{income}:25\}$ (age in years, income in £1000 per year) and by the default it is denied ($y = 0$), for instance, because those values are the medians in the dataset. The casebase is formed of the following cases:

$D = \{(\{\texttt{age}:20; \texttt{income}:100\}, 1), (\{\texttt{age}:40; \texttt{income}:30\}, 1),$
$(\{\texttt{age}:32; \texttt{income}:40\}, 1), (\{\texttt{age}:60; \texttt{income}:42\}, 0), (\{\texttt{age}:86; \texttt{income}:150\}, 1)\}$

More formally, we could say inputs are elements of $X = \mathbb{R}^2$, and outputs of $Y = \{0, 1\}$. We will interpret the partial order in the following way: $\{\texttt{age}:a; \texttt{income}:b\} \preceq \{\texttt{age}:a2; \texttt{income}:b2\}$ iff i) $30 \leq a \leq a2$ or $a2 \leq a \leq 30$; and ii) $25000 \leq b \leq b2$ or $b2 \leq b \leq 25000$. Essentially, a case α is smaller than β when, for both coordinates (features), the value of the coordinate for α is between β and the default value. Notice this means that the notion of atypicality, represented by the partial order $\preceq$, is being, at each coordinate, either larger at values larger than the mean or smaller at values smaller than it, in such way that values at different sides of this threshold are incomparable. This captures atypicality as being either "too large" or "too small" with respect to the mean. The corresponding framework is presented in Figure 5.3.

Having this setup in terms of case-based reasoning, we will continue the example as a classification problem in which $AA\text{-}CBR$ is used as the classifier. We will first illustrate a counterfactual-based explanation, proposing an interpretation to it. Our goal is later to compare this example with another type of explanation that we are proposing here, namely, ADT readaptation (Example 5.15).

Example 5.13 (continuing from Example 5.12)**.** Let us illustrate the framework with a model of counterfactual explanations. We will continue Example 5.12 and assume the classifier is a regular $AA\text{-}CBR_{\succeq}$ model mined from D. For illustrative reasons, we will assume this classifier is treated as a black-box model. As for the counterfactual explanations, we will not assume a particular algorithm, but we will assume it finds a counterfactual by

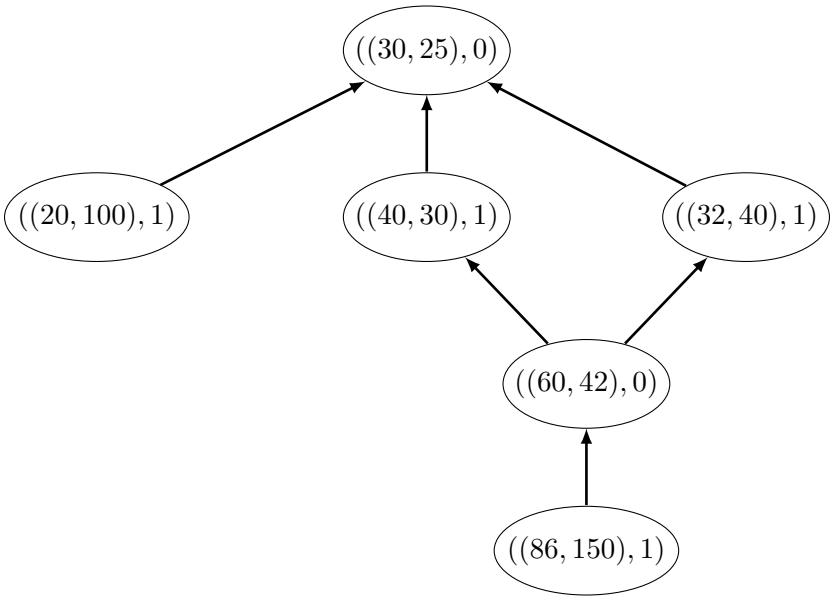


Figure 5.3: Casebase D from Example 5.12 structured as an AA framework (i.e. without a focus case).

heuristic search (Wachter et al., 2018; Hildebrandt et al., 2020). Besides, we will interpret it with a minimality assumption: a counterfactual is taken to be minimally different from the input, corresponding to the idea that it is as close as possible (Wachter et al., 2018).

This way, for a single input-output pair (x, y), the method returns a new input-output pair (x_{cf}, y_{cf}). The intended interpretation is: let $(\Delta_a, \Delta_b) = x - x_{cf}$, and (Δ'_a, Δ'_b) such that, for $0 \leq \kappa_a, \kappa_b \leq 1$, $\Delta'_a = \kappa_a \Delta_a$, and $\Delta'_b = \kappa_b \Delta_b$. For any such (Δ'_a, Δ'_b), $x' = x + (\Delta'_a, \Delta'_b)$ is intuitively a smaller change to x than x_{cf}, and thus is expected to be classified as y, and not as y_{cf}.

For example, for a case $x_0 = \{\texttt{age:}50; \texttt{income:}50\}$ the output is $y_0 = 1$, if the counterfactual is $e_0 = (\{\texttt{age:}75; \texttt{income:}43\}, 0)$, then for input $x_1 = \{\texttt{age:}65; \texttt{income:}45\}$ the expected output would still be 1.

Now suppose a user interacts with the system querying for x_0, receives y_0 as output and e_0 as an explanation. What would be expected if the user queries for x_1, but instead receives $y_1 = 0$ and $e_1 = (\{\texttt{age:}59; \texttt{income:}43\}, 1)$? Is this an inconsistency of the method? Could it be solved?

Let us try to solve those questions instantiating our framework with the problem in this example and the explanation method used.

Example 5.13 (continuing from p. 137)**.** An instantiation for counterfactuals as presented in this example is that an explanation is a tuple of original input, output, and counterfactual and counterfactual output (x, y, x_{cf}, y_{cf}), meaning that (x_{cf}, y_{cf}) is a counterfactual for (x, y)[4]. Since the explanation is derived only in terms of a specific example, we may extend for a sequence of examples by assuming that the explainer $\mathbb{E}$ is interaction-stable. That is, we have our counterfactual explanation as a single-instance explainer $\mathbb{E}_\bullet$ (as in Section 5.3). The entailment relation is then defined as:

For a single explanation $E = (x, y, x_{cf}, y_{cf})$, $E \models (x', y')$ iff:

1. $x' = x_{cf}$ and $y' = y_{cf}$; or
2. $x' = x$ and $y' = y$; or
3. $y' = y_{cf}$ and $x' = x + (\Delta'_a, \Delta'_b)$, for $(\Delta_a, \Delta_b) = x - x_{cf}$, $(\Delta'_a, \Delta'_b) = (\kappa_a \Delta_a, \kappa_b \Delta_b)$, for $0 \leq \kappa_a, \kappa_b \leq 1$.

For a sequence of explanations, a naïve aggregation can be used: $(E_i)_{i\in[n]} \models (x', y')$ iff there is i such that $E_i \models (x', y')$. Using this definition, the behaviour in Example 5.13 is inconsistent.

An alternative method could aggregate restricting by a specificity criterion, using the most sceptically specific relation (defined in Section 5.4.1): say E covers x' iff there is y' such that $E \models (x', y')$. We say that E_1 is more specific than E_2 if the set of inputs covered by E_1 is a subset of the set of inputs covered by E_2. Then redefine for sequences in the following way: $(E_i)_{i\in[n]} \models (x', y')$ iff there is i such that $E_i \models (x', y')$ and there is no E_j such that $E_j \models (x', y'')$, where $y' \neq y''$. By this definition, even if $e_0 \models (x_1, 1)$, $(e_0, e_1) \models (x_1, 0)$ and $(e_0, e_1) \not\models (x_1, 1)$. This would make such relation consistent and non-monotonic.

5.5.2 ADT readaptation as interactive explanations

While arbitrated dispute trees have been used as explanations in previous literature (Čyras et al., 2019a), here we instantiate our framework for interactive

[4]Repeating the input inside the explanation is necessary for the proper interpretation of the explanation, as the reader will now see.

explanations as ADTs. Our main question is: given an ADT, to what model behaviour does it "commit"? What should the $\models$ relation be for ADTs?

ADTs are way of presenting and explaining the behaviour of *AA-CBR* as a more succinct structure than the entire argumentation framework. Our idea here is simply that such presentation can be generalised to other cases: intuitively, an ADT for a new case x may show a set of sufficient reasons for justifying a decision for another new case $x2$. In a sense, $x2$ does not need more information than was already presented in that ADT in order to be decided, therefore the rest of the AF is unnecessary. This can be verified simply by the structure of the decision tree and by checking for relevance. Let us first define this formally.

Let D be a dataset, and $(Args, \rightsquigarrow) = AF_{\succeq}(D)$ be its corresponding argumentation framework. Let x be a new case, and $\mathcal{DT}$ be an ADT explanation for its prediction. Now let $x2$ be a second new case. Depending on the partial order relation, on the ADT and on the new cases, it might be possible to reuse the ADT $\mathcal{DT}$, without looking at $(Args, \rightsquigarrow)$ directly, to decide the outcome for $x2$. Indeed, in that case another ADT would be generated as well.

The two core ideas are: 1) if a past case is relevant to the new case, then every case smaller than it is also relevant; 2) every attacker of a W-node is in an ADT (as a L-node).

Theorem 5.14. *Let $\mathcal{DT}'$ be $\mathcal{DT}$ with all nodes labelled by $x1$ removed. For every leaf l of $\mathcal{DT}'$, let max_l be the node in the path from the root to l which is maximally relevant to $x2$ (that is, there is no node in the path greater than it such that it is also relevant to l).*

If all max_l are W-nodes, then the predicted outcome for $x2$ is the same as the predicted outcome for x. Besides, let $\mathcal{DT}_{x2}$ be the tree constructed by the following process: start with the subtree of $\mathcal{DT}'$ containing only max_l and their ancestors, add all L-nodes which are children of max_l in $\mathcal{DT}'$, and, finally, for each L-node added in this way, add as a child a new W-node labelled by $x2$. Then $\mathcal{DT}_{x2}$ is an ADT for the prediction on $x2$.

When this is the case for new cases x and $x2$, we will say $\mathcal{DT}$ can be **readapted** to $x2$. We will now prove the theorem, before proceeding to illustrate the process:

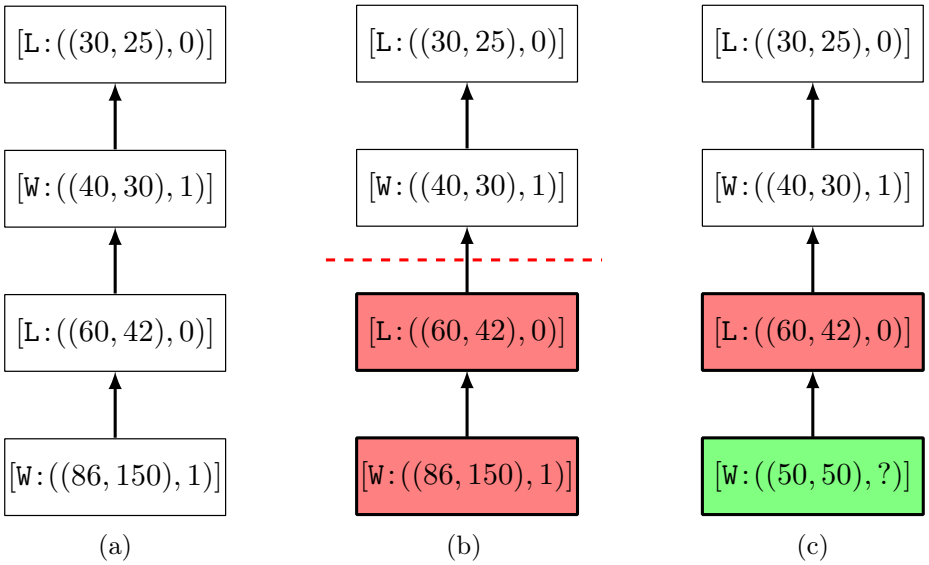


Figure 5.4: Example of ADT explanation in Example 5.15. Figure 5.4a shows the ADT explanation for input $x_2 = \{\texttt{age:}90; \texttt{income:}200\}$. Notice that x_2 itself does not appear in this ADT, since it does not participate in this chain of attacks. Indeed, this is not required by Definition 2.16, and the influence of x_2 in this explanation is in that every argument labelling a node on this ADT is not attacked by x_2. Figures 5.4b and 5.4c show the process of adapting the previous ADT to the new input $x_0 = \{\texttt{age:}50; \texttt{income:}50\}$. In Figure 5.4b, in red are the nodes labelled by arguments which are irrelevant to x_0. There is a single leaf, which is a W node. Therefore this ADT can be adapted. This is done in Figure 5.4c, with the (single) attacker of the leaf added back, but now attacked by the new case.

Proof of Theorem 5.14. For every l, max_l exists: by regularity, the default case (thus the root) is relevant to $x2$. Let us see that $\mathcal{DT}_{x2}$ is an ADT. Clearly the condition 1 is satisfied. Let us check conditions 3 and 4 for each node. For every branch, every node until max_l satisfies the conditions in Definition 2.16, since $\mathcal{DT}$ is an ADT. Now consider max_l. By assumption, every max_l is a W-node. Again, since $\mathcal{DT}$ is an ADT for x, for every max_l, each child of it is included in $\mathcal{DT}$ as a L-node child. Since those are also in $\mathcal{DT}_{x2}$, the condition is satisfied for each max_l. Next, each of those attackers require exactly one W-node as a child, attacking it. This is satisfied by every added W-node labelled by $x2$, which is an attacker since the arguments which label such L-nodes are irrelevant (otherwise max_l would not be maximally relevant). Finally, since $x2$ has no attackers in $AF^{\succeq}(D, x2,)$, it satisfies the conditions of a leaf. Condition 5 is clearly satisfied.

The last requirement to check is whether condition 2 is satisfied. Indeed it is, since, given that the other conditions are satisfied, then the set of arguments labelling W-nodes is in the grounded extension of $AF_{\succeq}(D, x2)$, which can be verified by induction (Čyras et al., 2019a, Prop. 3.3). Therefore if the root is a W-node, it is in the grounded extension and thus the prediction is δ_o. Otherwise, it is a L-node and then it has as a child which is a W-node, that is, the default argument is attacked by the grounded extension (and thus not in it, since it is conflict-free) and so the prediction of $x2$ is the $\bar{\delta_o}$. □

Example 5.15 (continuing from Example 5.12)**.** Now we will use ADT explanations for Example 5.12, instead of counterfactual explanations as in Example 5.13. Given our casebase D, suppose we would like to have the prediction for $x_2 = \{\texttt{age}:90; \texttt{income}:200\}$ and an ADT explanation. One would notice that the outcome is $y_2 = 1$ with the ADT in Figure 5.4a. The idea is, if one would like to predict for $x_0 = \{\texttt{age}:50; \texttt{income}:50\}$, this ADT could be readapted in a straightforward way. The process is illustrated in the remainder of Figure 5.4.

5.5.3 The $\models$ relation for dispute trees

Now we can instantiate our framework with dispute trees. We will define $\mathbb{E}_\bullet$ as a function that receives a new case x and its predicted outcome $AA\text{-}CBR_{\succeq}(D, x) = y$, and returns an ADT explanation $\mathcal{DT}$ for it. While the choice of the decision tree could matter for practical applications, for our purposes our concern is only that some decision tree is returned. Regarding existence, it requires the argumentation frameworks to be acyclic. This can be achieved by restricting ourselves to coherent casebases, or instead applying $cAA\text{-}CBR_{\succeq}$, a variation of the original method which guarantees acyclic AFs (Paulino-Passos and Toni, 2021). For the rest of the section we will assume those conditions.

We can then easily instantiate the remainder of the framework, extending for sequences as previously suggested $\mathbb{E}(((x_i, y_i))_{i\in[n]}) = (\mathbb{E}_\bullet(x_i, y_i))_{i\in[n]}$, and defining $\models$ as: $(\mathcal{DT}_i)_{i\in[n]} \models (x, y)$ iff there is i such that $\mathcal{DT}_i$ can be readapted to x. Indeed, this readapted tree $\mathcal{DT}_x$ is precisely an explanation such that $\mathcal{DT}_i \vdash_e \mathcal{DT}_x$.

The intuition for this relation is a possible answer to the question: what

does an ADT explanation say about other inputs, beyond the original one being explained? Surely it also reveals how the cases which label the appearing nodes are decided, but would that be all? What we show from Theorem 5.14 is precisely that, for a new input case, if, when filtering the tree for only nodes labelled by relevant arguments, all leaves are W-nodes, then not only we can guarantee what the output is, but also acquire an explanation for it via a transformation. Also notice that this all is done without running the prediction algorithm entirely from scratch. How could this be used to accelerate inference time in practice and what is the impact is outside the scope of this chapter.

Another interesting aspect to notice is that, if there is a single L-node as a maximally relevant node, then nothing can be said. This can be seen from Figure 5.4. For a new case $\{\texttt{age:}19; \texttt{income:}110\}$, the only relevant case is the root (the default case), which is a L-case. However the prediction for it is still 1 since the case $(\{\texttt{age:}20; \texttt{income:}100\}, 1)$ is relevant, and would attack the default case in the grounded extension. The slightly different new case $\{\texttt{age:}21; \texttt{income:}110\}$ would have no such attacker.

We can then state the following consequences:

Theorem 5.16. *Let* $\mathbb{E}(((x_i, y_i))_{i\in[n]}) = (E_i)_{i\in[n]}$, *and* $(x, y) \in X \times Y$ *Then:*

1. $\mathbb{E}$ *is interaction-stable;*
2. $(E_i)_{i\in[n]} \models (x, y)$ *iff there is an explanation* E *such that* $(E_i)_{i\in[n]} \vdash_e E$ *and* $E \models (x, y)$*;*
3. *(faithfulness) if* $(E_i)_{i\in[n]} \models (x, y)$, *then* $AA\text{-}CBR_{\succeq}(D, x) = y$.
4. $\models$ *is consistent;*
5. $\models$ *is monotonic.*

Those are straightforward from $\models$ being defined from existence of an explanation in the sequence and from the fact that all arguments labelling W-nodes are contained in the grounded extension (Prop. 3.3 Čyras et al., 2019a).

ADT explanations are simple, but provably faithful explanations for $AA\text{-}CBR$. This should not be surprising, since this inference relation is conservative: a single L-node as a maximally relevant node in the tree makes the entire ADT unusable.

5.6 Discussion

Consistency properties have already been advocated in the literature, as discussed in the introduction. A related but different direction is using interactions and explanations for changing the model itself or predictions for a particular instance (Sreedharan et al., 2021; Rago et al., 2021a). In opposition, we consider here the model fixed. Amgoud (2021) also presents explanations as functions that can be non-monotonic, showing examples on how an explanation can be non-monotonic and proposing an argumentation-based approach to create non-monotonic explanations without losing consistency. Our main differences from this work are: i) we base our definitions on sequences, which are more general and can capture the interactive scenario; ii) we make explicit the entailment relations, linking explicitly to NMR (at both input-output and explanation levels).

A line of work with connections to this chapter as well is on building model interpretation (global explanations) from local explanations (Setzu et al., 2021; Zhou et al., 2022). These works generate local explanations for a set of instances and then by essentially merging those local explanations into a single, global, set of explanations. While here we model explanations at a higher level of abstraction and thus do not attack the same problem, the idea of aggregating local explanations can be captured in our abstract formalism. Indeed, one possible instantiation of $\models$ is by generating local explanations $(E_i)_{i\in[n]}$ for $(x_i, y_i)_{i\in[n]}$, and then aggregating the resulting explanations into a single, global, explanation E_{agg}, and saying $(E_i)_{i\in[n]} \models (x, y)$ exactly when $E_{agg} \models (x, y)$.

Regarding ADT readaptation as an explanation, a possible discussion is whether the entire set of ADTs for an input could not be returned instead. While this is a possibility, in the case of a practical application, even if all this information is to be used, it would still need to be adapted and filtered in a way not to overwhelm the final user. This is especially the case since many different dispute trees could have overlaps, being partly redundant. Depending on context, this could be less informative than having the entire argumentation framework, since there is no repetition of cases in it. Thus having a single ADT seems a more effective strategy, or at least a simpler one. The condition is, as

previously mentioned, that either the casebase is coherent, or that $cAA\text{-}CBR_{\succeq}$ is used (see Chapter 3), since there is also a correspondence in that given an outcome, there is always a dispute for explaining that decision (Čyras et al., 2019a, Prop. 3.4).

5.7 Summary

In this chapter, we presented a first approach to modelling an interactive explanation scenario, and discussed possible properties of it, specially ones based on non-monotonic reasoning (NMR). We also show how the general concept of specificity in NMR may be used to extend entailment from single explanations to sequences of explanations. Finally, we applied those definitions to ADTs as explanations for $AA\text{-}CBR$. We showed that under certain hypotheses ADTs can be readapted to new cases, and that this allows an interpretation of ADTs as reasoning that turns out to be monotonic.

We believe considerations inspired by non-monotonic reasoning can inspire questions and angles of study of explainability methods. We leave to future work analyses of specific systems and algorithms, as well as empirical evaluations of the impact of the presence or absence of particular NMR properties with human users. Another interesting avenue for future work is defining entailment for sequence of explanations by aggregating, such as in Section 5.4.1, but using other approaches. A possible one is seeing each explanation as an argument and using argumentation-based approaches for defining the entailment for sequence of explanations (Čyras et al., 2021; Amgoud, 2021; Baroni et al., 2018b).

We leave for future work an exploration of how would ADTs be for real applications of $AA\text{-}CBR$, including how often ADTs can be reused, and whether this can create gains in inference time. Another aspect left for future work is user studies for evaluating the effectiveness of ADT as explanations. An interesting question is whether ADTs with maximally relevant L-nodes could be readapted in some way, even if fallible. An idea would be considering those *prima facie* explanations, resulting in a non-monotonic entailment relation from sequences of explanations.

Chapter 6

Conclusion

In this thesis, we explored connections between non-monotonic reasoning (NMR) and both case-based reasoning (CBR) and explanations, particularly towards applications on legal reasoning.

We will summarise our contributions to each objective (Section 6.1) and discuss future work (Section 6.2).

6.1 Summary of contributions

In this thesis we contributed to an understanding across the fields of NMR, CBR, and AI and law in the following ways.

On the connections between CBR and NMR (Objective 1), and the possibility of analysing CBR using NMR criteria, we proposed a formal NMR analysis of classification models, with a focus on CBR systems (Chapter 3). We analysed an existing argumentation-based form of CBR called $AA\text{-}CBR$ this way, proving it does not satisfy the NMR properties of cautious monotonicity and cut, and we proposed a variant to the model, $cAA\text{-}CBR_{\succeq}$, that does satisfy them. We showed that $cAA\text{-}CBR_{\succeq}$ empowers a principled treatment of noise in data, that is, of incoherent casebases. We empirically compared $cAA\text{-}CBR_{\succeq}$ with the original formulation $AA\text{-}CBR_{\succeq}$ on the COMPAS recidivism dataset (Chapter 4). This is a dataset on recidivism prediction, an application of AI used in the real world in the judiciary with a strong impact on human life (Larson et al., 2016; Northpointe Inc., 2019; Rudin et al., 2020). Our use of this dataset is limited to evaluate our algorithms in a real-world legal dataset,

but it should be noted that this dataset contains many biases and many social risks are created by the naïve use of it for deployed systems (Bao et al., 2021). We showed that in this application, with a particular way of defining past case relevance, the new variant $cAA\text{-}CBR_{\succeq}$ and the original one have comparable or even identical performances, which implies NMR properties can be enforced with no loss in task performance. In some situations, using the new model can result in superior performance, particularly in the presence of noise. We also showed they result in argumentation frameworks with fewer arguments, which might be advantageous in their interpretability and explainability.

On Objective 2, of studying the effects of NMR properties in legal reasoning, we performed a case study on a legal casebase from the literature by adapting the result model of precedential constraint to $AA\text{-}CBR$ and verified that there was no difference in using the cautiously monotonic version (Chapter 3). We also showed that a legal reasoning system that obeys its own previous decisions can be manipulated if it does not satisfy cautious monotonicity.

On Objective 3, of the pertinence of CBR in the context of data-centric AI, we presented methods for learning the notion of relevance in CBR and empirically evaluated them (Chapter 4). We showed a method based on decision trees that performs comparably to decision trees themselves. Furthermore, we presented initial results on an approach using autoencoders, a type of neural networks, in order to allow the usage of CBR methods on unstructured data.

Finally, on Objective 4, of improving theoretical analysis of interactive explanations, we contributed by modelling an interactive explanation scenario with NMR, thus providing a formalisation, and proposing how NMR properties can be adapted to it (Chapter 5). We also applied this modelling to a type of explanations called arbitrated dispute trees for $AA\text{-}CBR$, showing how a notion of entailment can be defined on them. This objective has also connections to data-centric AI insofar interactive explanations are used in their deployment.

6.2 Open questions and future work

We have mentioned avenues for future work and open questions in previous chapters, which we now summarise and expand.

6.2.1 Classification and CBR models as NMR

Other classification and CBR models. We have proposed a general framework of analysis of classification models as NMR, and applied this approach to *AA-CBR*. This leaves open other forms of classification, including many other CBR models. Our theory in Chapter 3 in principle could be applied to other CBR systems used to perform classification. Neuro-symbolic applications would be of particular interest (Kenny and Keane, 2019; Li et al., 2018; Barnett et al., 2021), as well as applications to more specific contexts, such as knowledge graphs (Das et al., 2020a; Das et al., 2020b; Das et al., 2022).

Non-deterministic classifiers. A restriction in our approach is that it requires the classifier to be deterministic. One generalisation is straightforward: if a particular property holds for any execution of the entire process (equivalently, for any random seed), one can say it holds generally. But another interesting way of considering the question is probabilistic: what is the probability that a non-deterministic classifier satisfies, for instance, cautious monotonicity? This can be measured empirically: over many random trials, how many resulting classifiers satisfy it?

Generalisation of properties. This discussion also suggests the idea of generalising those properties themselves. We treat those properties strictly: if there is a single counterexample, the model is said not to have the property in general. However, a more gradual measure could be useful in practice, for instance, considering the question: given a classifier (even if deterministic), over a dataset, how often is cautious monotonicity violated? A proper formalisation of those concepts and empirical evaluation of particular systems or methods would be an interesting direction of future work.

User studies. Finally, we have worked with the assumption that the NMR properties at some level reflect human practical reasoning and thus are important for CBR. Despite having shown the negative impacts of the violation of some properties, e.g., in legal reasoning, by allowing manipulability (Section 3.6), a more general investigation of in which scenarios those properties are

important is still needed. Indeed, another important issue we have left open for future work is the empirical evaluation with humans (i.e. user studies) of the desirability of NMR properties in CBR. We expect this to require experiments in different scenarios in order to rigorously evaluate the question of how important are those properties in practice.

6.2.2 Comparisons between $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$

On $AA\text{-}CBR$, we have compared $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$ as per their NMR properties and its empirical performance on prediction of recidivism on COMPAS using the decision tree methodology. This does not close the question of how those models compare. We leave to future work more comparisons between $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$, including possible theoretical comparisons in learnability and time or space complexity, as well as empirical comparison on other settings, including other datasets and partial orders. Future work could also better compare $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$ in terms of differences in interpretability or explainability. Evaluating the impact of NMR properties in other CBR models with respect to those criteria would also be fruitful.

6.2.3 On legal CBR models

Other legal CBR models as NMR. We have analysed $AA\text{-}CBR$ as NMR and modelled precedential constraint on it, but doing so invites the natural question of how other legal CBR would behave in terms of NMR, which we have not explored. This includes the original formal dialogue game model of Prakken and Sartor (1998), the ASPIC$^+$-based argumentation scheme approach of Prakken et al. (2015b), and the ADF-based modelling of Al-Abdulkarim et al. (2016) and Al-Abdulkarim (2017). We have briefly discussed cautious monotonicity for precedential constraint in the rule model in the end of Section 3.7, but a more formal analysis, including more alternative models of precedential constraint and incorporating, for instance, issues and dimensions, would still be important (Rigoni, 2015; Bench-Capon and Atkinson, 2021; Horty, 2021; Prakken, 2021).

AA-CBR **for legal CBR.** Another open question about the connection between the precedential constraint and this work is how to enforce precedential constraint in *AA-CBR* without a (potential) exponential expansion as the one of Section 3.6. This adaptation is a limitation in terms of scalability for a large number of factors and cases. It is also an open question how to adapt *AA-CBR* for issues and dimensions. Finally, many legal CBR approaches require (manual) knowledge engineering, which is labour-intensive. This also applies for *AA-CBR* with enforced precedential constraints. Developing models that overcome this challenge is still open work, for which the learning of partial orders for *AA-CBR* could be a promising approach.

6.2.4 Learning of partial orders

For the learning of partial orders in $AA\text{-}CBR_{\succeq}$, we have seen in Chapter 4 that feature splits from decision trees can be used to define a partial order, and we have shown initial results on using autoencoders. Those results are still preliminary and raise the question on how to achieve higher performance combining neural networks with $AA\text{-}CBR_{\succeq}$. That line of exploration is still open regarding different neural network techniques, including multi-task learning (Caruana, 1997; Ruder, 2017; Crawshaw, 2020), for using output information as a bias for creating potentially more useful hidden encodings for inducing the partial order. A clear direction for future work is exploring this methodology in unstructured data, such as images and text (including legal text, such as legal decisions), where such an approach could be more advantageous. Furthermore, developing a neural-network-based methodology such that a learned partial order (or more generally, a learned notion of relevance in CBR) is interpretable is still an open issue. This has connections to work that aims on ascribing previously built factors, an already interpretable form, to textual description of facts (Mumford et al., 2022), or uses data in combination with other types of knowledge in order to execute CBR (Grabmair, 2017b; Grabmair, 2016; Grabmair, 2017b).

6.2.5 Interactive explanations as NMR

On interactive explanations as NMR, we have presented a general framework and applied this modelling to arbitrated dispute trees for *AA-CBR*. However, this only scratches the surface of possibilities, since we do not apply or adapt the methodology to different explanation methods, including explanatory local rules (Ribeiro et al., 2018), prime implicants (Ignatiev et al., 2021), and feature attribution methods (Štrumbelj and Kononenko, 2014; Ribeiro et al., 2016a; Lundberg and Lee, 2017). We discuss examples with counterfactuals (Wachter et al., 2018), but we do not provide general ways of interpreting counterfactuals in our framework, an issue which is open for future work.

Thus we have here proposed a vocabulary for discussion, but we envisage future work analysing 1) how other explainable AI (XAI) methods can be seen as interactive, or adapted to an interactive scenario; and 2) what notions of (non-monotonic) entailment could be defined on those methods, and which properties would this reasoning have. Finally, empirical evaluations with users is still an essential requirement to ascertain the importance of NMR properties in the context of explainability.

6.3 Concluding remarks

Non-monotonic reasoning has been proposed as a way of capturing commonsense, human practical reasoning, including legal reasoning. Despite the contributions of this field in formalising intuitions about human reasoning, many of its insights have not been more broadly considered in AI. In this thesis, we start from the working hypothesis of a wider importance of non-monotonic reasoning insights in order to bring these lenses of analysis to other fields in AI. The application of these lenses allows both i) original methods of analyses of models to human-like reasoning; ii) the development of new models that satisfy properties considered essential for commonsense reasoning. We do so for two different subareas of AI: classification, with a focus on case-based reasoning; and interactive explanation models in explainable AI. We highlight law as an application domain for both case-based reasoning and explainable AI, as it is a domain a) with close ties to case-based reasoning; b) with stronger requirements of transparency, explainability and alignment to human values;

c) in which guarantees on systems are more strongly necessary. Indeed, despite their limitations in, e.g., scalability, models presented in this work (and much of the literature on which it draws) can be proved to behave in particular ways, and we have done so, for instance, by showing how precedential constraint can be enforced, or how reasoning properties can be certified. This is in striking opposition to recent developments, such as large language models, which are very flexible models, with surprisingly fluent outputs, but for which no guarantee whatsoever on their behaviours can be offered. We urge researchers and practitioners in AI in general, and in particular AI and law and legal technology, to consider the importance of assurances of conformance to human practical reasoning in their proposed systems.

Bibliography

A. C. Kakas, R. A. Kowalski and F. Toni (1998). 'The Role of Abduction in Logic Programming'. In: *Handbook of Logic in Artificial Intelligence and Logic Programming - Volume 5 - Logic Programming*. Ed. by Dov M. Gabbay, C. J. Hogger and J. A. Robinson. Oxford University Press (cit. on p. 31).

Aamodt, Agnar (1993). 'A Case-Based Answer to Some Problems of Knowledge-Based Systems'. In: *Fourth Scandinavian Conference on Artificial Intelligence, SCAI 1993, Stockholm, Sweden, May 4-7, 1993*. Ed. by Erik Sandewall and Carl Gustaf Jansson. Vol. 18. Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 168–182. ISBN: 90-5199-134-7 (cit. on p. 40).

Aamodt, Agnar and Enric Plaza (1994). 'Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches'. In: *AI Commun.* 7.1, pp. 39–59.
DOI: 10.3233/AIC-1994-7104. URL: https://doi.org/10.3233/AIC-1994-7104 (cit. on p. 33).

Al-Abdulkarim, Latifa (2017). 'Representation of case law for argumentative reasoning'. PhD thesis. University of Liverpool, UK. URL: http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.713388 (cit. on pp. 36, 40, 90, 150).

Al-Abdulkarim, Latifa, Katie Atkinson and Trevor J. M. Bench-Capon (June 2015a). 'Evaluating the use of abstract dialectical frameworks to represent case law'. In: *Proceedings of the 15th International Conference on Artificial Intelligence and Law*. DOI: 10.1145/2746090.2746111. URL: http://dx.doi.org/10.1145/2746090.2746111 (cit. on pp. 36, 40, 90).

— (2015b). 'Factors, issues and values: revisiting reasoning with cases'. In: *Proceedings of the 15th International Conference on Artificial Intelligence and Law* (cit. on pp. 36, 121).

Al-Abdulkarim, Latifa, Katie Atkinson and Trevor J. M. Bench-Capon (Mar. 2016). 'A methodology for designing systems to reason with legal cases using Abstract Dialectical Frameworks'. In: *Artificial Intelligence and Law* 24.1, pp. 1–49. ISSN: 1572-8382. DOI: 10.1007/s10506-016-9178-1. URL: http://dx.doi.org/10.1007/s10506-016-9178-1 (cit. on pp. 94, 150).

Abu-Mostafa, Yaser S., Malik Magdon-Ismail and Hsuan-Tien Lin (2012). *Learning From Data* (cit. on pp. 60, 71, 128).

Adadi, Amina and Mohammed Berrada (2018). 'Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)'. In: *IEEE Access* 6, pp. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052. URL: https://doi.org/10.1109/ACCESS.2018.2870052 (cit. on p. 37).

Albini, Emanuele, Antonio Rago, Pietro Baroni and Francesca Toni (2020). 'Relation-Based Counterfactual Explanations for Bayesian Network Classifiers'. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. Ed. by Christian Bessiere. ijcai.org, pp. 451–457.
DOI: 10.24963/ijcai.2020/63. URL: https://doi.org/10.24963/ijcai.2020/63 (cit. on p. 39).

Alchourrón, Carlos E. (1993). 'Philosophical foundations of Deontic Logic and the Logic of Defeasible Conditionals'. In: *Deontic Logic in Computer Science: Normative System Specification*. Ed. by J-J. Ch. Meyer and J.R. Wieringa. John Wiley & Sons (cit. on p. 32).

— (Dec. 1996). 'On Law and Logic'. In: *Ratio Juris* 9.4, pp. 331–348. ISSN: 1467-9337. DOI: 10.1111/j.1467-9337.1996.tb00250.x. URL: http://dx.doi.org/10.1111/j.1467-9337.1996.tb00250.x (cit. on p. 32).

Alchourrón, Carlos E., Peter Gärdenfors and David Makinson (1985). 'On the logic of theory change: Partial meet contraction and revision functions'. In: *Journal of Symbolic Logic* 50, pp. 510–530 (cit. on p. 32).

Aleven, Vincent (Nov. 2003a). 'Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment'. In: *Artificial Intelligence* 150.1-2, pp. 183–237. ISSN: 0004-3702. DOI: 10.1016/s0004-3702(03)00105-x. URL: http://dx.doi.org/10.1016/s0004-3702(03)00105-x (cit. on p. 35).

— (2003b). 'Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment'. In: *Artif. Intell.* 150.1-2, pp. 183–237. DOI: 10.1016/S0004-3702(03)00105-X. URL: https://doi.org/10.1016/S0004-3702(03)00105-X (cit. on p. 94).

Aleven, Vincent and Kevin D. Ashley (1997). 'Evaluating a Learning Environment for Case-Based Argumentation Skills'. In: *Proceedings of the Sixth International Conference on Artificial Intelligence and Law, ICAIL '97, Melbourne, Vicoria, Australia, June 30 - July 3, 1997*. Ed. by John Zeleznikow, Daniel Hunter and Karl Branting. ACM, pp. 170–179.
DOI: 10.1145/261618.261650. URL: https://doi.org/10.1145/261618.261650 (cit. on p. 35).

Algero, Mary Garvey (2004). 'The Sources of Law and the Value of Precedent: A Comparative and Empirical Study of a Civil Law State in a Common Law Nation'. In: *La. L. Rev.* 65, p. 775 (cit. on pp. 27, 57).

Amgoud, Leila (2021). 'Non-monotonic Explanation Functions'. In: *Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 16th European Conference, ECSQARU 2021, Prague, Czech Republic, September 21-24, 2021, Proceedings*. Ed. by Jirina Vejnarová and Nic Wilson. Vol. 12897. Lecture Notes in Computer Science. Springer, pp. 19–31. ISBN: 978-3-030-86771-3.
DOI: 10.1007/978-3-030-86772-0_2. URL: https://doi.org/10.1007/978-3-030-86772-0_2 (cit. on pp. 39, 144, 145).

Amgoud, Leila and Vivien Beuselinck (2022). 'Towards a Principle-Based Approach for Case-Based Reasoning'. In: *Scalable Uncertainty Management*. Ed. by Florence Dupin de Saint-Cyr, Meltem Öztürk-Escoffier and Nico Potyka. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 37–46. ISBN: 978-3-031-18843-5.
DOI: 10.1007/978-3-031-18843-5_3 (cit. on p. 96).

Amgoud, Leila, Claudette Cayrol, Marie-Christine Lagasquie-Schiex and P. Livet (2008). 'On bipolarity in argumentation frameworks'. In: *Int. J. Intell. Syst.* 23.10, pp. 1062–1093.
DOI: 10.1002/int.20307. URL: https://doi.org/10.1002/int.20307 (cit. on p. 32).

An, Jinwon and Sungzoon Cho (2015). *Variational Autoencoder based Anomaly Detection using Reconstruction Probability*. Technical Report. Seoul National University Data Mining Center. URL: http://dm.snu.ac.kr/static/docs/TR/SNUDM-TR-2015-03.pdf (cit. on p. 61).

Angelino, Elaine, Nicholas Larus-Stone, Daniel Alabi, Margo I. Seltzer and Cynthia Rudin (2017). 'Learning Certifiably Optimal Rule Lists for Categorical Data'. In: *J. Mach. Learn. Res.* 18, 234:1–234:78. URL: http://jmlr.org/papers/v18/17-716.html (cit. on p. 109).

Arieli, Ofer, AnneMarie Borg, Jesse Heyninck and Christian Straßer (2021). 'Logic-Based Approaches to Formal Argumentation'. In: *Journal of Applied Logics - IfCoLog Journal of Logics and their Applications* 8.6, pp. 1793–1898. URL: https://collegepublications.co.uk/ifcolog/?00048 (cit. on p. 32).

Ashley, K. D. (1989). 'Toward a Computational Theory of Arguing with Precedents'. In: *Proceedings of the 2nd International Conference on Artificial Intelligence and Law*. ICAIL '89. Vancouver, British Columbia, Canada: Association for Computing Machinery, pp. 93–102. ISBN: 0897913221. DOI: 10.1145/74014.74028. URL: https://doi.org/10.1145/74014.74028 (cit. on pp. 35, 59).

Ashley, Kevin D. (June 1991). 'Reasoning with cases and hypotheticals in HYPO'. In: *International Journal of Man-Machine Studies* 34.6, pp. 753–796. ISSN: 0020-7373. DOI: 10.1016/0020-7373(91)90011-u. URL: http://dx.doi.org/10.1016/0020-7373(91)90011-u (cit. on pp. 35, 93).

Ashley, Kevin D. and Stefanie Brüninghaus (2009). 'Automatically classifying case texts and predicting outcomes'. In: *Artif. Intell. Law* 17.2, pp. 125–165. DOI: 10.1007/s10506-009-9077-9. URL: https://doi.org/10.1007/s10506-009-9077-9 (cit. on p. 94).

Athakravi, Duangtida, Ken Satoh, Mark Law, Krysia Broda and Alessandra Russo (2015). 'Automated Inference of Rules with Exception from Past Legal Cases Using ASP'. In: *Logic Programming and Nonmonotonic Reasoning - 13th International Conference, LPNMR 2015, Lexington, KY, USA, September 27-30, 2015. Proceedings*. Ed. by Francesco Calimeri, Giovambattista Ianni and Miroslaw Truszczynski. Vol. 9345. Lecture Notes in Computer Science. Springer, pp. 83–96. DOI: 10.1007/978-3-319-23264-5_8. URL: https://doi.org/10.1007/978-3-319-23264-5%5C_8 (cit. on p. 67).

Atkinson, Katie and Trevor J. M. Bench-Capon (2021). 'Value-based Argumentation'. In: *Journal of Applied Logics - IfCoLog Journal of Logics and their Applications* 8.6, pp. 1543–1588. URL: https://collegepublications.co.uk/ifcolog/?00048 (cit. on p. 33).

Atkinson, Katie, Trevor J. M. Bench-Capon and Danushka Bollegala (2020). 'Explanation in AI and law: Past, present and future'. In: *Artif. Intell.* 289, p. 103387. DOI: 10.1016/j.artint.2020.103387. URL: https://doi.org/10.1016/j.artint.2020.103387 (cit. on pp. 38, 62, 126).

Bao, Michelle, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum and Suresh Venkatasubramanian (2021). 'It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks'. In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*. Ed. by Joaquin Vanschoren and Sai-Kit Yeung. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/92cc227532d17e56e07902b254dfad10-Abstract-round1.html (cit. on pp. 109, 148).

Barenstein, Matias (2019). 'ProPublica's COMPAS Data Revisited'. In: *ArXiv* abs/1906.04711 (cit. on p. 109).

Barnett, Alina Jade, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y. Lo and Cynthia Rudin (2021). 'A case-based interpretable deep learning model for classification of mass lesions in digital mammography'. In: *Nat. Mach. Intell.* 3.12, pp. 1061–1070. DOI: 10.1038/s42256-021-00423-x. URL: https://doi.org/10.1038/s42256-021-00423-x (cit. on pp. 35, 121, 149).

Baroni, Pietro, Martin Caminada and Massimiliano Giacomin (2018a). 'Abstract Argumentation Frameworks and Their Semantics'. In: *Handbook of Formal Argumentation*. Ed. by Pietro Baroni, Dov Gabbay, Massimiliano Giacomin and Leendert van der Torre. College Publications. ISBN: 1848902751 (cit. on pp. 51, 52).

Baroni, Pietro, Dov Gabbay, Massimiliano Giacomin and Leendert van der Torre, eds. (2018b). *Handbook of Formal Argumentation*. College Publications. ISBN: 1848902751 (cit. on pp. 26, 145).

Bell, John (1997). 'Comparing Precedent'. In: *Cornell Law Review* 82, pp. 1243–1278 (cit. on pp. 27, 57).

Bench-Capon, Trevor (Aug. 2022). 'Thirty years of Artificial Intelligence and Law: Editor's Introduction'. In: *Artificial Intelligence and Law* 30.4, pp. 475–479. ISSN: 1572-8382. DOI: 10.1007/s10506-022-09325-8. URL: http://dx.doi.org/10.1007/s10506-022-09325-8 (cit. on pp. 25, 26).

Bench-Capon, Trevor J. M. (2002). 'Value-Based Argumentation Frameworks'. In: *CoRR* cs.AI/0207059. URL: https://arxiv.org/abs/cs/0207059 (cit. on p. 33).

— (2017). 'HYPO'S legacy: introduction to the virtual special issue'. In: *Artif. Intell. Law* 25.2, pp. 205–250. DOI: 10.1007/s10506-017-9201-1. URL: https://doi.org/10.1007/s10506-017-9201-1 (cit. on pp. 35, 43, 89, 93).

— (2020). 'Before and after Dung: Argumentation in AI and Law'. In: *Argument Comput.* 11.1-2, pp. 221–238. DOI: 10.3233/AAC-190477. URL: https://doi.org/10.3233/AAC-190477 (cit. on pp. 32, 35).

Bench-Capon, Trevor J. M., Michal Araszkiewicz, Kevin D. Ashley, Katie Atkinson, Floris Bex, Filipe Borges, Danièle Bourcier, Paul Bourgine, Jack G. Conrad, Enrico Francesconi, Thomas F. Gordon, Guido Governatori, Jochen L. Leidner, David D. Lewis, Ronald Prescott Loui, L. Thorne McCarty, Henry Prakken, Frank Schilder, Erich Schweighofer, Paul Thompson, Alex Tyrrell, Bart Verheij, Douglas N. Walton and Adam Z. Wyner (2012). 'A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law'. In: *Artif. Intell. Law* 20.3, pp. 215–319. DOI: 10.1007/s10506-012-9131-x. URL: https://doi.org/10.1007/s10506-012-9131-x (cit. on p. 25).

Bench-Capon, Trevor J. M. and Katie Atkinson (2021). 'Precedential constraint: the role of issues'. In: *ICAIL '21: Eighteenth International Conference for Artificial Intelligence and Law, São Paulo Brazil, June 21 - 25, 2021*. Ed. by Juliano Maranhão and Adam Zachary Wyner. ACM, pp. 12–21. ISBN: 978-1-4503-8526-8. DOI: 10.1145/3462757.3466062. URL: https://doi.org/10.1145/3462757.3466062 (cit. on pp. 36, 56, 57, 59, 150).

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major and Shmargaret Shmitchell (3rd Mar. 2021). 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜'. In: *Proceedings of the 2021 ACM Con-*

ference on Fairness, Accountability, and Transparency, pp. 610–623. DOI: 10.1145/3442188.3445922. URL: https://dl.acm.org/doi/10.1145/3442188.3445922 (visited on 09/11/2023) (cit. on p. 36).

Bender, Emily M. and Alexander Koller (2020). 'Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, pp. 5185–5198. DOI: 10.18653/v1/2020.acl-main.463. URL: https://www.aclweb.org/anthology/2020.acl-main.463 (visited on 09/11/2023) (cit. on p. 36).

Besnard, Philippe, Alejandro Javier García, Anthony Hunter, Sanjay Modgil, Henry Prakken, Guillermo Ricardo Simari and Francesca Toni (2014). 'Introduction to structured argumentation'. In: *Argument Comput.* 5.1, pp. 1–4. DOI: 10.1080/19462166.2013.869764. URL: https://doi.org/10.1080/19462166.2013.869764 (cit. on p. 31).

Besnard, Philippe and Anthony Hunter (2008). *Elements of Argumentation*. MIT Press. ISBN: 978-0-262-02643-7. URL: http://mitpress.mit.edu/books/elements-argumentation (cit. on p. 32).

Bishop, Christopher M. (2007). *Pattern recognition and machine learning, 5th Edition*. Information science and statistics. Springer. ISBN: 9780387310732. URL: https://www.worldcat.org/oclc/71008143 (cit. on pp. 33, 59, 60).

Bondarenko, Andrei, Phan Minh Dung, Robert A. Kowalski and Francesca Toni (1997). 'An Abstract, Argumentation-Theoretic Approach to Default Reasoning'. In: *Artif. Intell.* 93, pp. 63–101. DOI: 10.1016/S0004-3702(97)00015-5. URL: https://doi.org/10.1016/S0004-3702(97)00015-5 (cit. on p. 31).

Bondarenko, Andrei, Francesca Toni and Robert A. Kowalski (1993). 'An Assumption-Based Framework for Non-Monotonic Reasoning'. In: *Logic Programming and Non-monotonic Reasoning, Proceedings of the Second International Workshop, Lisbon, Portugal, June 1993*. Ed. by Luís Moniz Pereira and Anil Nerode. MIT Press, pp. 171–189. ISBN: 0-262-66083-0 (cit. on p. 31).

Bowman, Samuel R. (2023). 'Eight Things to Know about Large Language Models'. Version 1. In: DOI: 10.48550/ARXIV.2304.00612. URL: https://arxiv.org/abs/2304.00612 (visited on 09/11/2023) (cit. on p. 36).

Branting, Karl (2017). 'Data-centric and logic-based models for automated legal problem solving'. In: *Artif. Intell. Law* 25.1, pp. 5–27. DOI: 10.1007/s10506-017-9193-x. URL: https://doi.org/10.1007/s10506-017-9193-x (cit. on p. 26).

Breiman, Leo, J. H. Friedman, R. A. Olshen and C. J. Stone (1984). *Classification and Regression Trees*. Wadsworth. ISBN: 0-534-98053-8 (cit. on pp. 60, 61, 101).

Brewka, Gerhard (1991). *Nonmonotonic reasoning - logical foundations of commonsense*. Vol. 12. Cambridge tracts in theoretical computer science. Cambridge University Press. ISBN: 978-0-521-38394-3 (cit. on p. 31).

Brewka, Gerhard, Paul E. Dunne and Stefan Woltran (2011). 'Relating the Semantics of Abstract Dialectical Frameworks and Standard AFs'. In: *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*. Ed. by Toby Walsh. IJCAI/AAAI, pp. 780–785. ISBN: 978-1-57735-516-8. DOI: 10.5591/978-1-57735-516-8/IJCAI11-137. URL: https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-137 (cit. on p. 32).

Brewka, Gerhard and Stefan Woltran (2010). 'Abstract Dialectical Frameworks'. In: *Principles of Knowledge Representation and Reasoning: Proceedings of the Twelfth International Conference, KR 2010, Toronto, Ontario, Canada, May 9-13, 2010*. Ed. by Fangzhen Lin, Ulrike Sattler and Miroslaw Truszczynski. AAAI Press. URL: http://aaai.org/ocs/index.php/KR/KR2010/paper/view/1294 (cit. on p. 32).

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. Henighan, Rewon Child, A. Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, S. Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever and Dario Amodei (28th May 2020). 'Language Models Are Few-Shot Learners'. In: *ArXiv*. URL: https://www.semanticscholar.

org / paper / Language - Models - are - Few - Shot - Learners - Brown - Mann / 6b85b63579a916f705a8e10a49bd8d849d91b1fc (visited on 09/11/2023) (cit. on p. 36).

Brüninghaus, Stefanie and Kevin D. Ashley (2003). 'Predicting Outcomes of Case-Based Legal Arguments'. In: *Proceedings of the 9th International Conference on Artificial Intelligence and Law, ICAIL 2003, Edinburgh, Scotland, UK, June 24-28, 2003*. Ed. by John Zeleznikow and Giovanni Sartor. ACM, pp. 233–242. ISBN: 1-58113-747-8. DOI: 10.1145/1047788.1047838. URL: https://doi.org/10.1145/1047788.1047838 (cit. on pp. 43, 89).

Bulygin, Eugenio (2003). 'Review of Jaap Hage's Law and Defeasibility'. In: *Artif. Intell. Law* 11.2-3, pp. 245–250. DOI: 10.1023/B:ARTI.0000046012.44321.27. URL: https://doi.org/10.1023/B:ARTI.0000046012.44321.27 (cit. on p. 33).

Byrne, Ruth M.J. (Feb. 1989). 'Suppressing valid inferences with conditionals'. In: *Cognition* 31.1, pp. 61–83. ISSN: 0010-0277. DOI: 10.1016/0010-0277(89)90018-8. URL: http://dx.doi.org/10.1016/0010-0277(89)90018-8 (cit. on pp. 30, 127, 133).

Camburu, Oana-Maria, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz and Phil Blunsom (July 2020). 'Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4157–4165. DOI: 10.18653/v1/2020.acl-main.382. URL: https://aclanthology.org/2020.acl-main.382 (cit. on pp. 125, 126).

Caminada, Martin (2006). 'On the Issue of Reinstatement in Argumentation'. In: *Logics in Artificial Intelligence, 10th European Conference, JELIA 2006, Liverpool, UK, September 13-15, 2006, Proceedings*. Ed. by Michael Fisher, Wiebe van der Hoek, Boris Konev and Alexei Lisitsa. Vol. 4160. Lecture Notes in Computer Science. Springer, pp. 111–123. ISBN: 3-540-39625-X. DOI: 10.1007/11853886_11. URL: https://doi.org/10.1007/11853886%5C_11 (cit. on pp. 51, 70).

— (2007). 'An Algorithm for Computing Semi-stable Semantics'. In: *Symbolic and Quantitative Approaches to Reasoning with Uncertainty, 9th European Conference, ECSQARU 2007, Hammamet, Tunisia, October 31 - November*

2, 2007, Proceedings. Ed. by Khaled Mellouli. Vol. 4724. Lecture Notes in Computer Science. Springer, pp. 222–234. ISBN: 978-3-540-75255-4. DOI: 10.1007/978-3-540-75256-1_22. URL: https://doi.org/10.1007/978-3-540-75256-1%5C_22 (cit. on pp. 51, 70).

Canavotto, Ilaria (2022). 'Precedential Constraint Derived from Inconsistent Case Bases'. In: *Legal Knowledge and Information Systems*. IOS Press, pp. 23–32. DOI: 10.3233/FAIA220445. URL: https://ebooks.iospress.nl/doi/10.3233/FAIA220445 (visited on 12/11/2023) (cit. on pp. 56, 59).

Canavotto, Ilaria and John Horty (19th June 2023). 'Reasoning with Hierarchies of Open-Textured Predicates'. In: *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*. ICAIL 2023: Nineteenth International Conference on Artificial Intelligence and Law. Braga Portugal: ACM, pp. 52–61. ISBN: 9798400701979. DOI: 10.1145/3594536.3595148. URL: https://dl.acm.org/doi/10.1145/3594536.3595148 (visited on 12/11/2023) (cit. on p. 59).

Caruana, Rich (1997). 'Multitask Learning'. In: *Mach. Learn.* 28.1, pp. 41–75. DOI: 10.1023/A:1007379606734. URL: https://doi.org/10.1023/A:1007379606734 (cit. on p. 151).

Casella, George and Roger L. Berger (2002). *Statistical inference*. 2nd ed. Duxbury Press. ISBN: 0534243126 (cit. on p. 103).

Cayrol, Claudette and Marie-Christine Lagasquie-Schiex (2005). 'On the Acceptability of Arguments in Bipolar Argumentation Frameworks'. In: *Symbolic and Quantitative Approaches to Reasoning with Uncertainty, 8th European Conference, ECSQARU 2005, Barcelona, Spain, July 6-8, 2005, Proceedings*. Ed. by Lluís Godo. Vol. 3571. Lecture Notes in Computer Science. Springer, pp. 378–389. ISBN: 3-540-27326-3. DOI: 10.1007/11518655_33. URL: https://doi.org/10.1007/11518655%5C_33 (cit. on p. 32).

Cerutti, Federico, Marcos Cramer, Mathieu Guillaume, Emmanuel Hadoux, Anthony Hunter and Sylwia Polberg (2021). 'Empirical cognitive studies about formal argumentation'. In: *Handbook of Formal Argumentation, Volume 2*. Ed. by Guillermo R. Simari Dov Gabbay Massimiliano Giacomin and Matthias Thimm. 2 vols. College Publications. ISBN: 1848903367 (cit. on p. 133).

Chen, Chaofan, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin and Jonathan Su (2019). 'This Looks Like That: Deep Learning for Interpretable Image Recognition'. In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox and Roman Garnett, pp. 8928–8939. URL: https://proceedings.neurips.cc/paper/2019/hash/adf7ee2dcf142b0e11888e72b43fcb75-Abstract.html (cit. on pp. 35, 121).

Chorley, Alison and Trevor J. M. Bench-Capon (2005). 'AGATHA: Using heuristic search to automate the construction of case law theories'. In: *Artif. Intell. Law* 13.1, pp. 9–51. DOI: 10.1007/s10506-006-9004-2. URL: https://doi.org/10.1007/s10506-006-9004-2 (cit. on p. 90).

Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le and Jason Wei (2022). 'Scaling Instruction-Finetuned Language Models'. Version 5. In: DOI: 10.48550/ARXIV.2210.11416. URL: https://arxiv.org/abs/2210.11416 (visited on 09/11/2023) (cit. on p. 36).

Clark, Keith L. (1977). 'Negation as Failure'. In: *Logic and Data Bases, Symposium on Logic and Data Bases, Centre d'études et de recherches de Toulouse, France, 1977*. Ed. by Hervé Gallaire and Jack Minker. Advances in Data Base Theory. New York: Plemum Press, pp. 293–322. ISBN: 0-306-40060-X.

DOI: 10.1007/978-1-4684-3384-5_11. URL: https://doi.org/10.1007/978-1-4684-3384-5%5C_11 (cit. on p. 31).

Cocarascu, Oana, Kristijonas Čyras and Francesca Toni (2018). 'Explanatory predictions with artificial neural networks and argumentation'. In: *2nd Workshop on XAI at the 27th IJCAI and the 23rd ECAI* (cit. on pp. 34, 39, 100).

Cocarascu, Oana, Antonio Rago and Francesca Toni (2020a). 'Explanation via Machine Arguing'. In: *Reasoning Web. Declarative Artificial Intelligence - 16th International Summer School 2020, Oslo, Norway, June 24-26, 2020, Tutorial Lectures*. Ed. by Marco Manna and Andreas Pieris. Vol. 12258. Lecture Notes in Computer Science. Springer, pp. 53–84. DOI: 10.1007/978-3-030-60067-9_3. URL: https://doi.org/10.1007/978-3-030-60067-9%5C_3 (cit. on p. 38).

Cocarascu, Oana, Andria Stylianou, Kristijonas Čyras and Francesca Toni (2020b). 'Data-Empowered Argumentation for Dialectically Explainable Predictions'. In: *ECAI 2020 - 24th European Conference on Artificial Intelligence, Santiago de Compostela, Spain, 10-12 June 2020* (cit. on pp. 34, 39, 53–55, 67, 74, 94, 97, 99, 100).

Cover, Thomas M. and Peter E. Hart (1967). 'Nearest neighbor pattern classification'. In: *IEEE Trans. Inf. Theory* 13.1, pp. 21–27. DOI: 10.1109/TIT.1967.1053964. URL: https://doi.org/10.1109/TIT.1967.1053964 (cit. on p. 33).

Cramer, Harald (1946). *Mathematical methods of statistics*. Nineteenth printing, 1999. Princeton University Press. ISBN: 0691005478,9780691005478 (cit. on pp. 103, 104).

Crawshaw, Michael (2020). *Multi-Task Learning with Deep Neural Networks: A Survey*. arXiv: 2009.09796v1 [cs.LG] (cit. on p. 151).

Čyras, Kristijonas, David Birch, Yike Guo, Francesca Toni, Rajvinder Dulay, Sally Turvey, Daniel Greenberg and Tharindi Hapuarachchi (2019a). 'Explanations by arbitrated argumentative dispute'. In: *Expert Syst. Appl.* 127, pp. 141–156 (cit. on pp. 39, 63, 64, 136, 139, 142, 143, 145).

Čyras, Kristijonas, Timotheus Kampik, Oana Cocarascu and Antonio Rago, eds. (2022). *1st International Workshop on Argumentation for eXplainable AI co-located with 9th International Conference on Computational Models of Argument (COMMA 2022), Cardiff, Wales, September 12, 2022*. Vol. 3209. CEUR Workshop Proceedings. CEUR-WS.org. URL: http://ceur-ws.org/Vol-3209.

Čyras, Kristijonas, Amin Karamlou, Myles Lee, Dimitrios Letsios, Ruth Misener and Francesca Toni (2020). 'AI-assisted Schedule Explainer for Nurse Rostering'. In: *Proceedings of the 19th International Conference on Autonom-*

ous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020. Ed. by Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An and Neil Yorke-Smith. International Foundation for Autonomous Agents and Multiagent Systems, pp. 2101–2103. DOI: 10.5555/3398761.3399089. URL: https://dl.acm.org/doi/10.5555/3398761.3399089 (cit. on p. 39).

Čyras, Kristijonas, Dimitrios Letsios, Ruth Misener and Francesca Toni (2019b). 'Argumentation for Explainable Scheduling'. In: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, pp. 2752–2759. DOI: 10.1609/aaai.v33i01.33012752. URL: https://doi.org/10.1609/aaai.v33i01.33012752 (cit. on p. 39).

Čyras, Kristijonas, Antonio Rago, Emanuele Albini, Pietro Baroni and Francesca Toni (2021). 'Argumentative XAI: A Survey'. In: *ArXiv* abs/2105.11266 (cit. on p. 145).

Čyras, Kristijonas, Ken Satoh and Francesca Toni (2016a). 'Abstract Argumentation for Case-Based Reasoning'. In: *KR 2016*, pp. 549–552 (cit. on pp. 34, 53, 54, 67, 74, 97).

— (2016b). 'Explanation for Case-Based Reasoning via Abstract Argumentation'. In: *Proceedings of COMMA 2016*, pp. 243–254 (cit. on pp. 34, 63, 67, 126, 136).

Čyras, Kristijonas and Francesca Toni (2015). 'Non-Monotonic Inference Properties for Assumption-Based Argumentation'. In: *Theory and Applications of Formal Argumentation - Third International Workshop, TAFA 2015, Buenos Aires, Argentina, July 25-26, 2015, Revised Selected Papers*. Ed. by Elizabeth Black, Sanjay Modgil and Nir Oren. Vol. 9524. Lecture Notes in Computer Science. Springer, pp. 92–111. ISBN: 978-3-319-28459-0. DOI: 10.1007/978-3-319-28460-6_6. URL: https://doi.org/10.1007/978-3-319-28460-6%5C_6 (cit. on pp. 32, 68).

— (2016). 'Properties of ABA+ for Non-Monotonic Reasoning'. In: *CoRR* abs/1603.08714. arXiv: 1603.08714. URL: http://arxiv.org/abs/1603.08714 (cit. on pp. 32, 68).

Das, Rajarshi, Ameya Godbole, Shehzaad Dhuliawala, Manzil Zaheer and Andrew McCallum (2020a). 'A Simple Approach to Case-Based Reasoning in Knowledge Bases'. In: *Conference on Automated Knowledge Base Construction, AKBC 2020, Virtual, June 22-24, 2020*. Ed. by Dipanjan Das, Hannaneh Hajishirzi, Andrew McCallum and Sameer Singh.
DOI: 10.24432/C52S3K. URL: https://doi.org/10.24432/C52S3K (cit. on pp. 35, 149).

Das, Rajarshi, Ameya Godbole, Nicholas Monath, Manzil Zaheer and Andrew McCallum (2020b). 'Probabilistic Case-based Reasoning in Knowledge Bases'. In: *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*. Ed. by Trevor Cohn, Yulan He and Yang Liu. Vol. EMNLP 2020. Findings of ACL. Association for Computational Linguistics, pp. 4752–4765.
DOI: 10.18653/v1/2020.findings-emnlp.427. URL: https://doi.org/10.18653/v1/2020.findings-emnlp.427 (cit. on pp. 35, 149).

Das, Rajarshi, Ameya Godbole, Ankita Naik, Elliot Tower, Manzil Zaheer, Hannaneh Hajishirzi, Robin Jia and Andrew Mccallum (July 2022). 'Knowledge Base Question Answering by Case-based Reasoning over Subgraphs'. In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu and Sivan Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 4777–4793. URL: https://proceedings.mlr.press/v162/das22a.html (cit. on pp. 35, 149).

Davey, Brian A. and Hilary A. Priestley (2002). *Introduction to Lattices and Order, Second Edition*. Cambridge University Press. ISBN: 978-0-521-78451-1. DOI: 10.1017/CBO9780511809088. URL: https://doi.org/10.1017/CBO9780511809088 (cit. on pp. 54, 105).

David, René and John E. C. Brierley (1988). *Major Legal Systems In The World Today. An Introduction to the Comparative Study of Law*. The Legal Classics Library. URL: https://archive.org/details/majorlegalsystem00davi (cit. on p. 27).

Defense Advanced Research Projects Agency (2016). *Broad Agency Announcement - Explainable Artificial Intelligence (XAI)*. DARPA-BAA-16-53. URL:

https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf (cit. on p. 37).

Dejl, Adam, Peter He, Pranav Mangal, Hasan Mohsin, Bogdan Surdu, Eduard Voinea, Emanuele Albini, Piyawat Lertvittayakumjorn, Antonio Rago and Francesca Toni (2021). 'Argflow: A Toolkit for Deep Argumentative Explanations for Neural Networks'. In: *AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021*. Ed. by Frank Dignum, Alessio Lomuscio, Ulle Endriss and Ann Nowé. ACM, pp. 1761–1763. URL: https://dl.acm.org/doi/10.5555/3463952.3464229 (cit. on p. 39).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2019). 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran and Thamar Solorio. Association for Computational Linguistics, pp. 4171–4186. ISBN: 978-1-950737-13-0. URL: https://aclweb.org/anthology/papers/N/N19/N19-1423/ (cit. on p. 36).

Dieterich, William, Christina Mendoza and Tim Brennan (2016). *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity Performance of the COMPAS Risk Scales in Broward County*. Research report. Northpointe Inc. URL: https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf (cit. on p. 99).

Doran, Derek, Sarah Schulz and Tarek R. Besold (2017). 'What Does Explainable AI Really Mean? A New Conceptualization of Perspectives'. In: *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017), Bari, Italy, November 16th and 17th, 2017*. Ed. by Tarek R. Besold and Oliver Kutz. Vol. 2071. CEUR Workshop Proceedings. CEUR-WS.org. URL: http://ceur-ws.org/Vol-2071/CExAIIA%5C_2017%5C_paper%5C_2.pdf (cit. on pp. 37, 41).

Dung, Phan Minh (1995). 'On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games'. In: *Artificial Intelligence* 77.2, pp. 321–357 (cit. on pp. 26, 31, 34, 50–52, 55, 69, 88).

— (2014). 'An Axiomatic Analysis of Structured Argumentation for Prioritized Default Reasoning'. In: *ECAI 2014 - 21st European Conference on Artificial Intelligence, 18-22 August 2014, Prague, Czech Republic - Including Prestigious Applications of Intelligent Systems (PAIS 2014)*. Ed. by Torsten Schaub, Gerhard Friedrich and Barry O'Sullivan. Vol. 263. Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 267–272. ISBN: 978-1-61499-418-3. DOI: 10.3233/978-1-61499-419-0-267. URL: https://doi.org/10.3233/978-1-61499-419-0-267 (cit. on pp. 32, 68).

— (Feb. 2016). 'An axiomatic analysis of structured argumentation with priorities'. In: *Artificial Intelligence* 231, pp. 107–150. ISSN: 0004-3702. DOI: 10.1016/j.artint.2015.10.005. URL: http://dx.doi.org/10.1016/j.artint.2015.10.005 (cit. on pp. 32, 68).

Dung, Phan Minh, Robert A. Kowalski and Francesca Toni (2006). 'Dialectic proof procedures for assumption-based, admissible argumentation'. In: *Artif. Intell.* 170.2, pp. 114–159. DOI: 10.1016/j.artint.2005.07.002. URL: https://doi.org/10.1016/j.artint.2005.07.002 (cit. on p. 63).

Dung, Phan Minh, Paolo Mancarella and Francesca Toni (2007). 'Computing ideal sceptical argumentation'. In: *Artif. Intell.* 171.10-15, pp. 642–674. DOI: 10.1016/j.artint.2007.05.003. URL: https://doi.org/10.1016/j.artint.2007.05.003 (cit. on p. 63).

Dung, Phan Minh and Tran Cao Son (2001). 'An argument-based approach to reasoning with specificity'. In: *Artif. Intell.* 133.1-2, pp. 35–85. DOI: 10.1016/S0004-3702(01)00134-5. URL: https://doi.org/10.1016/S0004-3702(01)00134-5 (cit. on p. 136).

Dung, Phan Minh and Phan Minh Thang (2018). 'Representing the semantics of abstract dialectical frameworks based on arguments and attacks'. In: *Argument Comput.* 9.3, pp. 249–267. DOI: 10.3233/AAC-180427. URL: https://doi.org/10.3233/AAC-180427 (cit. on p. 32).

Durán, Juan M. (2021). 'Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare'. In: *Artif. Intell.* 297, p. 103498.

DOI: 10.1016/j.artint.2021.103498. URL: https://doi.org/10.1016/j.artint.2021.103498 (cit. on p. 38).

Equality Act (2010). URL: https://www.legislation.gov.uk/ukpga/2010/15/contents (visited on 10/11/2023) (cit. on pp. 110, 115).

Espino, Orlando and Ruth M. J. Byrne (2020). 'The Suppression of Inferences From Counterfactual Conditionals'. In: *Cognitive science* 44 4, e12827. URL: https://api.semanticscholar.org/CorpusID:215773481 (cit. on p. 133).

European Commission (2020). *On Artificial Intelligence - A European approach to excellence and trust*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%5C%3A52020DC0065 (cit. on p. 37).

Fabris, Alessandro, Stefano Messina, Gianmaria Silvello and Gian Antonio Susto (2022). 'Tackling Documentation Debt: A Survey on Algorithmic Fairness Datasets'. In: *Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO 2022, Arlington, VA, USA, October 6-9, 2022*. ACM, 2:1–2:13. ISBN: 978-1-4503-9477-2. DOI: 10.1145/3551624.3555286. URL: https://doi.org/10.1145/3551624.3555286 (cit. on p. 110).

Fugard, Andrew J. B., Niki Pfeifer, Bastian Mayerhofer and Gernot D. Kleiter (2011). 'How people interpret conditionals: Shifts toward the conditional event.' In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37.3, pp. 635–648. ISSN: 0278-7393. DOI: 10.1037/a0022329. URL: http://dx.doi.org/10.1037/a0022329 (cit. on pp. 127, 133).

Fuller, Lon L. (1958). 'Positivism and Fidelity to Law: A Reply to Professor Hart'. In: *Harvard Law Review* 71.4, pp. 630–672. ISSN: 0017-811X. DOI: 10.2307/1338226. JSTOR: 1338226. URL: https://www.jstor.org/stable/1338226 (visited on 12/11/2023) (cit. on p. 52).

— (1969). *The Morality of Law*. Revised edition. New Haven and London: Yale University Press (cit. on p. 27).

Fungwacharakorn, Wachara, Kanae Tsushima and Ken Satoh (Dec. 2021). 'Resolving Counterintuitive Consequences in Law Using Legal Debugging'. In: *Artificial Intelligence and Law* 29.4, pp. 541–557. ISSN: 0924-8463, 1572-8382. DOI: 10.1007/s10506-021-09283-7. URL: https://link.springer.com/10.1007/s10506-021-09283-7 (visited on 14/11/2023) (cit. on pp. 67, 95).

— (1st Apr. 2022). 'Diagnosing and Treating Effect of Legal Rule-Based Revision'. In: *New Generation Computing* 40.1, pp. 25–45. ISSN: 1882-7055.

DOI: 10.1007/s00354-022-00157-3. URL: https://doi.org/10.1007/s00354-022-00157-3 (visited on 14/11/2023) (cit. on pp. 67, 95).

Gabbay, Dov M. (1984). 'Theoretical Foundations for Non-Monotonic Reasoning in Expert Systems'. In: *Logics and Models of Concurrent Systems - Conference proceedings, Colle-sur-Loup (near Nice), France, 8-19 October 1984*. Ed. by Krzysztof R. Apt. Vol. 13. NATO ASI Series. Springer, pp. 439–457. ISBN: 978-3-642-82455-5. DOI: 10.1007/978-3-642-82453-1_15. URL: https://doi.org/10.1007/978-3-642-82453-1%5C_15 (cit. on pp. 29, 31, 48).

Gabbay, Dov M., C. J. Hogger and J. A. Robinson, eds. (1994). *Handbook of Logic in Artificial Intelligence and Logic Programming - Volume 3 - Nonmonotonic Reasoning and Uncertain Reasoning*. Oxford University Press (cit. on pp. 26, 31).

— eds. (1998). *Handbook of Logic in Artificial Intelligence and Logic Programming - Volume 5 - Logic Programming*. Oxford University Press (cit. on p. 31).

Garner, Bryan A. (2009). *Black's Law Dictionary: Deluxe Ninth Edition*. 9th ed. West. ISBN: 0314199500,9780314199508 (cit. on p. 26).

Gentzen, Gerhard (1934). 'Investigation into logical deduction'. In: *The collected papers of Gerhard Gentzen*. Studies in logic and the foundations of mathematics. North-Holland Pub. Co (cit. on p. 30).

— (1969). *The collected papers of Gerhard Gentzen*. Studies in logic and the foundations of mathematics. North-Holland Pub. Co (cit. on p. 48).

Ghosh, Partha, Mehdi S. M. Sajjadi, Antonio Vergari, Michael J. Black and Bernhard Schölkopf (2020). 'From Variational to Deterministic Autoencoders'. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL: https://openreview.net/forum?id=S1g7tpEYDS (cit. on pp. 62, 102, 105).

Glorot, Xavier, Antoine Bordes and Yoshua Bengio (2011). 'Deep Sparse Rectifier Neural Networks'. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*. Ed. by Geoffrey J. Gordon, David B. Dunson and Miroslav Dudík. Vol. 15. JMLR Proceedings. JMLR.org, pp. 315–

323. URL: http://proceedings.mlr.press/v15/glorot11a/glorot11a.pdf (cit. on p. 119).

Gniewkowski, Mateusz, Paweł Walkowiak, Piotr Syga, Marek Klonowski and Tomasz Walkowiak (2023). 'Do Not Trust Me: Explainability Against Text Classification'. In: *ECAI 2023*. IOS Press, pp. 875–882. DOI: 10.3233/FAIA230356. URL: https://ebooks.iospress.nl/doi/10.3233/FAIA230356 (visited on 08/11/2023) (cit. on p. 126).

Goodfellow, Ian J., Yoshua Bengio and Aaron C. Courville (2016). *Deep Learning*. Adaptive computation and machine learning. MIT Press. ISBN: 978-0-262-03561-3. URL: http://www.deeplearningbook.org/ (cit. on p. 62).

Google (2023). *Bard: An important next step on our AI journey*. URL: https://blog.google/technology/ai/bard-google-ai-search-updates/ (cit. on p. 36).

Grabmair, Matthias (2016). 'Modeling purposive legal argumentation and case outcome prediction using argument schemes in the value judgment formalism'. PhD thesis. University of Pittsburgh, USA. URL: http://d-scholarship.pitt.edu/27608/7/MGrabmair-ETD-v2.pdf (cit. on pp. 21, 35, 36, 90, 121, 151, 209).

— (2017a). 'Predicting trade secret case outcomes using argument schemes and learned quantitative value effect tradeoffs'. In: *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law* (cit. on pp. 35, 36).

— (June 2017b). 'Predicting trade secret case outcomes using argument schemes and learned quantitative value effect tradeoffs'. In: *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*. DOI: 10.1145/3086512.3086521. URL: http://dx.doi.org/10.1145/3086512.3086521 (cit. on pp. 94, 121, 151).

Grabmair, Matthias and Kevin D. Ashley (2011). 'Facilitating case comparison using value judgments and intermediate legal concepts'. In: *International Conference on Artificial Intelligence and Law* (cit. on p. 36).

Greenleaf, Graham, Andrew Mowbray and Philip Chung (2018). 'Building sustainable free legal advisory systems: Experiences from the history of AI & law'. In: *Comput. Law Secur. Rev.* 34.2, pp. 314–326. DOI: 10.1016/j.clsr.

2018.02.007. URL: https://doi.org/10.1016/j.clsr.2018.02.007 (cit. on p. 26).

Greenspan, P. S. (1975). 'Conditional Oughts and Hypothetical Imperatives'. In: *Journal of Philosophy* 72.10, pp. 259–276. DOI: 10.2307/2024734 (cit. on pp. 41, 93).

Grgic-Hlaca, Nina, M. B. Zafar, K. Gummadi and Adrian Weller (2016). 'The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making'. In: NIPS Symposium on Machine Learning and the Law. URL: https://www.semanticscholar.org/paper/The-Case-for-Process-Fairness-in-Learning%3A-Feature-Grgic-Hlaca-Zafar/fdb6a159cb65f4d1147224998d56e67f0398948b (visited on 14/12/2023) (cit. on pp. 110, 111, 115).

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti and Dino Pedreschi (2019). 'A Survey of Methods for Explaining Black Box Models'. In: *ACM Comput. Surv.* 51.5, 93:1–93:42. DOI: 10.1145/3236009. URL: https://doi.org/10.1145/3236009 (cit. on p. 37).

Gunning, David (2019). 'DARPA's explainable artificial intelligence (XAI) program'. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI 2019, Marina del Ray, CA, USA, March 17-20, 2019*. Ed. by Wai-Tat Fu, Shimei Pan, Oliver Brdiczka, Polo Chau and Gaelle Calvary. ACM. ISBN: 978-1-4503-6272-6. DOI: 10.1145/3301275.3308446. URL: https://doi.org/10.1145/3301275.3308446 (cit. on p. 126).

Gunning, David and David W. Aha (2019). 'DARPA's Explainable Artificial Intelligence (XAI) Program'. In: *AI Mag.* 40.2, pp. 44–58. DOI: 10.1609/AIMAG.V40I2.2850. URL: https://doi.org/10.1609/aimag.v40i2.2850 (cit. on pp. 37, 41).

Gut, Allan (1995). *An Intermediate Course in Probability*. Springer New York. ISBN: 9781475724318. DOI: 10.1007/978-1-4757-2431-8. URL: http://dx.doi.org/10.1007/978-1-4757-2431-8 (cit. on pp. 106, 107).

Guyon, Isabelle, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan and Roman Garnett, eds. (2017). *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. URL: https://proceedings.neurips.cc/paper/2017.

Gyevnar, Balint, Nick Ferguson and Burkhard Schafer (2023). 'Bridging the Transparency Gap: What Can Explainable AI Learn from the AI Act?' In: *ECAI 2023*. IOS Press, pp. 964–971.
DOI: 10.3233/FAIA230367. URL: https://ebooks.iospress.nl/doi/10.3233/FAIA230367 (visited on 08/11/2023) (cit. on p. 126).

Hart, H. L. A. (1958). 'Positivism and the Separation of Law and Morals'. In: *Harvard Law Review* 71.4, pp. 593–629. ISSN: 0017-811X. DOI: 10.2307/1338225. JSTOR: 1338225. URL: https://www.jstor.org/stable/1338225 (visited on 12/11/2023) (cit. on p. 52).

— (1994). *The Concept of Law*. Second ed. Oxford: Clarendon (cit. on p. 27).

Heyninck, Jesse, Gabriele Kern-Isberner, Tjitze Rienstra, Kenneth Skiba and Matthias Thimm (Apr. 2023). 'Revision, defeasible conditionals and non-monotonic inference for abstract dialectical frameworks'. In: *Artificial Intelligence* 317, p. 103876. ISSN: 0004-3702. DOI: 10.1016/j.artint.2023.103876. URL: http://dx.doi.org/10.1016/j.artint.2023.103876 (cit. on p. 26).

Hildebrandt, Mireille, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor and Gabriela Zanfir-Fortuna, eds. (2020). *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*. ACM. ISBN: 978-1-4503-6936-7.
DOI: 10.1145/3351095. URL: https://doi.org/10.1145/3351095 (cit. on p. 138).

Hitzler, Pascal, Federico Bianchi, Monireh Ebrahimi and Md Kamruzzaman Sarker (31st Jan. 2020). 'Neural-Symbolic Integration and the Semantic Web'. In: *Semantic Web* 11.1. Ed. by Krzysztof Janowicz, pp. 3–11. ISSN: 22104968, 15700844. DOI: 10.3233/SW-190368. URL: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-190368 (visited on 14/11/2023) (cit. on p. 94).

Horty, John F. (1994). 'Some Direct Theories of Nonmonotonic Inheritance'. In: *Handbook of Logic in Artificial Intelligence and Logic Programming - Volume 3 - Nonmonotonic Reasoning and Uncertain Reasoning*. Ed. by Dov M. Gabbay, C. J. Hogger and J. A. Robinson. Oxford University Press, pp. 111–188 (cit. on p. 136).

Horty, John F. (Mar. 2004). 'The Result Model Of Precedent'. In: *Legal Theory* 10.1, pp. 19–31. ISSN: 1469-8048. DOI: 10.1017/s1352325204000151. URL: http://dx.doi.org/10.1017/s1352325204000151 (cit. on pp. 36, 56–58).

— (15th Sept. 2015). 'Constraint and Freedom in the Common Law'. In: *Philosopher's Imprint* 15.25. URL: https://api.semanticscholar.org/CorpusID:15038751 (visited on 13/11/2023) (cit. on p. 57).

— (Mar. 2019). 'Reasoning with dimensions and magnitudes'. In: *Artificial Intelligence and Law* 27.3, pp. 309–345. ISSN: 1572-8382. DOI: 10.1007/s10506-019-09245-0. URL: http://dx.doi.org/10.1007/s10506-019-09245-0 (cit. on pp. 36, 91).

— (2021). 'Modifying the reason model'. In: *Artif. Intell. Law* 29.2, pp. 271–285.
DOI: 10.1007/s10506-020-09275-z. URL: https://doi.org/10.1007/s10506-020-09275-z (cit. on pp. 56, 57, 59, 150).

Horty, John F. and Trevor J. M. Bench-Capon (May 2012). 'A factor-based definition of precedential constraint'. In: *Artificial Intelligence and Law* 20.2, pp. 181–214. ISSN: 1572-8382. DOI: 10.1007/s10506-012-9125-8. URL: http://dx.doi.org/10.1007/s10506-012-9125-8 (cit. on pp. 36, 56, 57, 91, 96).

Hunter, Anthony (2010). 'Base Logics in Argumentation'. In: *Computational Models of Argument: Proceedings of COMMA 2010, Desenzano del Garda, Italy, September 8-10, 2010*. Ed. by Pietro Baroni, Federico Cerutti, Massimiliano Giacomin and Guillermo Ricardo Simari. Vol. 216. Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 275–286. ISBN: 978-1-60750-618-8.
DOI: 10.3233/978-1-60750-619-5-275. URL: https://doi.org/10.3233/978-1-60750-619-5-275 (cit. on p. 68).

Ignatiev, Alexey, Nina Narodytska, Nicholas Asher and Joao Marques-Silva (2021). 'From Contrastive to Abductive Explanations and Back Again'. In: *AIxIA 2020 – Advances in Artificial Intelligence*. Ed. by Matteo Baldoni and Stefania Bandini. Cham: Springer International Publishing, pp. 335–355. ISBN: 978-3-030-77091-4. DOI: 10.1007/978-3-030-77091-4_21 (cit. on p. 152).

Ignatiev, Alexey, Nina Narodytska and João Marques-Silva (2019). 'On Relating Explanations and Adversarial Examples'. In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox and Roman Garnett, pp. 15857–15867. URL: https://proceedings.neurips.cc/paper/2019/hash/7392ea4ca76ad2fb4c9c3b6a5c6e31e3-Abstract.html (cit. on pp. 62, 126).

Information Commissioner's Office and The Alan Turing Institute (2022). *Explaining decisions made with AI*. URL: https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-artificial-intelligence-1-0.pdf (cit. on p. 37).

Israel, David J. (1993). 'The role(s) of logic in artificial intelligence'. In: *Handbook of Logic in Artificial Intelligence and Logic Programming - Volume 1 - Logical Foundations*. Ed. by Dov M. Gabbay, C. J. Hogger and J. A. Robinson. Oxford University Press, pp. 1–30 (cit. on p. 30).

Kakas, Antonis C., Robert A. Kowalski and Francesca Toni (1992). 'Abductive Logic Programming'. In: *Journal of Logic and Computation* 2.6, pp. 719–770. ISSN: 1465-363X. DOI: 10.1093/logcom/2.6.719. URL: http://dx.doi.org/10.1093/logcom/2.6.719 (cit. on p. 31).

Kaminski, Margot E. (2019). 'The Right to Explanation, Explained'. In: *Berkeley Technology Law Journal*. DOI: 10.15779/Z38TD9N83H (cit. on p. 38).

Keilson, J. and H. Gerber (June 1971). 'Some Results for Discrete Unimodality'. In: *Journal of the American Statistical Association* 66.334, pp. 386–389. ISSN: 1537-274X. DOI: 10.1080/01621459.1971.10482273. URL: http://dx.doi.org/10.1080/01621459.1971.10482273 (cit. on p. 104).

Kenny, Eoin M. and Mark T. Keane (2019). 'Twin-Systems to Explain Artificial Neural Networks using Case-Based Reasoning: Comparative Tests of Feature-Weighting Methods in ANN-CBR Twins for XAI'. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. Ed. by Sarit Kraus. ijcai.org, pp. 2708–2715. DOI: 10.24963/ijcai.2019/376. URL: https://doi.org/10.24963/ijcai.2019/376 (cit. on pp. 126, 149).

Kenny, Eoin M. and Mark T. Keane (2021). 'Explaining Deep Learning using examples: Optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in XAI'. In: *Knowl. Based Syst.* 233, p. 107530. DOI: 10.1016/j.knosys.2021.107530. URL: https://doi.org/10.1016/j.knosys.2021.107530 (cit. on p. 126).

Kim, Been, Cynthia Rudin and Julie A. Shah (2014). 'The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification'. In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. Ed. by Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence and Kilian Q. Weinberger, pp. 1952–1960. URL: https://proceedings.neurips.cc/paper/2014/hash/390e982518a50e280d8e2b535462ec1f-Abstract.html (cit. on pp. 35, 121).

Kingma, Diederik P. and Jimmy Ba (2015). 'Adam: A Method for Stochastic Optimization'. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. URL: http://arxiv.org/abs/1412.6980 (cit. on p. 119).

Kingma, Diederik P. and Max Welling (2014). 'Auto-Encoding Variational Bayes'. In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. URL: http://arxiv.org/abs/1312.6114 (cit. on p. 61).

Klauer, Karl Christoph, Christoph Stahl and Edgar Erdfelder (2007). 'The abstract selection task: New data and an almost comprehensive model.' In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33.4, pp. 680–703. ISSN: 0278-7393. DOI: 10.1037/0278-7393.33.4.680. URL: http://dx.doi.org/10.1037/0278-7393.33.4.680 (cit. on p. 127).

Kolodner, Janet L. (1992). 'An introduction to case-based reasoning'. In: *Artif. Intell. Rev.* 6.1, pp. 3–34.
DOI: 10.1007/BF00155578. URL: https://doi.org/10.1007/BF00155578 (cit. on pp. 33, 34, 39).

Kraus, Sarit, Daniel Lehmann and Menachem Magidor (1990). 'Nonmonotonic Reasoning, Preferential Models and Cumulative Logics'. In: *Artif. Intell.*

44.1-2, pp. 167–207. DOI: 10.1016/0004-3702(90)90101-5. URL: https://doi.org/10.1016/0004-3702(90)90101-5 (cit. on pp. 29, 132, 134).

Krishnapuram, Balaji, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen and Rajeev Rastogi, eds. (2016). *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. ACM. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672. URL: https://doi.org/10.1145/2939672.

Kumar, Sachin, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos and Yulia Tsvetkov (2022). 'Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey'. In: arXiv. DOI: 10.48550/ARXIV.2210.07700. URL: https://arxiv.org/abs/2210.07700 (visited on 09/11/2023) (cit. on p. 36).

Kusner, Matt J., Joshua R. Loftus, Chris Russell and Ricardo Silva (2017). 'Counterfactual Fairness'. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan and Roman Garnett, pp. 4066–4076. URL: https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html (cit. on pp. 110, 111, 115).

Lakkaraju, Himabindu, Stephen H. Bach and Jure Leskovec (2016). 'Interpretable Decision Sets: A Joint Framework for Description and Prediction'. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. Ed. by Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen and Rajeev Rastogi. ACM, pp. 1675–1684. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939874. URL: https://doi.org/10.1145/2939672.2939874 (cit. on pp. 128, 129).

Langer, Markus, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing and Kevin Baum (1st July 2021). 'What Do We Want from Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research'. In: *Artificial Intelligence* 296, p. 103473. ISSN: 0004-3702. DOI: 10.1016/j.artint.2021.103473. URL: https://www.sciencedirect.com/

science/article/pii/S0004370221000242 (visited on 10/11/2023) (cit. on p. 126).

Larson, Jeff, Surya Mattu, Lauren Kirchner and Julia Angwin (2016). *How We Analyzed the COMPAS Recidivism Algorithm.* URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (cit. on pp. 99, 109, 110, 147).

Law, Mark, Alessandra Russo and Krysia Broda (2019). 'Logic-Based Learning of Answer Set Programs'. In: *Reasoning Web. Explainable Artificial Intelligence - 15th International Summer School 2019, Bolzano, Italy, September 20-24, 2019, Tutorial Lectures.* Ed. by Markus Krötzsch and Daria Stepanova. Vol. 11810. Lecture Notes in Computer Science. Springer, pp. 196–231. ISBN: 978-3-030-31422-4. DOI: 10.1007/978-3-030-31423-1_6. URL: https://doi.org/10.1007/978-3-030-31423-1%5C_6 (cit. on p. 39).

Leen, Todd K., Thomas G. Dietterich and Volker Tresp, eds. (2001). *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*. MIT Press. URL: https://proceedings.neurips.cc/paper/2000 (cit. on p. 97).

Lewis, Sebastian (Mar. 2021). 'Precedent and the Rule of Law'. In: *Oxford Journal of Legal Studies* 41.4, pp. 873–898. ISSN: 1464-3820. DOI: 10.1093/ojls/gqab007. URL: http://dx.doi.org/10.1093/ojls/gqab007 (cit. on pp. 27, 57).

Li, Oscar, Hao Liu, Chaofan Chen and Cynthia Rudin (2018). 'Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions'. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018.* Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, pp. 3530–3537. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17082 (cit. on pp. 35, 121, 126, 149).

Lin, Fangzhen and Yoav Shoham (1989). 'Argument Systems: A Uniform Basis for Nonmonotonic Reasoning'. In: *Proceedings of the 1st International Conference on Principles of Knowledge Representation and Reasoning (KR'89).*

Toronto, Canada, May 15-18 1989. Ed. by Ronald J. Brachman, Hector J. Levesque and Raymond Reiter. Morgan Kaufmann, pp. 245–255. ISBN: 1-55860-032-9 (cit. on p. 31).

London Ambulance Service NHS Trust (1st Sept. 2017). *Cycle Responder - London Ambulance Service NHS Trust*. London Ambulance Service NHS Trust - accidents, traffic accidents, car, vehicle. URL: https://www.londonambulance.nhs.uk/calling-us/who-will-treat-you/single-responder/cycle-responder/ (visited on 12/11/2023) (cit. on p. 53).

López, Beatriz (2013). *Case-Based Reasoning: A Concise Introduction*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers. ISBN: 9781627050074.
DOI: 10.2200/S00490ED1V01Y201303AIM020. URL: https://doi.org/10.2200/S00490ED1V01Y201303AIM020 (cit. on p. 33).

Lundberg, Scott M. and Su-In Lee (2017). 'A Unified Approach to Interpreting Model Predictions'. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan and Roman Garnett, pp. 4765–4774. URL: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (cit. on pp. 63, 152).

MacCormick, D. Neil and Robert S. Summers, eds. (2016). *Interpreting Precedents: A Comparative Study*. Originally published in 1997. London: Routledge. 598 pp. ISBN: 978-1-315-25190-5. DOI: 10.4324/9781315251905 (cit. on p. 57).

Makinson, David (1994). 'General Patterns in Nonmonotonic Reasoning'. In: *Handbook of Logic in Artificial Intelligence and Logic Programming - Volume 3 - Nonmonotonic Reasoning and Uncertain Reasoning*. Ed. by Dov M. Gabbay, C. J. Hogger and J. A. Robinson. Oxford University Press, pp. 35–110 (cit. on pp. 29, 31, 48, 132, 134, 136).

Maranhão, Juliano (2006). 'Why was Alchourrón afraid of snakes'. In: *Análisis Filosófico* 26, pp. 62–92. URL: http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S1851-96362006000100004 (cit. on pp. 32, 33).

Marcos Cramer Steffen Hölldobler, Marco Ragni (2021). 'Modeling Human Reasoning About Conditionals'. 19th International Workshop on Non-Monotonic Reasoning. Informal proceedings. Hanoi, Vietnam. URL: https://drive.google.com/open?id=1WSIl3TOrXBhaWhckWN4NLXoD9AVFKp5R (visited on 13/11/2023) (cit. on pp. 30, 127, 133).

Maslej, Nestor, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark and Raymond Perrault (2023). 'Artificial Intelligence Index Report 2023'. In: Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.2310.03715. URL: https://arxiv.org/abs/2310.03715 (visited on 08/11/2023) (cit. on p. 36).

Merrer, Erwan Le and Gilles Trédan (2020). 'Remote explainability faces the bouncer problem'. In: *Nat. Mach. Intell.* 2, pp. 529–539 (cit. on pp. 125, 126).

Michaelson, Greg J. (1989). *An introduction to functional programming through lambda calculus*. International computer science series. Addison-Wesley. ISBN: 978-0-201-17812-8 (cit. on p. 71).

Miller, Tim (2019). 'Explanation in artificial intelligence: Insights from the social sciences'. In: *Artif. Intell.* 267, pp. 1–38. DOI: 10.1016/j.artint.2018.07.007. URL: https://doi.org/10.1016/j.artint.2018.07.007 (cit. on pp. 37, 41, 62, 126).

Mitchell, Tom M. (1997). *Machine learning*. McGraw Hill series in computer science. McGraw-Hill. ISBN: 978-0-07-042807-2. URL: http://www.worldcat.org/oclc/61321007 (cit. on pp. 33, 34, 59, 60, 71).

Mittelstadt, Brent D., Chris Russell and Sandra Wachter (2019). 'Explaining Explanations in AI'. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*. Ed. by danah boyd and Jamie H. Morgenstern. ACM, pp. 279–288. DOI: 10.1145/3287560.3287574. URL: https://doi.org/10.1145/3287560.3287574 (cit. on pp. 37, 41).

Modgil, Sanjay and Martin Caminada (2009). 'Proof Theories and Algorithms for Abstract Argumentation Frameworks'. In: *Argumentation in Artificial Intelligence*. Ed. by Guillermo Ricardo Simari and Iyad Rahwan. Springer, pp. 105–129. ISBN: 978-0-387-98196-3. DOI: 10.1007/978-0-387-98197-

0_6. URL: https://doi.org/10.1007/978-0-387-98197-0%5C_6 (cit. on pp. 51, 70, 207).

Modgil, Sanjay and Henry Prakken (Feb. 2013). 'A general account of argumentation with preferences'. In: *Artificial Intelligence* 195, pp. 361–397. ISSN: 00043702.
DOI: 10.1016/j.artint.2012.10.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S0004370212001361 (visited on 08/11/2023) (cit. on p. 31).

Molnar, Christoph (2022). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2nd ed. URL: https://christophm.github.io/interpretable-ml-book (cit. on pp. 34, 37, 62, 63, 101).

Monath, Nicholas, Manzil Zaheer, Kumar Avinava Dubey, Amr Ahmed and Andrew McCallum (2021). 'DAG-Structured Clustering by Nearest Neighbors'. In: *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*. Ed. by Arindam Banerjee and Kenji Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, pp. 2854–2862. URL: http://proceedings.mlr.press/v130/monath21a.html (cit. on p. 121).

Mosqueira-Rey, Eduardo, Elena Hernández-Pereira, David Alonso-Ríos, Jose Ramon Bobes-Bascaran and Ángel Fernández-Leal (2022). 'Human-in-the-loop machine learning: a state of the art'. In: *Artificial Intelligence Review* (cit. on p. 41).

Mueller, Shane T., Robert R. Hoffman, William J. Clancey, Abigail Emrey and Gary Klein (2019). 'Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI'. In: *CoRR* abs/1902.01876. arXiv: 1902.01876. URL: http://arxiv.org/abs/1902.01876 (cit. on pp. 27, 37, 41, 62, 126).

Muggleton, Stephen H. and Luc De Raedt (1994). 'Inductive Logic Programming: Theory and Methods'. In: *J. Log. Program.* 19/20, pp. 629–679.
DOI: 10.1016/0743-1066(94)90035-3. URL: https://doi.org/10.1016/0743-1066(94)90035-3 (cit. on p. 39).

Mumford, Jack, Katie Atkinson and Trevor J. M. Bench-Capon (2022). 'Reasoning with Legal Cases: A Hybrid ADF-ML Approach'. In: *Legal Knowledge and Information Systems - JURIX 2022: The Thirty-fifth Annual Conference, Saarbrücken, Germany, 14-16 December 2022*. Ed. by Enrico

Francesconi, Georg Borges and Christoph Sorge. Vol. 362. Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 93–102.
DOI: 10.3233/FAIA220452. URL: https://doi.org/10.3233/FAIA220452 (cit. on pp. 120, 151).

Nair, Vinod and Geoffrey E. Hinton (2010). 'Rectified Linear Units Improve Restricted Boltzmann Machines'. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*. Ed. by Johannes Fürnkranz and Thorsten Joachims. Omnipress, pp. 807–814. URL: https://icml.cc/Conferences/2010/papers/432.pdf (cit. on p. 119).

Naudts, Laurens, Pierre Dewitte and Jef Ausloos (22nd Apr. 2022). 'Meaningful Transparency through Data Rights: A Multidimensional Analysis'. In: *Research Handbook on EU Data Protection Law*. Ed. by Eleni Kosta, Ronald Leenes and Irene Kamara. Edward Elgar Publishing. ISBN: 978-1-80037-168-2 978-1-80037-167-5.
DOI: 10.4337/9781800371682.00030. URL: https://www.elgaronline.com/view/edcoll/9781800371675/9781800371675.00030.xml (visited on 09/11/2023) (cit. on p. 38).

Neves, Rui Da Silva, Jean-François Bonnefon and Eric Raufaste (2004). 'An Empirical Test of Patterns for Nonmonotonic Inference'. In: *Annals of Mathematics and Artificial Intelligence* 34, pp. 107–130 (cit. on pp. 30, 127, 133).

Nguyen, Ha-Thanh, Randy Goebel, Francesca Toni, Kostas Stathis and Ken Satoh (2023). 'How well do SOTA legal reasoning models support abductive reasoning?' In: *Proceedings of the International Conference on Logic Programming 2023 Workshops co-located with the 39th International Conference on Logic Programming (ICLP 2023), London, United Kingdom, July 9th and 10th, 2023*. Ed. by Joaquín Arias, Sotiris Batsakis, Wolfgang Faber, Gopal Gupta, Francesco Pacenza, Emmanuel Papadakis, Livio Robaldo, Kilian Rückschloß, Elmer Salazar, Zeynep Gozen Saribatur, Ilias Tachmazidis, Felix Weitkämper and Adam Z. Wyner. Vol. 3437. CEUR Workshop Proceedings. CEUR-WS.org. URL: https://ceur-ws.org/Vol-3437/paper1LPLR.pdf (cit. on p. 36).

Nofal, Samer, Katie Atkinson and Paul E. Dunne (2021). 'Computing Grounded Extensions Of Abstract Argumentation Frameworks'. In: *Comput. J.* 64.1, pp. 54–63. DOI: 10.1093/comjnl/bxz138. URL: https://doi.org/10.1093/comjnl/bxz138 (cit. on p. 51).

Northpointe Inc. (2019). *Practitioner's Guide to COMPAS Core*. URL: https://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf (cit. on pp. 109, 147).

Nugent, Conor and Padraig Cunningham (2005). 'A Case-Based Explanation System for Black-Box Systems'. In: *Artif. Intell. Rev.* 24.2, pp. 163–178. DOI: 10.1007/s10462-005-4609-5. URL: https://doi.org/10.1007/s10462-005-4609-5 (cit. on pp. 39, 126).

Nute, Donald and Xiaochang Yu (1997). 'Introduction'. In: *Defeasible Deontic Logic*. Ed. by Donald Nute. Springer Netherlands, pp. 1–16. ISBN: 9789401588515.
DOI: 10.1007/978-94-015-8851-5. URL: http://dx.doi.org/10.1007/978-94-015-8851-5 (cit. on p. 41).

O'Donnell, Michael J. (1998). 'Introduction: Logic and Logic Programming Languages'. In: *Handbook of Logic in Artificial Intelligence and Logic Programming - Volume 5 - Logic Programming*. Ed. by Dov M. Gabbay, C. J. Hogger and J. A. Robinson. Oxford University Press.
DOI: 10.1093/oso/9780198537922.003.0004. URL: http://dx.doi.org/10.1093/oso/9780198537922.003.0004 (cit. on p. 31).

OpenAI (2022). *ChatGPT: Optimizing Language Models for Dialogue*. URL: https://openai.com/blog/chatgpt (cit. on p. 36).

— (15th Mar. 2023). 'GPT-4 Technical Report'. In: *ArXiv*. URL: https://www.semanticscholar.org/paper/GPT-4-Technical-Report-OpenAI/8ca62fdf4c276ea3052dc96dcfd8ee96ca425a48 (visited on 09/11/2023) (cit. on p. 36).

Organisation for Economic Co-operation and Development (2019). *Recommendation of the Council on Artificial Intelligence*. URL: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (cit. on p. 37).

Parent, Xavier (2001). 'Cumulativity, Identity and Time in Deontic Logic'. In: *Fundam. Informaticae* 48.2-3, pp. 237–252. URL: http://content.iospress.com/articles/fundamenta-informaticae/fi48-2-3-08 (cit. on p. 41).

Parent, Xavier (2014). 'Maximality vs. Optimality in Dyadic Deontic Logic'. In: *J. Philos. Log.* 43.6, pp. 1101–1128. DOI: 10.1007/s10992-013-9308-0. URL: https://doi.org/10.1007/s10992-013-9308-0 (cit. on p. 41).

Paulino-Passos, Guilherme and Francesca Toni (2020). 'Cautious Monotonicity in Case-Based Reasoning with Abstract Argumentation'. In: *CoRR* abs/2007.05284. 18th International Workshop On Non-Monotonic Reasoning. Informal proceedings. arXiv: 2007.05284. URL: https://arxiv.org/abs/2007.05284 (cit. on p. 44).

— (2021). 'Monotonicity and Noise-Tolerance in Case-Based Reasoning with Abstract Argumentation'. In: *Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning, KR 2021, Online event, November 3-12, 2021*. Ed. by Meghyn Bienvenu, Gerhard Lakemeyer and Esra Erdem, pp. 508–518. ISBN: 978-1-956792-99-7. DOI: 10.24963/kr.2021/48. URL: https://doi.org/10.24963/kr.2021/48 (cit. on pp. 44, 142).

— (2022a). 'On Interactive Explanations as Non-Monotonic Reasoning'. In: *CoRR* abs/2208.00316. Workshop on Explainable Artificial Intelligence (XAI) at IJCAI 2022. DOI: 10.48550/ARXIV.2208.00316. arXiv: 2208.00316. URL: https://doi.org/10.48550/arXiv.2208.00316 (cit. on p. 44).

— (2022b). 'On Interactive Explanations as Reasoning'. XLoKR 2022: The Third Workshop on Explainable Logic-Based Knowledge Representation at KR2022. Informal proceedings. URL: https://drive.google.com/file/d/1eWsB1up6Il_-L81GaTClW06W3oKV9tEw/view?usp=sharing (cit. on p. 45).

— (2022c). 'On Monotonicity of Dispute Trees as Explanations for Case-Based Reasoning with Abstract Argumentation'. In: *1st International Workshop on Argumentation for eXplainable AI co-located with 9th International Conference on Computational Models of Argument (COMMA 2022), Cardiff, Wales, September 12, 2022*. Ed. by Kristijonas Čyras, Timotheus Kampik, Oana Cocarascu and Antonio Rago. Vol. 3209. CEUR Workshop Proceedings. CEUR-WS.org. URL: http://ceur-ws.org/Vol-3209/8465.pdf (cit. on p. 45).

— (2023). 'Learning Case Relevance in Case-Based Reasoning with Abstract Argumentation'. In: *Legal Knowledge and Information Systems - JURIX*

2023: The Thirty-sixth Annual Conference, Maastricht, the Netherlands, 18-20 December 2023. (Accepted, forthcoming.) (cit. on p. 44).

Perlman, Andrew M. (5th Dec. 2022). *The Implications of ChatGPT for Legal Services and Society.* DOI: 10.2139/ssrn.4294197. URL: https://papers.ssrn.com/abstract=4294197 (visited on 09/11/2023). preprint (cit. on p. 36).

Peters, Joeri GT, Floris J. Bex, Henry Prakken, Kristijonas Čyras, Timotheus Kampik, Oana Cocarascu and Antonio Rago (2022). 'Justifications Derived from Inconsistent Case Bases Using Authoritativeness'. In: *Proceedings of the 1st International Workshop on Argumentation for eXplainable AI (ArgXAI 2022) Co-Located with 9th International Conference on Computational Models of Argument (COMMA 2022) Cardiff, Wales, September 12, 2022.* Vol. 3209. CEUR WS, pp. 1–13. URL: https://dspace.library.uu.nl/handle/1874/424498 (visited on 12/11/2023) (cit. on p. 59).

Poole, David (1985). 'On the Comparison of Theories: Preferring the Most Specific Explanation'. In: *Proceedings of the 9th International Joint Conference on Artificial Intelligence. Los Angeles, CA, USA, August 1985.* Ed. by Aravind K. Joshi. Morgan Kaufmann, pp. 144–147. URL: http://ijcai.org/Proceedings/85-1/Papers/026.pdf (cit. on p. 136).

Popper Karl, Karl Popper (21st Feb. 2002). *The Logic of Scientific Discovery.* 2nd ed. London: Routledge. 544 pp. ISBN: 978-0-203-99462-7. DOI: 10.4324/9780203994627 (cit. on p. 39).

Potyka, Nico, Xiang Yin and Francesca Toni (2022). 'On the Tradeoff Between Correctness and Completeness in Argumentative Explainable AI'. In: *1st International Workshop on Argumentation for eXplainable AI co-located with 9th International Conference on Computational Models of Argument (COMMA 2022), Cardiff, Wales, September 12, 2022.* Ed. by Kristijonas Čyras, Timotheus Kampik, Oana Cocarascu and Antonio Rago. Vol. 3209. CEUR Workshop Proceedings. CEUR-WS.org. URL: http://ceur-ws.org/Vol-3209/8151.pdf (cit. on pp. 39, 41).

Prakken, Henry (1997). 'Logical Tools for Modelling Legal Argument'. In: *Law and Philosophy Library.* ISSN: 1572-4395. DOI: 10.1007/978-94-015-8975-8. URL: http://dx.doi.org/10.1007/978-94-015-8975-8 (cit. on pp. 33, 40, 95).

Prakken, Henry (June 2010). 'An abstract framework for argumentation with structured arguments'. In: *Argument & Computation* 1.2, pp. 93–124. ISSN: 1946-2166, 1946-2174.
DOI: 10.1080/19462160903564592. URL: http://content.iospress.com/doi/10.1080/19462160903564592 (visited on 07/11/2023) (cit. on p. 31).

— (2018). 'Historical Overview of Formal Argumentation'. In: *Handbook of Formal Argumentation*. Ed. by Pietro Baroni, Dov Gabbay, Massimiliano Giacomin and Leendert van der Torre. College Publications. ISBN: 1848902751 (cit. on p. 31).

— (Feb. 2021). 'A formal analysis of some factor- and precedent-based accounts of precedential constraint'. In: *Artificial Intelligence and Law*. ISSN: 1572-8382. DOI: 10.1007/s10506-021-09284-6. URL: http://dx.doi.org/10.1007/s10506-021-09284-6 (cit. on pp. 36, 56, 57, 59, 91, 150).

Prakken, Henry and Sanjay Modgil (2018). 'Abstract Rule-Based Argumentation'. In: *Handbook of Formal Argumentation*. Ed. by Pietro Baroni, Dov Gabbay, Massimiliano Giacomin and Leendert van der Torre. College Publications. ISBN: 1848902751 (cit. on pp. 40, 95).

Prakken, Henry and Rosa Ratsma (2022a). 'A top-level model of case-based argumentation for explanation: Formalisation and experiments'. In: *Argument Comput.* 13, pp. 159–194 (cit. on p. 39).

— (2022b). 'A top-level model of case-based argumentation for explanation: Formalisation and experiments'. In: *Argument Comput.* 13.2, pp. 159–194. DOI: 10.3233/AAC-210009. URL: https://doi.org/10.3233/AAC-210009 (cit. on pp. 67, 91, 94).

Prakken, Henry and Giovanni Sartor (1996). 'A Dialectical Model of Assessing Conflicting Arguments in Legal Reasoning'. In: *Artif. Intell. Law* 4.3-4, pp. 331–368.
DOI: 10.1007/BF00118496. URL: https://doi.org/10.1007/BF00118496 (cit. on p. 33).

— (1998). 'Modelling Reasoning with Precedents in a Formal Dialogue Game'. In: *Artif. Intell. Law* 6.2-4, pp. 231–287 (cit. on pp. 35, 40, 57, 150).

— (Oct. 2015). 'Law and Logic'. In: *Artif. Intell.* 227.C, pp. 214–245. ISSN: 0004-3702. DOI: 10.1016/j.artint.2015.06.005. URL: http://dx.doi.org/10.1016/j.artint.2015.06.005 (cit. on p. 32).

Prakken, Henry, Adam Z. Wyner, Trevor J. M. Bench-Capon and Katie Atkinson (2015a). 'A formalization of argumentation schemes for legal case-based reasoning in ASPIC+'. In: *J. Log. Comput.* 25.5, pp. 1141–1166. DOI: 10.1093/logcom/ext010. URL: https://doi.org/10.1093/logcom/ext010 (cit. on pp. 34, 95).

— (2015b). 'A formalization of argumentation schemes for legal case-based reasoning in ASPIC+'. In: *J. Log. Comput.* 25.5, pp. 1141–1166. DOI: 10.1093/logcom/ext010. URL: https://doi.org/10.1093/logcom/ext010 (cit. on pp. 35, 40, 150).

Rabold, Johannes, Michael Siebers and Ute Schmid (2018). 'Explaining Black-Box Classifiers with ILP - Empowering LIME with Aleph to Approximate Non-linear Decisions with Relational Rules'. In: *Inductive Logic Programming - 28th International Conference, ILP 2018, Ferrara, Italy, September 2-4, 2018, Proceedings*. Ed. by Fabrizio Riguzzi, Elena Bellodi and Riccardo Zese. Vol. 11105. Lecture Notes in Computer Science. Springer, pp. 105–117. ISBN: 978-3-319-99959-3. DOI: 10.1007/978-3-319-99960-9_7. URL: https://doi.org/10.1007/978-3-319-99960-9%5C_7 (cit. on p. 39).

Raczyński, Jakub, Mateusz Lango and Jerzy Stefanowski (2023). 'The Problem of Coherence in Natural Language Explanations of Recommendations'. In: *ECAI 2023*. IOS Press, pp. 1922–1929.
DOI: 10.3233/FAIA230482. URL: https://ebooks.iospress.nl/doi/10.3233/FAIA230482 (visited on 08/11/2023) (cit. on p. 126).

Ragni, Marco, Christian Eichhorn, Tanja Bock, Gabriele Kern-Isberner and Alice Ping Ping Tse (2017). 'Formal Nonmonotonic Theories and Properties of Human Defeasible Reasoning'. In: *Minds and Machines* 27, pp. 79–117. URL: https://api.semanticscholar.org/CorpusID:22117331 (cit. on pp. 127, 133).

Ragni, Marco, Christian Eichhorn and Gabriele Kern-Isberner (2016). 'Simulating Human Inferences in the Light of New Information: A Formal Analysis'. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*. Ed. by Subbarao Kambhampati. IJCAI/AAAI Press, pp. 2604–2610. ISBN: 978-1-57735-770-4. URL: http://www.ijcai.org/Abstract/16/370 (cit. on pp. 30, 127, 133).

Ragni, Marco, Gabriele Kern-Isberner, Christoph Beierle and Kai Sauerwald (2020). 'Cognitive Logics - Features, Formalisms, and Challenges'. In: *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*. Ed. by Giuseppe De Giacomo, Alejandro Catalá, Bistra Dilkina, Michela Milano, Senén Barro, Alberto Bugarín and Jérôme Lang. Vol. 325. Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 2931–2932.
DOI: 10.3233/FAIA200459. URL: https://doi.org/10.3233/FAIA200459 (cit. on p. 133).

Rago, Antonio, Pietro Baroni and Francesca Toni (2022). 'Explaining Causal Models with Argumentation: the Case of Bi-variate Reinforcement'. In: *Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning, KR 2022, Haifa, Israel. July 31 - August 5, 2022*. Ed. by Gabriele Kern-Isberner, Gerhard Lakemeyer and Thomas Meyer. ISBN: 978-1-956792-01-0. DOI: 10.24963/kr.2022. URL: https://proceedings.kr.org/2022/52/ (cit. on p. 39).

Rago, Antonio, Oana Cocarascu, Christos Bechlivanidis, David A. Lagnado and Francesca Toni (2021a). 'Argumentative explanations for interactive recommendations'. In: *Artif. Intell.* 296, p. 103506. DOI: 10.1016/j.artint.2021.103506. URL: https://doi.org/10.1016/j.artint.2021.103506 (cit. on pp. 38, 144).

— (2021b). 'Argumentative explanations for interactive recommendations'. In: *Artif. Intell.* 296, p. 103506 (cit. on p. 41).

Rago, Antonio, Oana Cocarascu, Christos Bechlivanidis and Francesca Toni (Sept. 2020a). 'Argumentation as a Framework for Interactive Explanations for Recommendations'. In: *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning*, pp. 805–815.
DOI: 10.24963/kr.2020/83. URL: https://doi.org/10.24963/kr.2020/83 (cit. on p. 38).

— (2020b). 'Argumentation as a Framework for Interactive Explanations for Recommendations'. In: *International Conference on Principles of Knowledge Representation and Reasoning* (cit. on p. 41).

Raz, Joseph (18th June 2009). *The Authority of Law: Essays on Law and Morality*. Second Edition, Second Edition. Oxford, New York: Oxford University Press. 368 pp. ISBN: 978-0-19-957356-1 (cit. on p. 57).

Rezende, Danilo Jimenez, Shakir Mohamed and Daan Wierstra (2014). 'Stochastic Backpropagation and Approximate Inference in Deep Generative Models'. In: *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*. Vol. 32. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 1278–1286. URL: http://proceedings.mlr.press/v32/rezende14.html (cit. on p. 61).

Ribeiro, Marco Túlio, Sameer Singh and Carlos Guestrin (2016a). '"Why Should I Trust You?": Explaining the Predictions of Any Classifier'. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. Ed. by Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen and Rajeev Rastogi. ACM, pp. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778. URL: https://doi.org/10.1145/2939672.2939778 (cit. on pp. 63, 126, 152).

— (2016b). 'Model-Agnostic Interpretability of Machine Learning'. In: *CoRR* abs/1606.05386. Human Interpretability in Machine Learning workshop, ICML 2016. arXiv: 1606.05386. URL: http://arxiv.org/abs/1606.05386 (cit. on p. 62).

— (2018). 'Anchors: High-Precision Model-Agnostic Explanations'. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, pp. 1527–1535. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16982 (cit. on pp. 62, 63, 152).

Richter, Michael M. and Rosina O. Weber (2013). *Case-Based Reasoning - A Textbook*. Springer. ISBN: 978-3-642-40166-4. DOI: 10.1007/978-3-642-40167-1. URL: https://doi.org/10.1007/978-3-642-40167-1 (cit. on pp. 26, 33, 40, 59).

Riesbeck, Christopher K. and Roger C. Schank (1989). 'Inside Case-Based Reasoning'. In: (cit. on p. 34).

Rigoni, Adam (2015). 'An improved factor based approach to precedential constraint'. In: *Artificial Intelligence and Law* 23, pp. 133–160 (cit. on pp. 36, 56, 57, 59, 150).

Rissland, Edwina L. and Kevin D. Ashley (1987). 'A Case-Based System for Trade Secrets Law'. In: *Proceedings of the First International Conference on Artificial Intelligence and Law, ICAIL '87, Boston, MA, USA, May 27-29, 1987*. ACM, pp. 60–66. ISBN: 0-89791-230-6. DOI: 10.1145/41735.41743. URL: https://doi.org/10.1145/41735.41743 (cit. on pp. 35, 43, 89).

Rissland, Edwina L., Kevin D. Ashley and Karl Branting (2005). 'Case-based reasoning and law'. In: *The Knowledge Engineering Review* 20, pp. 293–298 (cit. on p. 35).

Rissland, Edwina L. and David B. Skalak (1989). 'Combining Case-Based and Rule-Based Reasoning: A Heuristic Approach'. In: *Proceedings of the 11th International Joint Conference on Artificial Intelligence. Detroit, MI, USA, August 1989*. Ed. by N. S. Sridharan. Morgan Kaufmann, pp. 524–530. URL: http://ijcai.org/Proceedings/89-1/Papers/084.pdf (cit. on p. 35).

Rooij, Robert van and Katrin Schulz (2011). 'Non-Monotonic Reasoning in Interpretation'. In: 2nd ed. Elsevier. ISBN: 9780444537263. DOI: 10.1016/c2010-0-65666-5. URL: http://dx.doi.org/10.1016/c2010-0-65666-5 (cit. on p. 31).

Rotolo, Antonino and Giovanni Sartor (2018). 'Deductive and Deontic Reasoning'. In: *Handbook of Legal Reasoning and Argumentation*, pp. 243–274. DOI: 10.1007/978-90-481-9452-0_10. URL: http://dx.doi.org/10.1007/978-90-481-9452-0_10 (cit. on p. 40).

Ruder, Sebastian (2017). *An Overview of Multi-Task Learning in Deep Neural Networks*. arXiv: 1706.05098v1 [cs.LG] (cit. on p. 151).

Rudin, Cynthia, Caroline Wang and Beau Coker (Jan. 2020). 'The Age of Secrecy and Unfairness in Recidivism Prediction'. In: *Harvard Data Science Review* 2.1. DOI: 10.1162/99608f92.6ed64b30. URL: http://dx.doi.org/10.1162/99608f92.6ed64b30 (cit. on pp. 109, 147).

Russell, Stuart and Peter Norvig (2020). *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson. ISBN: 9780134610993. URL: http://aima.cs.berkeley.edu/ (cit. on pp. 25, 36).

Sartor, Giovanni (2018). 'Defeasibility in Law'. In: *Handbook of Legal Reasoning and Argumentation*, pp. 315–364. DOI: 10.1007/978-90-481-9452-0_12. URL: http://dx.doi.org/10.1007/978-90-481-9452-0_12 (cit. on pp. 26, 40).

Schauer, Frederick (2006). '(Re)Taking Hart'. In: *Harvard Law Review* 119.3. Ed. by Nicola Lacey, pp. 852–883. ISSN: 0017-811X. JSTOR: 4093593. URL: https://www.jstor.org/stable/4093593 (visited on 12/11/2023) (cit. on p. 52).

— (2008). 'A Critical Guide to Vehicles in the Park'. In: *NEW YORK UNIVERSITY LAW REVIEW* 83.4 (cit. on p. 52).

— (2009). *Thinking Like a Lawyer: A New Introduction to Legal Reasoning*. Cambridge, Massachusetts: Harvard University Press. ISBN: 0674032705 (cit. on pp. 27, 69, 70).

— (Sept. 2012). 'Is Defeasibility an Essential Property of Law?' In: *The Logic of Legal Requirements*, pp. 77–88.
DOI: 10.1093/acprof:oso/9780199661640.003.0005. URL: http://dx.doi.org/10.1093/acprof:oso/9780199661640.003.0005 (cit. on pp. 26, 33).

— (2014). 'Precedent'. In: *The Routledge Companion to Philosophy of Law*. DOI: 10.4324/9780203124352.ch9. URL: http://dx.doi.org/10.4324/9780203124352.ch9 (cit. on pp. 56, 57, 70).

Schlag, Pierre (1st Jan. 1999). 'No Vehicles in the Park'. In: *Seattle University Law Review* 23.2, p. 381. ISSN: 1078-1927. URL: https://digitalcommons.law.seattleu.edu/sulr/vol23/iss2/5 (cit. on p. 52).

Schmid, Ute, Christina Zeller, Tarek R. Besold, Alireza Tamaddoni-Nezhad and Stephen H. Muggleton (2016). 'How Does Predicate Invention Affect Human Comprehensibility?' In: *Inductive Logic Programming - 26th International Conference, ILP 2016, London, UK, September 4-6, 2016, Revised Selected Papers*. Ed. by James Cussens and Alessandra Russo. Vol. 10326. Lecture Notes in Computer Science. Springer, pp. 52–67. ISBN: 978-3-319-63341-1.

DOI: 10.1007/978-3-319-63342-8_5. URL: https://doi.org/10.1007/978-3-319-63342-8%5C_5 (cit. on p. 39).

Selbst, Andrew D and Julia Powles (Nov. 2017). 'Meaningful information and the right to explanation'. In: *International Data Privacy Law* 7.4, pp. 233–242. ISSN: 2044-4001.
DOI: 10.1093/idpl/ipx022. URL: http://dx.doi.org/10.1093/idpl/ipx022 (cit. on p. 38).

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh and Dhruv Batra (2017). 'Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization'. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, pp. 618–626. ISBN: 978-1-5386-1032-9.
DOI: 10.1109/ICCV.2017.74. URL: https://doi.org/10.1109/ICCV.2017.74 (cit. on p. 62).

Setzu, Mattia, Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi and Fosca Giannotti (2021). 'GLocalX - From Local to Global Explanations of Black Box AI Models'. In: *Artif. Intell.* 294, p. 103457.
DOI: 10.1016/j.artint.2021.103457. URL: https://doi.org/10.1016/j.artint.2021.103457 (cit. on p. 144).

Shecaira, Fábio Perin (2015). 'Sources of Law Are Not Legal Norms'. In: *Ratio Juris* 28.1, pp. 15–30. ISSN: 1467-9337. DOI: 10.1111/raju.12053. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/raju.12053 (visited on 12/11/2023) (cit. on p. 52).

Shepherdson, J. C. (1998). 'Negation as Failure, Completion and Stratification'. In: *Handbook of Logic in Artificial Intelligence and Logic Programming - Volume 5 - Logic Programming*. Ed. by Dov M. Gabbay, C. J. Hogger and J. A. Robinson. Oxford University Press (cit. on p. 31).

Shih, Andy, Arthur Choi and Adnan Darwiche (2019). 'Compiling Bayesian Network Classifiers into Decision Graphs'. In: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 7966–

7974. DOI: 10.1609/aaai.v33i01.33017966. URL: https://doi.org/10.1609/aaai.v33i01.33017966 (cit. on p. 72).

Smilkov, Daniel, Nikhil Thorat, Been Kim, Fernanda B. Viégas and Martin Wattenberg (2017). 'SmoothGrad: removing noise by adding noise'. In: *CoRR* abs/1706.03825. arXiv: 1706.03825. URL: http://arxiv.org/abs/1706.03825 (cit. on p. 63).

Smullyan, Raymond M. (1995). *First-Order Logic*. New York: Dover Publications Inc. 176 pp. ISBN: 978-0-486-68370-6 (cit. on p. 49).

Sokol, Kacper and Peter A. Flach (2020). 'One Explanation Does Not Fit All'. In: *KI - Künstliche Intelligenz* 34, pp. 235–250 (cit. on p. 41).

Spiegel, Lars-Phillip, Gabriele Kern-Isberner and Marco Ragni (2019). 'Rational Inference Patterns'. In: *PRICAI 2019: Trends in Artificial Intelligence - 16th Pacific Rim International Conference on Artificial Intelligence, Cuvu, Yanuca Island, Fiji, August 26-30, 2019, Proceedings, Part I*. Ed. by Abhaya C. Nayak and Alok Sharma. Vol. 11670. Lecture Notes in Computer Science. Springer, pp. 405–417. ISBN: 978-3-030-29907-1. DOI: 10.1007/978-3-030-29908-8_33. URL: https://doi.org/10.1007/978-3-030-29908-8%5C_33 (cit. on pp. 127, 133).

Sreedharan, Sarath, Tathagata Chakraborti and Subbarao Kambhampati (2021). 'Foundations of explanations as model reconciliation'. In: *Artif. Intell.* 301, p. 103558. DOI: 10.1016/j.artint.2021.103558. URL: https://doi.org/10.1016/j.artint.2021.103558 (cit. on p. 144).

Stenning, Keith and Michiel van Lambalgen (Nov. 2005). 'Semantic Interpretation as Computation in Nonmonotonic Logic: The Real Meaning of the Suppression Task'. In: *Cognitive Science* 29.6, pp. 919–960. ISSN: 0364-0213. DOI: 10.1207/s15516709cog0000_36. URL: http://dx.doi.org/10.1207/s15516709cog0000_36 (cit. on pp. 127, 133).

Stolzenburg, Frieder, Alejandro Javier García, Carlos Iván Chesñevar and Guillermo Ricardo Simari (2003). 'Computing Generalized Specificity'. In: *J. Appl. Non Class. Logics* 13.1, pp. 87–113. DOI: 10.3166/jancl.13.87-113. URL: https://doi.org/10.3166/jancl.13.87-113 (cit. on p. 136).

Strasser, Christian and G. Aldo Antonelli (2019). 'Non-monotonic Logic'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2019. Metaphysics Research Lab, Stanford University (cit. on p. 31).

Struchiner, Noel, Ivar R. Hannikainen and Guilherme da F. C. F. de Almeida (May 2020). 'An Experimental Guide to Vehicles in the Park'. In: *Judgment and Decision Making* 15.3, pp. 312–329. ISSN: 1930-2975. DOI: 10.1017/S1930297500007130. URL: https://www.cambridge.org/core/journals/judgment-and-decision-making/article/an-experimental-guide-to-vehicles-in-the-park/163BA48F57F5C4DD8BC886D143043AF7 (visited on 12/11/2023) (cit. on pp. 30, 52, 127).

Štrumbelj, Erik and Igor Kononenko (2010). 'An Efficient Explanation of Individual Classifications Using Game Theory'. In: *Journal of Machine Learning Research* 11.1, pp. 1–18. ISSN: 1533-7928. URL: http://jmlr.org/papers/v11/strumbelj10a.html (visited on 12/11/2023) (cit. on p. 63).

— (1st Dec. 2014). 'Explaining Prediction Models and Individual Predictions with Feature Contributions'. In: *Knowledge and Information Systems* 41.3, pp. 647–665. ISSN: 0219-3116. DOI: 10.1007/s10115-013-0679-x. URL: https://doi.org/10.1007/s10115-013-0679-x (visited on 10/11/2023) (cit. on pp. 63, 126, 152).

Sukpanichnant, Purin, Antonio Rago, Piyawat Lertvittayakumjorn and Francesca Toni (2021). 'Neural QBAFs: Explaining Neural Networks Under LRP-Based Argumentation Frameworks'. In: *AIxIA 2021 - Advances in Artificial Intelligence - 20th International Conference of the Italian Association for Artificial Intelligence, Virtual Event, December 1-3, 2021, Revised Selected Papers*. Ed. by Stefania Bandini, Francesca Gasparini, Viviana Mascardi, Matteo Palmonari and Giuseppe Vizzari. Vol. 13196. Lecture Notes in Computer Science. Springer, pp. 429–444. ISBN: 978-3-031-08420-1. DOI: 10.1007/978-3-031-08421-8_30. URL: https://doi.org/10.1007/978-3-031-08421-8%5C_30 (cit. on p. 39).

Sundararajan, Mukund and Amir Najmi (21st Nov. 2020). 'The Many Shapley Values for Model Explanation'. In: *Proceedings of the 37th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 9269–9278. URL: https://proceedings.mlr.press/v119/sundararajan20b.html (visited on 12/11/2023) (cit. on p. 63).

Tan, Jinzhe, H. Westermann and Karim Benyekhlef (2023). 'ChatGPT as an Artificial Lawyer?' In: AI4AJ@ICAIL. URL: https://ceur-ws.org/Vol-3435/short2.pdf (visited on 09/11/2023) (cit. on p. 36).

Tarski, Alfred (1930). 'On some fundamental concepts of metamathematics'. In: *Logic, Semantics, Metamathematics. Papers from 1923 to 1938*. Trans. by J. H. Woodger. Oxford Univ. Press (cit. on p. 30).

— (1936). 'On the concept of logical consequence'. In: *Logic, Semantics, Metamathematics. Papers from 1923 to 1938*. Trans. by J. H. Woodger. Oxford Univ. Press (cit. on p. 30).

— (1956). *Logic, Semantics, Metamathematics. Papers from 1923 to 1938*. Trans. by J. H. Woodger. Oxford Univ. Press (cit. on p. 48).

Teso, Stefano and Kristian Kersting (2019). 'Explanatory Interactive Machine Learning'. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*. Ed. by Vincent Conitzer, Gillian K. Hadfield and Shannon Vallor. ACM, pp. 239–245. ISBN: 978-1-4503-6324-2. DOI: 10.1145/3306618.3314293. URL: https://doi.org/10.1145/3306618.3314293 (cit. on pp. 41, 126).

Timmer, Sjoerd T., John-Jules Ch. Meyer, Henry Prakken, Silja Renooij and Bart Verheij (2015). 'Explaining Bayesian Networks Using Argumentation'. In: *Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 13th European Conference, ECSQARU 2015, Compiègne, France, July 15-17, 2015. Proceedings*. Ed. by Sébastien Destercke and Thierry Denoeux. Vol. 9161. Lecture Notes in Computer Science. Springer, pp. 83–92. ISBN: 978-3-319-20806-0. DOI: 10.1007/978-3-319-20807-7_8. URL: https://doi.org/10.1007/978-3-319-20807-7%5C_8 (cit. on p. 39).

Toni, Francesca (2014). 'A tutorial on assumption-based argumentation'. In: *Argument Comput.* 5.1, pp. 89–117. DOI: 10.1080/19462166.2013.869878. URL: https://doi.org/10.1080/19462166.2013.869878 (cit. on p. 31).

UK Cabinet Office (2021). *Ethics, Transparency and Accountability Framework for Automated Decision-Making*. URL: https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making/ethics-transparency-and-accountability-framework-for-automated-decision-making#case-studies (cit. on p. 37).

UK Department for Business, Energy & Industrial Strategy (2022a). *Establishing a pro-innovation approach to regulating AI*. URL: https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-

regulating-ai/establishing-a-pro-innovation-approach-to-regulating-ai-policy-statement#fn:17 (cit. on p. 37).

UK Department for Business, Energy & Industrial Strategy (2022b). *National AI Strategy*. URL: https://www.gov.uk/government/publications/national-ai-strategy/national-ai-strategy-html-version (cit. on p. 37).

Van Woerkom, Wijnand, Davide Grossi, Henry Prakken and Bart Verheij (19th June 2023). 'Hierarchical Precedential Constraint'. In: *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pp. 333–342. DOI: 10.1145/3594536.3595154. URL: https://dl.acm.org/doi/10.1145/3594536.3595154 (visited on 12/11/2023) (cit. on pp. 56, 59).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin (2017). 'Attention is All you Need'. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan and Roman Garnett, pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (cit. on p. 36).

Verheij, Bart (2017). 'Formalizing arguments, rules and cases'. In: *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law* (cit. on p. 34).

— (1st June 2020). 'Artificial Intelligence as Law'. In: *Artificial Intelligence and Law* 28.2, pp. 181–206. ISSN: 1572-8382. DOI: 10.1007/s10506-020-09266-0. URL: https://doi.org/10.1007/s10506-020-09266-0 (visited on 15/11/2023) (cit. on p. 69).

Wachter, Sandra, Brent Mittelstadt and Luciano Floridi (May 2017). 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation'. In: *International Data Privacy Law* 7.2, pp. 76–99. ISSN: 2044-4001. DOI: 10.1093/idpl/ipx005. URL: http://dx.doi.org/10.1093/idpl/ipx005 (cit. on p. 38).

Wachter, Sandra, Brent Mittelstadt and Chris Russell (2018). 'Counterfactual Explanations Without Opening The Black Box: Automated Decisions and

the GDPR'. In: *Harvard Journal of Law & Technology* 31.2 (cit. on pp. 126, 138, 152).

Wallach, Hanna M., Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox and Roman Garnett, eds. (2019). *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. URL: https://proceedings.neurips.cc/paper/2019.

Watson, David S., Limor Gultchin, Ankur Taly and Luciano Floridi (2021). 'Local explanations via necessity and sufficiency: unifying theory and practice'. In: *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*. Ed. by Cassio P. de Campos, Marloes H. Maathuis and Erik Quaeghebeur. Vol. 161. Proceedings of Machine Learning Research. AUAI Press, pp. 1382–1392. URL: https://proceedings.mlr.press/v161/watson21a.html (cit. on p. 41).

Watson, Ian D. and Farhi Marir (1994). 'Case-based reasoning: A review'. In: *Knowl. Eng. Rev.* 9.4, pp. 327–354. DOI: 10.1017/S0269888900007098. URL: https://doi.org/10.1017/S0269888900007098 (cit. on pp. 33, 35, 40).

Weidinger, Laura, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving and Iason Gabriel (21st June 2022). 'Taxonomy of Risks Posed by Language Models'. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229. DOI: 10.1145/3531146.3533088. URL: https://dl.acm.org/doi/10.1145/3531146.3533088 (visited on 09/11/2023) (cit. on p. 36).

Wirth, Claus-Peter and Frieder Stolzenburg (2016). 'A series of revisions of David Poole's specificity'. In: *Ann. Math. Artif. Intell.* 78.3-4, pp. 205–258. DOI: 10.1007/s10472-015-9471-9. URL: https://doi.org/10.1007/s10472-015-9471-9 (cit. on p. 136).

Woerkom, Wijnand van, Davide Grossi, Henry Prakken and Bart Verheij (2022). 'Landmarks in Case-Based Reasoning: From Theory to Data'. In: *HHAI 2022: Augmenting Human Intellect - Proceedings of the First International Conference on Hybrid Human-Artificial Intelligence, Amsterdam, The Neth-*

erlands, 13-17 June 2022. Ed. by Stefan Schlobach, María Pérez-Ortiz and Myrthe Tielman. Vol. 354. Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 212–224. ISBN: 978-1-64368-308-9.
DOI: 10.3233/FAIA220200. URL: https://doi.org/10.3233/FAIA220200 (cit. on pp. 109, 121).

Ye, Xi and Greg Durrett (16th May 2022). 'The Unreliability of Explanations in Few-shot Prompting for Textual Reasoning'. In: Advances in Neural Information Processing Systems. URL: https://openreview.net/forum?id=Bct2f8fRd8S (visited on 09/11/2023) (cit. on p. 36).

Zhong, Qiaoting, Xiuyi Fan, Xudong Luo and Francesca Toni (2019). 'An explainable multi-attribute decision model based on argumentation'. In: *Expert Syst. Appl.* 117, pp. 42–61. DOI: 10.1016/j.eswa.2018.09.038. URL: https://doi.org/10.1016/j.eswa.2018.09.038 (cit. on p. 39).

Zhou, Yilun, Marco Túlio Ribeiro and Julie Shah (2022). 'ExSum: From Local Explanations to Model Understanding'. In: *CoRR* abs/2205.00130.
DOI: 10.48550/arXiv.2205.00130. arXiv: 2205.00130. URL: https://doi.org/10.48550/arXiv.2205.00130 (cit. on pp. 126, 144).

Appendices

Appendix A

Proofs of Theorems

A.1 Correctness of algorithms - Proof of Theorem 3.18

We will need a lemma about the addition of new cases. We will first present and prove this lemma, and then move to the proof of the main theorem.

A.1.1 Addition of new cases

The next result characterises the set of past cases/arguments attacked when the dataset is extended with a new labelled case/argument. In particular, this result compares the effect of predicting the outcome of some N_2 from D alone and from D extended with (N_1, o_1), when there is no case in D with characterisation N_1 already and moreover D is coherent.

This result is interesting in its own right as it shows that, any argument attacked by the "newly added" case (N_1, o_1) is easily identified in the sets G_0 and G_1 in the grounded extension $\mathbb{G}$, being sufficient to check those rather than the entire casebase D. Notice that we require D to be coherent.

Lemma A.1. *Let D be coherent, $N_1, N_2 \in X$, $o_1 \in Y$, and suppose that there is no case in D with characterisation N_1. Consider $AF_1 = AF_{\succeq}(D, N_1)$ and $AF_2 = AF_{\succeq}(D \cup \{(N_1, o_1)\}, N_2)$. Finally, let $\mathbb{G}(AF_1)$and $\mathbb{G}(AF_2)$ be the respective grounded extensions. Let $\beta \in D$ be such that $(N_1, o_1) \rightsquigarrow \beta$ in AF_2. Then,*

1. *for every γ that attacks β in AF_1, $N_1 \not\sim \gamma$ (that is, γ is irrelevant to N_1 and, by regularity, $N_1 \not\succeq \gamma$);*

2. *in AF_1, $(N_1, ?)$ defends β;*

3. *$\beta \in \mathbb{G}(AF_1)$ and, for $\mathbb{G}(AF_1) = \bigcup_{i \geqslant 0} G_i$, β is either in G_0 (that it, it is unattacked), or in G_1.*

4. *For every $\eta = (\eta_C, \eta_o) \in D$ such that $(N_1, ?)$ defends η in AF_1, if $\eta_o \neq o_1$, then, in AF_2, $(N_1, o_1) \rightsquigarrow \eta$.*

Proof. 1. Let $\beta = (\beta_C, \beta_o)$. From the definition of attack: (i) $N_1 \succ \beta_C$ (this is strict due to coherence and since no case in D has characterisation N_1), (ii) $o_1 \neq \beta_o$, and (iii) there is no (α_C, x_o) such that $x_o = o_1$ and $N_1 \succ \alpha_C \succ \beta_C$. Consider $\epsilon = (\epsilon_C, \epsilon_o)$ such that ϵ attacks β in AF_1 (if there is no such ϵ then the result trivially holds). Assume by contradiction that ϵ is relevant to N_1. Then by regularity $N_1 \succeq \epsilon_C$. But since D is coherent and $(N_1, o_1) \notin D$, ϵ and N_1 are distinct, and thus $N_1 \succ \epsilon_C$. As ϵ attacks β, $\epsilon_o \neq \beta_o$, but this in turn implies that $\epsilon_o = o_1$, since (N_1, o_1) also attacks β, in AF_2. But then $N_1 \succ \epsilon_C \succ \beta_C$, with $\epsilon_o = o_1$. This contradicts requirement 3 in the second bullet of Definition 2.10 of the attack between (N_1, o_1) and β. Therefore, ϵ is not relevant to N_1, as we wanted to prove.

2. Trivially true, by 1 (as, if ϵ is an attacker β, then $N_1 \not\sim \epsilon$; but then $(N_1, ?) \rightsquigarrow \epsilon$).

3. Trivially true, by 2.

4. Since, in AF_1, $(N_1, ?)$ defends η, then any attacker ϵ of η is irrelevant to N_1, and by regularity, $N_1 \not\succeq \epsilon$. Thus requirement 3 in the second bullet of Definition 2.10 is satisfied. Requirement 1 is the hypothesis and requirement 2 is satisfied since $(N_1, ?)$ defends η in AF_1. □

A.1.2 Correctness

Now we can show the theorem. We will first prove that Algorithm 1 is correct.

Theorem A.2 (Correctness of Algorithm 1)**.** *Every execution of Algorithm 1 (simple_add*$((Args, \rightsquigarrow), next\_case)$*) inside Algorithm 2 correctly returns* $AF_{\succeq}(Args \cup \{next\_case\})$.

Proof. This is essentially a consequence of Lemma A.1. We know that there will never be an argument in *Args* with the same characterisation as *next_case*, since they will occur in the same stratum, thus the lemma applies. The lemma guarantees that Algorithm 1 adds all attacks that need to be added and only those. Finally, we need to check that it will never be necessary to remove an attack. This is true due to requirement 3 in the second bullet of Definition 2.10, and since arguments are added following the partial order. Therefore the only modifications on the set of attacks are the ones in *simple_add*. □

We can finally go to the proof of Theorem 3.18, on the correctness of Algorithm 2.

Proof of Theorem 3.18. 1. Every case in D' is surprising w.r.t. D'.

Algorithm 2 explicitly only adds a case if it is surprising. Since coherence is maintained at every step of the algorithm and it proceeds by following the partial order, one can see that after a case (α_C, α_o) is added, the outcome predicted for a new case $(\alpha_C, ?)$ is always α_o, by Lemma 3.9. Thus every case in D' is surprising w.r.t. D'.

2. If D has a concise subset, it is found.

Now, suppose that D has a concise subset D''. We already know by Theorem 3.16 that it is the unique concise subset. Now we need to show that $D' = D''$. We will show it by contradiction, in a similar fashion to the proof of Theorem 3.16:

Let then $(x, y) \in (D' \setminus D'') \cup (D'' \setminus D')$ (the symmetric difference between D' and D'') such that (x, y) is $\preceq$-minimal in this set. It exists since the sets are different (there is at least one element in one of the sets and not in the other) and this set is finite. Then we have that:

$$\{(x', y') \in D' \mid (x', y') \prec (x, y)\} =$$
$$\{(x', y') \in D'' \mid (x', y') \prec (x, y)\}$$

otherwise (x, y) would not be minimal in the symmetrical difference.

However, every concise set is coherent. Besides, at every step, the currently selected subset of D is also coherent. Therefore in the equality above we can replace $\prec$ with $\preceq$, and by applying Lemma 3.9 we will conclude that either (x, y) is includable w.r.t. both D' and D'' or to neither. Since D'' is concise, it must be includable w.r.t. both, and thus $(x, y) \in D''$.

Now consider the moment when (x, y) was considered in Algorithm 2. At that point, every case in $\{(x', y') \in D' \mid (x', y') \prec (x, y)\}$ was already in *Args*, since Algorithm 2 follows the partial order. Thus it was considered to be includable if and only if it is includable w.r.t. D'. However, we already know that it is includable w.r.t. D'. Thus it was considered to be includable at that point. Thus, it was added, and then $(x, y) \in D'$. This is absurd, since we defined (x, y) as an element of $(x, y) \in (D' \setminus D'') \cup (D'' \setminus D')$.

3. A coherent dataset has a concise subset.

   We will show this by showing that, if the input dataset is coherent, then the dataset underpinning the AF resulting from Algorithm 2 is concise.

   In order to prove that, for the returned *Args*, $Args \setminus \{(\delta_C, \delta_o)\}$ is concise, we just need to prove that at the end of each loop $Args \setminus \{(\delta_C, \delta_o)\}$ is concise w.r.t. the set of all seen examples.

   As the base case, before the loop is entered, this is clearly the case, as the only seen argument is the default.

   As the induction step, we know that every case previously added is still includable, since the new cases added are not less specific than them according to the partial order, and thus by Lemma 3.9 their prediction is not changed, that is, they keep being includable. The same is true for every case previously not added: adding more cases afterwards does not change their prediction. For the cases added at this new iteration, by definition the includable ones are added and the not includable ones are not. Regarding the order in which cases of the same stratum are added, each of the includable cases will be included and the not includable ones

will not be. It can be seen that the order is irrelevant as, since they are all $\preceq$-minimal and the dataset is coherent, they are incomparable, so each case in the list is irrelevant with respect to the other. Thus, for every case seen until this point, it is in the AF iff it is includable. As this is true for every iteration, it is true for the final, returned AF. □

A.2 Spikes - Proof of Theorems 3.24 and 3.25

Using previous results, we are now ready to present a proof sketch of Theorem 3.24 and a full proof of Theorem 3.25, the more interesting result.

Proof sketch of Theorem 3.24. One may remember that the definition of the grounded extension is of the form $\mathbb{G} = \bigcup_{i \geqslant 0} G_i$ and verify by induction on i that whether an argument is on the grounded extension or not is a function of whether the arguments that can reach it are or not in the grounded extension. This may be more clear using the method of labellings (Modgil and Caminada, 2009). Thus whether $(\delta_C, \delta_o) \in \mathbb{G}$ or not does not depend on a spike. □

Proof of Theorem 3.25. We will prove it by contradiction. Let α be a spike and assume $\alpha \in \mathit{Args}$. Then, during the execution of Algorithm 2, at some moment α was added to `to_add`. Indeed, assume additionally that α is the first such spike to be added. In order to this to happen, the predicted outcome at the moment for α was not α_o. By (the contrapositive of) Theorem 3.7, this implies that there is a case β which is a nearest case to α with the different outcome, that is, $\beta_o \neq \alpha_o$. It is straightforward to check that every nearest case of a new case is defended by it. Thus, by Lemma A.1, item 4, when α is added to *Args*, α attacks β.

Now notice that, since Algorithm 2 goes by strata, and β is in the AF at this point of the algorithm, then it also is at the final AF, that is, $\beta \in \mathit{Args}$, as is the attack from α to β. However, since α is a spike, it has no path to the default case, and thus β also has no path to the default case, that is, it is also a spike. This is absurd, since it contradicts our assumption that α is the first spike to be added to *Args*. Therefore, no spike is added to *Args*. □

Appendix B

Case Study - Description of Factors

Table B.1: List of factors for the U.S. Trade Secrets domain, with their descriptions, as presented by Grabmair (2016).

Feature	Description
F_{Δ}^{1}	plaintiff disclosed its product information in negotiations with defendant
F_{Π}^{2}	defendant paid plaintiff's former employee to switch employment, apparently in an attempt to induce the employee to bring plaintiff's information
F_{Δ}^{3}	defendant's employee was the sole developer of plaintiff's product
F_{Π}^{4}	defendant entered into a nondisclosure agreement with plaintiff
F_{Δ}^{5}	the nondisclosure agreement did not specify which information was to be treated as confidential
F_{Π}^{6}	plaintiff took active measures to limit access to and distribution of its information
F_{Π}^{7}	plaintiff's former employee brought product development information to defendant
F_{Π}^{8}	defendant's access to plaintiff's product information saved it time or expense
F_{Δ}^{10}	plaintiff disclosed its product information to outsiders

Table B.1: List of factors for the U.S. Trade Secrets domain (continued).

Feature	Description
F_{Δ}^{11}	plaintiff's information was about customers and suppliers (i.e. it may have been available independently from customers or even in directories)
F_{Π}^{12}	plaintiff's disclosures to outsiders were subject to confidentiality restrictions
F_{Π}^{13}	plaintiff and defendant entered into a noncompetition agreement
F_{Π}^{14}	defendant used materials that were subject to confidentiality restrictions
F_{Π}^{15}	plaintiff's information was unique in that plaintiff was the only manufacturer making the product
F_{Δ}^{16}	plaintiff's product information could be learned by reverse-engineering
F_{Δ}^{17}	defendant developed its product by independent research
F_{Π}^{18}	defendant's product was identical to plaintiff's
F_{Δ}^{19}	plaintiff did not adopt any security measures
F_{Δ}^{20}	plaintiff's information was known to competitors
F_{Π}^{21}	defendant obtained plaintiff's information although he knew that plaintiff's information was confidential
F_{Π}^{22}	defendant used invasive techniques to gain access to plaintiff's information
F_{Δ}^{23}	plaintiff entered into an agreement waiving confidentiality
F_{Δ}^{24}	the information could be obtained from publicly available sources
F_{Δ}^{25}	defendant discovered plaintiff's information through reverse engineering
F_{Π}^{26}	defendant obtained plaintiff's information through deception
F_{Δ}^{27}	plaintiff disclosed its information in a public forum

Appendix C

Learning Relevance in Case-Based Reasoning: Supplementary Result Tables

Table C.1: Comparison between $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$ for percentage accuracy, over different choices of decision tree hyperparameters (maximum depth and maximum number of leaves) and over different strategies on dealing with incoherence. Comparisons are done over a single train-test split. Columns representing difference calculate the difference between the $cAA\text{-}CBR_{\succeq}$ and the $AA\text{-}CBR_{\succeq}$ values.

	strategy	keep			removal			majority		
	model	$AA\text{-}CBR_{\succeq}$	$cAA\text{-}CBR_{\succeq}$	diff.	$AA\text{-}CBR_{\succeq}$	$cAA\text{-}CBR_{\succeq}$	diff.	$AA\text{-}CBR_{\succeq}$	$cAA\text{-}CBR_{\succeq}$	diff.
DT max depth	DT max leaves									
3	4	45.57	50.65	5.08	54.43	54.43	0.00	67.22	67.22	0.00
	8	45.57	46.98	1.40	56.80	56.80	0.00	67.98	67.98	0.00
5	4	45.57	50.65	5.08	54.43	54.43	0.00	67.22	67.22	0.00
	8	45.57	49.46	3.89	56.80	56.80	0.00	68.09	68.09	0.00
	32	51.94	53.02	1.08	63.93	61.88	-2.05	63.93	63.55	-0.38
7	4	45.57	50.65	5.08	54.43	54.43	0.00	67.22	67.22	0.00
	8	45.57	49.46	3.89	56.80	56.80	0.00	68.09	68.09	0.00
	32	46.87	51.03	4.16	59.88	61.45	1.57	65.01	64.96	-0.05
	128	52.97	55.29	2.32	56.91	57.51	0.59	58.64	58.75	0.11
9	4	45.57	50.65	5.08	54.43	54.43	0.00	67.22	67.22	0.00
	8	45.57	49.46	3.89	56.80	56.80	0.00	68.09	68.09	0.00
	32	46.87	51.03	4.16	59.88	61.45	1.57	65.01	64.96	-0.05
	128	54.43	57.83	3.40	58.48	58.80	0.32	60.53	60.75	0.22
	512	54.81	55.62	0.81	58.32	57.67	-0.65	58.75	58.15	-0.59
11	4	45.57	50.65	5.08	54.43	54.43	0.00	67.22	67.22	0.00
	8	45.57	49.46	3.89	56.80	56.80	0.00	68.09	68.09	0.00
	32	46.87	51.03	4.16	59.88	61.45	1.57	65.01	64.96	-0.05
	128	54.21	57.83	3.62	56.91	57.72	0.81	59.99	59.83	-0.16
	512	55.35	55.94	0.59	57.67	57.02	-0.65	58.80	57.94	-0.86
	2048	55.35	55.94	0.59	57.67	57.02	-0.65	58.80	57.94	-0.86
13	4	45.57	50.65	5.08	54.43	54.43	0.00	67.22	67.22	0.00
	8	45.57	49.46	3.89	56.80	56.80	0.00	68.09	68.09	0.00
	32	46.87	51.03	4.16	59.88	61.45	1.57	65.01	64.96	-0.05
	128	53.46	55.51	2.05	57.56	58.10	0.54	59.23	58.75	-0.49
	512	55.29	55.45	0.16	57.45	57.02	-0.43	58.48	57.61	-0.86
	2048	55.29	55.62	0.32	57.34	57.02	-0.32	58.21	57.45	-0.76

Table C.2: Comparison between $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$ for number of nodes, over different choices of decision tree hyperparameters (maximum depth and maximum number of leaves) and over different strategies on dealing with incoherence. Comparisons are done over a single train-test split. Columns representing ratio calculate the ratio of the $cAA\text{-}CBR_{\succeq}$ to the $AA\text{-}CBR_{\succeq}$ values.

	strategy	keep			removal			majority		
	model	$AA\text{-}CBR_{\succeq}$	$cAA\text{-}CBR_{\succeq}$	ratio	$AA\text{-}CBR_{\succeq}$	$cAA\text{-}CBR_{\succeq}$	ratio	$AA\text{-}CBR_{\succeq}$	$cAA\text{-}CBR_{\succeq}$	ratio
DT max depth	DT max leaves									
3	4	13	7	54%	1	1	100%	6	4	67%
	8	22	12	55%	2	2	100%	7	6	86%
5	4	13	7	54%	1	1	100%	6	4	67%
	8	47	24	51%	2	2	100%	13	7	54%
	32	1697	863	51%	606	345	57%	959	570	59%
7	4	13	7	54%	1	1	100%	6	4	67%
	8	47	24	51%	2	2	100%	13	7	54%
	32	1152	581	50%	338	206	61%	594	346	58%
	128	2633	1484	56%	1843	1128	61%	2088	1254	60%
9	4	13	7	54%	1	1	100%	6	4	67%
	8	47	24	51%	2	2	100%	13	7	54%
	32	1152	581	50%	338	206	61%	594	346	58%
	128	2856	1547	54%	2160	1212	56%	2391	1328	56%
	512	2946	1681	57%	2378	1393	59%	2565	1498	58%
11	4	13	7	54%	1	1	100%	6	4	67%
	8	47	24	51%	2	2	100%	13	7	54%
	32	1152	581	50%	338	206	61%	594	346	58%
	128	2802	1591	57%	2116	1267	60%	2359	1390	59%
	512	3023	1733	57%	2492	1468	59%	2685	1576	59%
	2048	3023	1733	57%	2492	1468	59%	2685	1576	59%
13	4	13	7	54%	1	1	100%	6	4	67%
	8	47	24	51%	2	2	100%	13	7	54%
	32	1152	581	50%	338	206	61%	594	346	58%
	128	2773	1547	56%	2087	1219	58%	2322	1350	58%
	512	3040	1759	58%	2528	1497	59%	2716	1604	59%
	2048	3043	1759	58%	2533	1498	59%	2720	1606	59%

Table C.3: Comparison between $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$ for number of attacks, over different choices of decision tree hyperparameters (maximum depth and maximum number of leaves) and over different strategies on dealing with incoherence. Comparisons are done over a single train-test split. Columns representing ratio calculate the ratio of the $cAA\text{-}CBR_{\succeq}$ to the $AA\text{-}CBR_{\succeq}$ values.

	strategy	keep			removal			majority		
	model	$AA\text{-}CBR_{\succeq}$	$cAA\text{-}CBR_{\succeq}$	ratio	$AA\text{-}CBR_{\succeq}$	$cAA\text{-}CBR_{\succeq}$	ratio	$AA\text{-}CBR_{\succeq}$	$cAA\text{-}CBR_{\succeq}$	ratio
DT max depth	DT max leaves									
3	4	29	8	28%	0	0	–	6	3	50%
	8	51	16	31%	1	1	100%	7	5	71%
5	4	29	8	28%	0	0	–	6	3	50%
	8	134	44	33%	1	1	100%	14	6	43%
	32	7121	3273	46%	2207	1276	58%	3569	2189	61%
7	4	29	8	28%	0	0	–	6	3	50%
	8	134	44	33%	1	1	100%	14	6	43%
	32	4186	1861	44%	1011	613	61%	1865	1041	56%
	128	11815	6428	54%	8215	4601	56%	9282	5434	59%
9	4	29	8	28%	0	0	–	6	3	50%
	8	134	44	33%	1	1	100%	14	6	43%
	32	4186	1861	44%	1011	613	61%	1865	1041	56%
	128	15059	7964	53%	11689	6167	53%	12810	6970	54%
	512	15200	8430	55%	12520	6673	53%	13417	7389	55%
11	4	29	8	28%	0	0	–	6	3	50%
	8	134	44	33%	1	1	100%	14	6	43%
	32	4186	1861	44%	1011	613	61%	1865	1041	56%
	128	13674	7522	55%	10464	5968	57%	11550	6695	58%
	512	16280	9048	56%	13652	7559	55%	14589	8260	57%
	2048	16280	9048	56%	13652	7559	55%	14589	8260	57%
13	4	29	8	28%	0	0	–	6	3	50%
	8	134	44	33%	1	1	100%	14	6	43%
	32	4186	1861	44%	1011	613	61%	1865	1041	56%
	128	13558	7497	55%	10266	5767	56%	11346	6588	58%
	512	16584	9369	56%	14051	7854	56%	14941	8586	57%
	2048	16593	9356	56%	14065	7868	56%	14947	8591	57%

Table C.4: Comparison between $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$ for maximum depth of the AF, over different choices of decision tree hyperparameters (maximum depth and maximum number of leaves) and over different strategies on dealing with incoherence. Comparisons are done over a single train-test split. Columns representing difference calculate the difference between the $cAA\text{-}CBR_{\succeq}$ and the $AA\text{-}CBR_{\succeq}$ values.

	strategy	keep			removal			majority		
	model	$AA\text{-}CBR_{\succeq}$	$cAA\text{-}CBR_{\succeq}$	diff.	$AA\text{-}CBR_{\succeq}$	$cAA\text{-}CBR_{\succeq}$	diff.	$AA\text{-}CBR_{\succeq}$	$cAA\text{-}CBR_{\succeq}$	diff.
DT max depth	DT max leaves									
3	4	4	4	0	0	0	0	3	3	0
	8	6	6	0	1	1	0	3	5	2
5	4	4	4	0	0	0	0	3	3	0
	8	7	7	0	1	1	0	5	6	1
	32	9	8	-1	7	5	-2	8	6	-2
7	4	4	4	0	0	0	0	3	3	0
	8	7	7	0	1	1	0	5	6	1
	32	9	7	-2	8	7	-1	7	7	0
	128	6	6	0	6	4	-2	6	4	-2
9	4	4	4	0	0	0	0	3	3	0
	8	7	7	0	1	1	0	5	6	1
	32	9	7	-2	8	7	-1	7	7	0
	128	6	4	-2	6	4	-2	6	4	-2
	512	6	5	-1	5	4	-1	6	4	-2
11	4	4	4	0	0	0	0	3	3	0
	8	7	7	0	1	1	0	5	6	1
	32	9	7	-2	8	7	-1	7	7	0
	128	6	4	-2	6	4	-2	6	4	-2
	512	6	5	-1	5	4	-1	6	4	-2
	2048	6	5	-1	5	4	-1	6	4	-2
13	4	4	4	0	0	0	0	3	3	0
	8	7	7	0	1	1	0	5	6	1
	32	9	7	-2	8	7	-1	7	7	0
	128	6	5	-1	6	4	-2	6	4	-2
	512	6	5	-1	5	4	-1	6	4	-2
	2048	6	5	-1	5	4	-1	6	4	-2

Table C.5: Comparison between $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$ for average depth of the AF, over different choices of decision tree hyperparameters (maximum depth and maximum number of leaves) and over different strategies on dealing with incoherence. Comparisons are done over a single train-test split. Columns representing difference calculate the difference between the $cAA\text{-}CBR_{\succeq}$ and the $AA\text{-}CBR_{\succeq}$ values.

	strategy	keep			removal			majority		
	model	$AA\text{-}CBR_{\succeq}$	$cAA\text{-}CBR_{\succeq}$	diff.	$AA\text{-}CBR_{\succeq}$	$cAA\text{-}CBR_{\succeq}$	diff.	$AA\text{-}CBR_{\succeq}$	$cAA\text{-}CBR_{\succeq}$	diff.
DT max depth	DT max leaves									
3	4	2.00	2.14	0.14	0.00	0.00	0.00	1.50	1.50	0.00
	8	2.95	3.17	0.21	0.50	0.50	0.00	1.71	2.50	0.79
5	4	2.00	2.14	0.14	0.00	0.00	0.00	1.50	1.50	0.00
	8	3.49	3.83	0.34	0.50	0.50	0.00	2.38	3.00	0.62
	32	4.62	4.05	-0.57	3.55	3.18	-0.37	4.11	3.64	-0.48
7	4	2.00	2.14	0.14	0.00	0.00	0.00	1.50	1.50	0.00
	8	3.49	3.83	0.34	0.50	0.50	0.00	2.38	3.00	0.62
	32	4.75	4.11	-0.64	3.42	3.07	-0.35	3.97	3.46	-0.51
	128	3.12	2.75	-0.37	2.83	2.51	-0.32	2.91	2.60	-0.32
9	4	2.00	2.14	0.14	0.00	0.00	0.00	1.50	1.50	0.00
	8	3.49	3.83	0.34	0.50	0.50	0.00	2.38	3.00	0.62
	32	4.75	4.11	-0.64	3.42	3.07	-0.35	3.97	3.46	-0.51
	128	3.08	2.74	-0.34	2.87	2.54	-0.33	2.94	2.61	-0.33
	512	2.92	2.57	-0.35	2.75	2.38	-0.36	2.80	2.45	-0.35
11	4	2.00	2.14	0.14	0.00	0.00	0.00	1.50	1.50	0.00
	8	3.49	3.83	0.34	0.50	0.50	0.00	2.38	3.00	0.62
	32	4.75	4.11	-0.64	3.42	3.07	-0.35	3.97	3.46	-0.51
	128	3.00	2.66	-0.34	2.78	2.45	-0.33	2.86	2.54	-0.32
	512	2.91	2.56	-0.35	2.74	2.38	-0.36	2.80	2.45	-0.35
	2048	2.91	2.56	-0.35	2.74	2.38	-0.36	2.80	2.45	-0.35
13	4	2.00	2.14	0.14	0.00	0.00	0.00	1.50	1.50	0.00
	8	3.49	3.83	0.34	0.50	0.50	0.00	2.38	3.00	0.62
	32	4.75	4.11	-0.64	3.42	3.07	-0.35	3.97	3.46	-0.51
	128	3.02	2.66	-0.36	2.81	2.46	-0.35	2.88	2.54	-0.34
	512	2.90	2.56	-0.34	2.74	2.39	-0.35	2.79	2.45	-0.34
	2048	2.90	2.55	-0.35	2.74	2.38	-0.36	2.79	2.45	-0.34