A Fortiori Case-Based Reasoning

Formal Studies with Applications in Artificial Intelligence and Law

Wijnand van Woerkom

A Fortiori Case-Based Reasoning

Formal Studies with Applications in Artificial Intelligence and Law

Wijnand van Woerkom

Universiteit Utrecht

SIKS Dissertation Series No. 2025–56
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

DOI:	10.33540/2994
ISBN:	978-90-393-7963-9
Printed by:	Scharlau
Cover by:	Riccardo Rigo

A Fortiori Case-Based Reasoning

Formal Studies with Applications in Artificial Intelligence and Law

A Fortiori Casusgebaseerd Redeneren

Formele Studies met Toepassingen in Kunstmatige Intelligentie en Recht

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht

op gezag van de rector magnificus, prof. dr. ir. W. Hazeleger,

ingevolge het besluit van het college voor promoties

in het openbaar te verdedigen

op woensdag 15 oktober 2025 des middags te 12.15 uur

door

Wijnand van Woerkom

geboren op 8 augustus 1992 te Zaandam

Promotoren:

Prof. dr. H. Prakken

Prof. dr. D. Grossi

Prof. dr. H.B. Verheij

Beoordelingscommissie:

Prof. dr. F.J. Bex

Prof. dr. ir. J. Broersen

Prof. dr. J. Horty

Prof. dr. G. Sartor

Prof. dr. A. ten Teije

This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, grant number 024.004.022.

Voor mijn ouders

Contents

Chapter 1

Introduction

ASE-BASED REASONING (CBR) involves comparing a problem to prior cases to guide decision-making and draw conclusions (Ashley, 1992). This type of reasoning is fundamental to *common law* jurisdictions, where courts are constrained to abide by precedent cases according to the *stare decisis* principle. In these systems, attorneys compare current disputes to past cases to construct arguments, while judges use past cases to justify and explain their conclusions. More recently, CBR is found in data-driven machine learning systems, which rely on datasets of input-output pairs to produce output for new input in a way that appropriately generalizes the previously seen data. This has been compared to the way in which courts abide by precedent (Čyras et al., 2016; Prakken & Ratsma, 2022).

It comes as no surprise, then, that CBR has been a focal point of the literature on artificial intelligence (AI) & law. A prototypical example in this line of work is the HYPO computer program, designed to generate legal arguments about who should win a dispute on trade secret law (Ashley, 1991; Rissland & Ashley, 1987). It does so chiefly on the basis of citations of prior cases called precedents. Cases are represented by the use of *factors*—in the words of Ashley (1991, Section 2):

> Factors [...] are generalizations. Unlike rules, they do not specify necessary and sufficient conditions for a conclusion. Instead, they designate collections of facts, commonly observed in cases, that tend to strengthen or weaken a plaintiff's argument in favor of a conclusion, such as a legal conclusion that the plaintiff has a trade secret.

It is difficult to define exactly what a trade secret is, and so factors are used to describe aspects of the situation at hand which make the case stronger or weaker for the party claiming to have the trade secret. A few examples of factors in this domain are:

- The extent to which the information is known to outsiders.
- The value of the information to the owner of the information, and to competitors.
- The amount of effort or money expended by the owner on developing the information.

The HYPO program can generate a 3-ply legal argument for both sides of a dispute, by sequentially *analogizing*, *distinguishing*, and *rebutting* citations of precedent cases, on the basis of similarities and differences in the factor representations of the precedent cases and some novel problem situation. This innovative use of factors as a knowledge representation device, on the basis of which precedents can be compared and contrasted with novel fact situations, has become an essential aspect of many subsequent works.

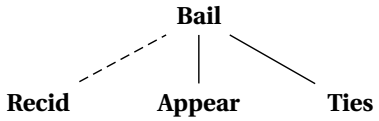
Several programs akin to HYPO were developed in its wake. What could be considered its immediate successor was the CATO program—an intelligent learning environment designed to teach case-based legal reasoning to law students (Aleven, 2003; Aleven & Ashley, 1997). In many ways, CATO is comparable to HYPO: It builds arguments for either side of a legal dispute by citing precedent cases, which are represented using factors. However, in contrast to HYPO, the CATO program constructs multi-case arguments organized around *issues*—the key legal questions that arise from the facts of a case, and that courts often address when explaining their decisions. Such arguments are constructed by CATO through the use of middle-level normative background knowledge for the domain of law under consideration, which is represented as a *factor hierarchy*. The idea of such a hierarchy is that the factors that are used to represent cases provide support or opposition to higher-level normative concepts called *abstract factors*. For example, the factor that "the information under consideration in a trade secrets dispute is not known outside of the plaintiff's business" has influence on the more abstract factor that this information is valuable, which in turn bears on the legal issue of whether the information is a trade secret. Legal reasoning, according to this model, proceeds in a stepwise fashion from the lower-level factors through the higher-level factors, ultimately deciding on the legal issues, which then form the basis of a decision for either the plaintiff or the defendant. As HYPO pioneered the use of factors as a knowledge representation device for cases, so did CATO pioneer the use of factor hierarchies. Continuing this line of work, the IBP and VJAP programs were presented by Bruninghaus and Ashley (2003) and Grabmair (2017), developed for the purpose of predicting legal case outcomes, and both of which operate on the basis of factor hierarchies.

Aside from spurring the further development of programs designed to produce legal arguments, or even to predict case outcomes, HYPO has also inspired work of a more theoretical character. In particular, a line of work developed which focused on giving a formal account of *precedential constraint*—the way in which precedent cases constrain future decision-making according to the *stare decisis* principle. Notably, building on the reason-based logic of Hage and Verheij (1994) and the formal theory of legal CBR by Prakken and Sartor (1998), Roth (2003) represented cases using a logical language and formulated a fundamental *a fortiori* principle of constraint. This principle roughly states that in a novel fact situation the same decision should be reached as in a precedent case, if the novel fact situation has equal-or-more support for that decision than the precedent case did, as measured in terms of their factor representations. Concurrently, Horty (2004, 2011) concisely isolated this same principle in his *result model* (RM) of precedential constraint, again using a factor-based representation of cases. The RM, and extensions thereof, have been influential in the literature on formal precedential constraint, and are foundational to this thesis.

1.1 A theory of a fortiori case-based reasoning

The RM describes the type of reasoning performed by a decision-maker, such as a court or a judge, when it tries to adhere to past decisions called precedent cases. The model describes when a new decision is, or is not, consistent with respect to the precedents. In other words, it normatively describes the way in which a set of precedents constrains future decision-making. The RM works on the basis of a knowledge representation using factors, which, as mentioned, are legally relevant fact patterns that are assumed to favor either a decision for the plaintiff or for the defendant of the case.

Example 1.1. Consider, as an example, the decision of a judge on whether to release a subject on bail. Bail is a sum of money that the defendant must pay to the court as a guarantee that they will appear at their trial—if the defendant does not appear, the bail is forfeited. The decision to grant bail is influenced by several factors: Is the person at risk of recidivism? Do they have a history of appearing for trial? Do they have strong communal ties? Let us assume, for the sake of simplicity, that these are the only relevant factors to consider. To come to a decision on whether to release the defendant on bail, the judge will then weigh the answers to these questions, each of which either support or oppose the conclusion to grant bail. In this example, the decision to grant bail is opposed by the presence of a recidivism risk, whereas it is supported by a history of appearance and strong communal ties. The situation can be depicted graphically as follows, where a dashed line indicates opposition and a solid line indicates support:



(1.1.1)

Now, suppose that the judge has previously decided to grant bail to a person who was not at risk of recidivism, and had strong communal ties, but did not have a history of appearing for trial. Then, a fortiori, the judge should also grant bail to a defendant sharing these same characteristics, with the exception that they *do* have a history of appearing for trial. This obligation is called precedential constraint, and it is what the RM is designed to formalize.

The RM generalizes the situation described in Example 1.1. It assumes that a decision is made for one of two sides of a dispute, say a plaintiff and a defendant, based on two sets, Pro and Con, containing factors supporting or opposing a decision for the plaintiff, respectively. A fact situation is described as a pair (X, Y) of subsets $X \subseteq \text{Pro}$ and $Y \subseteq \text{Con}$, indicating which of the factors apply in the fact situation. The RM prescribes a formal principle of a fortiori constraint in this setting: Once a fact situation (X, Y) is decided for the plaintiff, any fact situation (X', Y') with $X \subseteq X' \subseteq \text{Pro}$ and $Y' \subseteq Y \subseteq \text{Con}$ should also be decided for the plaintiff. A dual principle applies to decisions for the defendant.

The RM is elegant in its simplicity—but in many instances it is too simple (Bench-Capon, 2024; Canavotto & Horty, 2023b; Horty, 2011, 2019). In particular, some shortcomings of its factor-based representation of cases have been pointed out, three of which we will now discuss. The first is that not all legally relevant information can be captured as a binary proposition, as it may be multivalued. As a matter of fact, HYPO did allow for the

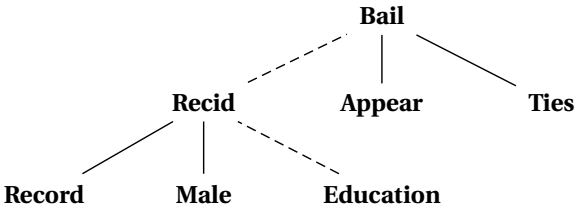
representation of multivalued information in the form of what its authors called *dimensions*, but many of the subsequent works that sprung from HYPO, such as CATO, only used factors. To account for multivalued information, Horty (2019) presented an extended version of the RM which uses a knowledge representation that allows the legally relevant facts to take values in ordered sets, also called dimensions. For the purpose of comparison we will refer to this version as the *dimensional result model* (DRM).

Example 1.2. To apply the RM to the domain discussed in Example 1.1 we should assume that its factors are binary, meaning they either apply or do not apply in any particular situation. However, it is easy to see that in practice it could be beneficial to employ a more fine-grained representation to describe these factors. For example, the factor **Appear**, which corresponds to a history of appearing for trial, could also be described as the relative frequency of past trial appearances, which clearly carries more information. Likewise, we might imagine that that the recidivism risk factor **Recid** is quantified as a number ranging from 1 (lowest) to 10 (highest), rather than as a binary "risk or no risk" judgment.

In the DRM, dimensions are assumed to be ordered, and this order expresses the preference that values of the dimension have towards either of the two possible outcomes of a case: A value in a dimension is "greater than" another if it provides greater support for the plaintiff. In other words, in Horty's model, values do not directly favor an outcome, but instead may be more or less in favor of an outcome compared to other values. Cases in the DRM are modelled as assignments of values to all the dimensions describing the legal domain under consideration, and precedential constraint is defined using the same a fortiori principle that is used for the plain RM: An outcome is forced for a side by a precedent case if the novel fact situation has equal-or-greater support for that side.

A second shortcoming of the RM, as also pointed out by Horty (2011), is that in practice factors often have a hierarchical structure, which the RM does not take into account. A court uses this hierarchical structure to move from low-level factors through a series of intermediate concepts, called abstract factors, before arriving at some final conclusion. This hierarchical structure was utilized by CATO to construct multi-case arguments for a conclusion. The RM, on the other hand, makes a simplifying assumption that the precedential constraint proceeds directly from the base-level factors to a decision for either of the two possible case outcomes, based on the comparison with a single precedent case.

Example 1.3. The factors of Example 1.1, depicted in (1.1.1), can be understood as fitting into a larger factor hierarchy, by recognizing that a recidivism risk assessment itself follows a weighing of pro and con factors. Much research has been done on the factors influencing recidivism—see e.g. Yukhnenko et al. (2020) for a recent meta-study. Examples of such factors include: Does the person have a criminal record? Are they male? Did they obtain a high school diploma? Appending these factors to the graph depicted in (1.1.1) we obtain the following factor hierarchy, culminating in a bail decision:



(1.1.2)

Thirdly, and lastly, we note that the representation of cases used by the RM assumes that every factor either applies or does not apply—but in practice it might be unknown, or irrelevant, whether a factor applies. Examples of this arise naturally in the context of factor hierarchies, where precedential reasoning involves multiple steps, moving through intermediate concepts to an eventual decision. Modeling this process involves representing partial information, wherein the status of higher-level concepts may not (yet) be determined.

Example 1.4. Consider the factor hierarchy depicted in (1.1.2). Initially, a judge deciding whether to grant bail to a defendant would be presented with information on some, or all, of the lowest level factors, such as whether the person has finished high school. A representation of this initial situation should not yet contain information on whether the person presents a recidivism risk. Based on the initial information, the judge may decide to assess the recidivism risk, and then come to a bail decision. Alternatively, the judge might decide, for example, that a lack of communal ties and a history of nonappearance already provide sufficient reason to deny bail, and forego a recidivism risk assessment. The knowledge representation framework should be able to describe these various scenarios.

The RM concisely formulates a principle of a fortiori constraint for cases using a factor-based representation. However, this representation makes simplifying assumptions that do not always apply in practice. This leads us to the first of our research questions:

Research question 1: Can we extend the result model of precedential constraint to a general theory of a fortiori case-based reasoning? In particular:

- A How can incomplete, dimensional, and hierarchical information be incorporated in the knowledge representation, and what should the corresponding notion of constraint be?
- B How can the models developed in response to 1A be formally compared, and what are their differences and similarities?
- C What is the relation between this theory and other reasoning formalisms, such as logic?

Part I of this thesis will be devoted to answering Research question 1 and its subquestions. Question 1A is addressed in the first three chapters of this part. In Chapter 2 we review the RM in detail, laying the groundwork for subsequent chapters, and we address Question 1A regarding the representation of incomplete information. In Chapter 3 we discuss the dimension-based extension of the RM, the DRM, and position it in a broader context using order theory and many-sorted logic, to answer Question 1C. Then, in Chapter 4, we propose an extension of the RM that incorporates factors with hierarchical structure. In Chapter 5 we introduce the notion of *dimension hierarchy*—simply put, a set of hierarchically structured dimensions—and propose an extension of the HRM which operates on a dimension hierarchy instead of a factor hierarchy. We call the resulting model the DHRM. After each model is introduced, it is formally compared to the other ones, in order to answer Question 1B.

The RM and the DRM have also met with criticism with regards to their applicability to modeling the common law doctrine of precedent (see e.g. Horty, 2011; Rigoni, 2018). However, the RM and its extensions can still be usefully applied more generally as a model of a fortiori reasoning—particularly in the context of AI, as we will see in the next section.

1.2 Applications in artificial intelligence and law

Much present-day research is focused on increasing the interpretability of AI systems, i.e., to enable humans to understand why a complex AI system behaves in the way it does. This research is partially done in response to mounting concerns that uninterpretable algorithms, so-called *black box* AI, are making high-impact decisions—such as those with legal, social, or ethical consequences—in an unfair or irresponsible manner. A prominent example of such a system is the proprietary software Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), developed by Northepointe (now called Equivant) for automatic risk assessment of various forms of recidivism, which has seen nationwide use in the United States (Equivant, 2019). High-profile allegations by Angwin et al. (2016) that COMPAS racially discriminates in its decision-making process have led to a host of follow-up research and discussions. Since then, significant problems in the analysis by Angwin et al. (2016) have been identified (Barenstein, 2019; Dieterich et al., 2016; Flores et al., 2016)—but as Rudin et al. (2020a) point out, this situation is symptomatic of the larger problem that the use of such black box systems is obstructing independent assessment of bias, regardless of the veracity of the allegations in this particular instance.

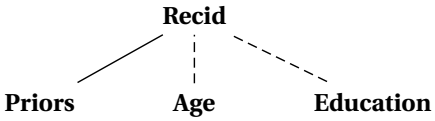
Many different kinds of solutions have been proposed, among which those to make AI inherently more transparent (Rudin, 2019), to formulate appropriate regulations (Wachter et al., 2017), to monitor the systems and measure bias over time (Kurita et al., 2019), and to develop *post-hoc* explanation methods, in which the black-box system is analyzed after it has been trained and little to no access to the way it functions is assumed. There are in turn many types of post hoc explanation methods, see e.g. the work by Koh and Liang (2017), Ribeiro et al. (2016), and Wachter et al. (2018).

Another approach is to consider the problems and solutions studied in the field of AI & law, of which explainability has always been a core aspect (Atkinson et al., 2020; Verheij, 2020). Indeed, we have seen several applications of models of case-based reasoning from the AI & law literature to the improvement of interpretability, and value alignment of AI systems (Canavotto & Horty, 2022; Čyras et al., 2016, 2019; Liu et al., 2022; Prakken & Ratsma, 2022). The idea of a CBR explanation of a decision is to provide an analogy between the decision and relevant training examples. Proponents of the CBR-explanation approach, such as Nugent and Cunningham (2005), argue that explanations of this form are natural to humans: they are simple, people are well-acquainted with reasoning by analogy, and these explanations draw on real evidence in the sense that training examples typically serve as a gold standard that a machine learning system adheres to.

The dimension-based result model (DRM), which we discussed in the previous section, was conceived primarily as a formal model of legal CBR, but since its inception it has found applications to research on interpretable AI, based on the analogy of *training examples as cases*: Just as a court draws on precedent cases to decide on a novel fact situation, machine learning systems draw on training data to decide on novel data points. Based on this analogy, Prakken and Ratsma (2022) combined the DRM with HYPO style argumentation to develop an a fortiori case-based argumentation (AF-CBA) method for justifying the decisions of data-driven machine learning systems. The AF-CBA method conceptually tries to mimic the arguments used by lawyers with respect to case law. In such discussions, precedent cases are cited by both sides as a means of arguing that the present (focus) case should be decided similarly as the precedent. Both sides may attack the other's citations, by pointing

to important differences between the citation and the focus case; and they may defend themselves against such attacks, by pointing to aspects of the focus case which compensates for these differences.

Example 1.5. We consider, as an example, the aforementioned COMPAS program, which is designed to predict recidivism, and has been at the center of debates on the fairness of decisions made by black-box AI systems. Suppose, for the sake of simplicity, that it makes decisions based on three input features of a person:



(1.2.1)

Here, **Priors** represents a number of prior convictions, **Age** the age of the person, and **Education** a binary feature indicating whether the person finished high school. The number of priors is positively correlated with a high-risk assessment, as indicated by the solid line above; while the person's age and level of education is negatively correlated, as indicated by the dashed line.

The COMPAS program was trained to predict recidivism based on a dataset of such features together with labels indicating whether a recidivism offense took place (Brennan et al., 2009). Given a particular input-output pair, called the *focus case*, generated by COMPAS, the AF-CBA method can generate an explanation of the focus case in the form of an argumentative dialogue. This dialogue starts with the citation of a most-similar precedent in the dataset used to train the model—the *precedent case*. The AF-CBA method then proceeds by comparing the focus case to this precedent. If the precedent and focus case share the same outcome, the explanation emphasizes their similarities and addresses any differences by arguing that these do not significantly affect the decision. For instance, if the focus case involves a younger individual but still results in a high-risk prediction, the explanation might argue that the higher number of prior convictions compensates for the mitigating effect of age. Conversely, if the precedent has an opposite outcome, the method identifies relevant differences and uses those to justify the opposing outcomes.

The method by Prakken and Ratsma (2022) uses the DRM to interpret a machine learning dataset as a set of precedent cases, on the basis of which to generate explanatory dialogues. In other words, the data used to train the AI is taken as the set of precedent cases. Another option is to take a set of decisions made by the AI, and use that as the set of precedent to compare decisions to. In that setting, the notion of precedential constraint formulated by the DRM can be used as a measure of the internal consistency of the decisions made by the AI system.

Example 1.6. Consider again the set of features depicted in (1.2.1) of Example 1.5. Suppose that a 20 year old with 2 prior offenses and no high-school diploma was deemed to be at a high risk of recidivism by COMPAS. According to the a fortiori constraint principle of the DRM, this sets a precedent which constrains future decision-making: Anyone who is at or below the age of 20, has 2 or more prior offenses, and has not completed high-school, should be deemed at high risk of recidivism. Any decision made by COMPAS which violates this constraint is considered inconsistent by the model, and given a set of decisions by

COMPAS the relative frequency of inconsistent decisions within it can be used as a concrete measure of its decision fairness.

In a similar vein, DRM has also been implemented as part of a human-in-the-loop decision support system for classification of fraudulent web shops at the Dutch National Police Lab AI (Odekerken & Bex, 2020; Odekerken et al., 2023b). To facilitate this and similar applications, Odekerken et al. (2023b) have further developed the theory of the DRM by adding formal notions of *justification*, *stability*, and *relevance*. These notions allow the model to be applied to fact situations in which the values of the dimensions are not known precisely, but are only known to lie within a certain subset.

In sum, models of CBR that stem from the literature on AI & law are increasingly being applied in the context of artificial intelligence, to analyze, justify, or even make data-driven decisions. These applications lead us to the second set of research questions:

Research question 2: How can the models of a fortiori precedential constraint be applied to artificial intelligence? In particular:

- A How can the theory of a fortiori precedential constraint be used to formalize compensation and citability, to aid in justifying data-driven decisions?
- B How can we write capable and efficient computer implementations of these models?
- C Is precedential constraint useful as a measure of data-driven decision consistency?

Part II of this thesis will be devoted to answering Research question 2 and its subquestions. To start, we consider the application of the DRM to post-hoc explanation in the AF-CBA style of Example 1.5. We review the method in detail, and propose extending it with formal notions of compensation and citability, to answer Question 2A. In Chapter 7 we turn to Question 2B. Specifically, we draw on the logical theory which will be developed in Chapter 3 to implement the DRM in the satisfiability modulo theories (SMT) solver Z3 (de Moura & Bjørner, 2008). This implementation is then put to work to analyze the consistency of several machine learning datasets. Lastly, to answer Question 2C in Chapter 8, we use Z3 to implement an extended version of the DRM which will be developed in Chapter 5, and use this implementation to analyze a dataset of recidivism risk assessments made by the COMPAS program, and measure the internal consistency of its decisions.

1.3 Thesis outline

This thesis is divided into two parts and contains a total of 9 chapters, including this one.

Part I: A Theory of a Fortiori Case-Based Reasoning In this part we will answer research question 1, regarding the development of a general theory of a fortiori reasoning. In Chapter 2 we review the RM, upon which the rest of the chapters in this part will be build. In Chapter 3 we review the DRM, and position it within the broader literature on

order theory and many-sorted logic. In Chapter 4 we present an extension of the RM which operates on a factor hierarchy, as opposed to a flat set of factors, as a first step towards answering research question 1. Lastly, in Chapter 5, we will present an extension of the HRM, called the dimensional hierarchical result model (DHRM), which operates on a dimension hierarchy, as the answer to research question 1.

Part II: Applications in Artificial Intelligence and Law In this part we will answer research question 2, regarding the implementation and application of models of a fortiori precedential constraint to AI and law. In Chapter 6 we will review the AF-CBA method for post-hoc explanations and propose extensions to it. In Chapter 7 we will implement the DRM using the Z3 SMT solver and analyze the consistency of several machine learning datasets. In Chapter 8 we will implement the DHRM and use it to analyze a dataset of recidivism risk assessments made by the COMPAS program, measuring the internal consistency of its decisions.

Conclusion Lastly, in Chapter 9, we conclude the thesis by summarizing and discussing our answers to the research questions, and outlining directions for future work.

Part I

A Theory of a Fortiori Case-Based Reasoning

> Es wird . . . im Sinne der Bestrebungen unserer Gesellschaft liegen, wenn in ihren Tagungen die alte Wahrheit . . . Würdigung findet, daß . . . dem Anwenden das Erkennen vorausgehen muß.
>
> —Max Planck, *Das Wesen des Lichts*, 1919

FORTIORI REASONING was described by Horty through his result model (RM) of precedential constraint. Part I of this thesis is devoted to extending the RM, in order to progressively develop increasingly expressive models of a fortiori case-based reasoning. We begin by extending the RM, which uses a plain factor-based representation of cases in the style of HYPO, to make use of a hierarchical factor-based representation in the style of CATO; we call the resulting model the *hierarchical result model* (HRM). Then, just as Horty extended his RM to the DRM in order to account for dimensional information, we extend the HRM to account for dimensional information; we call the resulting model the *dimensional hierarchical result model* (DHRM). We motivate and illustrate the applicability of these various a fortiori models by use of a running example from the legal domain of criminal sentencing. More specifically, we consider the tasks of judging recidivism risk and granting bail. We show that a fortiori reasoning is applicable to these tasks, and requires the use of a knowledge representation incorporating both dimensional and hierarchical information. Criminal sentencing is a highly relevant domain for our purposes—decisions surrounding criminal sentences have the potential to greatly affect peoples' lives, and AI is increasingly being used to complement or even replace human decision-making for these tasks.

Chapter 2

Modeling Factor-Based Constraint

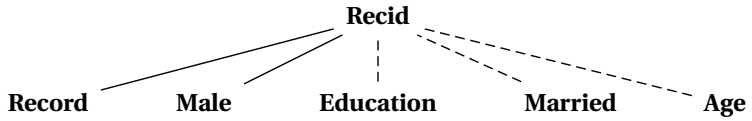
We start part I with a chapter discussing Horty's (2004, 2011) result model of precedential constraint.[1] In Section 2.1 we explain the concept of factors through the use of a running example based on the legal domain of criminal sentencing. We then formalize the knowledge representation used by Horty's model in Section 2.2, and subsequently give his notion of constraint in Section 2.3. In Section 2.4, we generalize Horty's theory by allowing fact situations to assign a truth value to only a subset of the factors of the given domain, and adapt the notion of constraint to this setting. We refer to this as the "result model" (RM), and it will serve as the basis for the models we present in the subsequent sections. In Section 2.5 we consider the notion of case base consistency, which states that the cases in the case base adhere to the constraint they induce. We then consider in Section 2.6 whether the type of reasoning represented by the RM is monotonic or not. We consider two types of monotonicity; that in the addition of new cases to the case base, and that in the addition of new information to the focus fact situation. We find that the RM is monotonic with respect to both of these types of additions. As the last of our considerations on this topic, we compare Horty's (2011) result model to his reason model in Section 2.7. We hope that this bridge between these two models might allow our findings regarding the result model to find applications to the reason model, which has continued to receive attention in the literature (Bench-Capon, 2024; Canavotto & Horty, 2023a, 2023b). We end the chapter in Section 2.8 with some concluding remarks.

2.1 An example of factors

We illustrate the various kinds of models discussed in this work through a running example in the criminal sentencing domain. In this case, we consider a judgment of whether a

[1] The material in this chapter—with the exception of Sections 2.5 and 2.7, which are new—stems from van Woerkom et al. (2025).

convict is at low or high risk of recidivism. Much research has been done on the factors influencing recidivism; see e.g. the work by Yukhnenko et al. (2020) for a recent meta-study. Below is a graphical representation of a number of such factors:



(2.1.1)

The factors in the bottom row respectively indicate whether the defendant has a criminal record (**Record**), is male (**Male**), has a high-school diploma (**Education**), is married (**Married**), and is over the age of 21 (**Age**). A solid line between a factor and the **Recid** node indicates that the presence of that factor suggests a higher risk of recidivism, while a dotted line indicates that its presence suggests a lower risk. For instance, having a criminal record indicates a higher risk, while being married indicates a lower risk. Now, suppose a 30 year-old unmarried male defendant with a pre-existing criminal record and no high-school diploma was judged to be at high risk of recidivism. Given our assumption that older people tend to recidivate less, it follows *a fortiori* that a defendant who is on all accounts similar, but is 20 years old instead of 30, should also be judged to be at high risk of recidivism.

2.2 Knowledge representation

A *factor* is a propositional variable, i.e. a variable which is either true (denoted **t**) or false (denoted **f**). We denote factors using lowercase letters p, q, r, etc. The domain is modeled by a *factor partition* $F = \text{Pro} \cup \text{Con}$, where Pro and Con are sets of factors satisfying $\text{Pro} \cap \text{Con} = \emptyset$. A *fact situation* is a valuation of F, i.e. a function $X : F \to \{\mathbf{t}, \mathbf{f}\}$ assigning true or false to every factor in F. We use upper case letters X, Y, Z etc. to denote fact situations, and write $X \vDash p$ for $X(p) = \mathbf{t}$ and $X \vDash \neg p$ for $X(p) = \mathbf{f}$. We may combine statements of this form with set notation, so for instance by $X \vDash \{p, \neg q, r\}$ we mean $X \vDash p$, $X \vDash \neg q$, and $X \vDash r$.

Cases are decided for either of two *sides*: the plaintiff, denoted by π, or the defendant, denoted by δ. Each factor $p \in F$ has a *preference* for exactly one of the two sides, which is modeled by two sets Pro and Con, which represent the factors supporting or opposing a decision for the plaintiff, respectively. If a factor is pro-π (δ) we assume it is con-δ (π), and so we write $\text{Pro}(\pi) = \text{Con}(\delta) = \text{Pro}$ and $\text{Con}(\pi) = \text{Pro}(\delta) = \text{Con}$. It is often useful to just speak of a generic side $s \in \{\pi, \delta\}$, in which case we denote the "other" side by $\bar{s}$; so $\bar{\pi} = \delta$ and $\bar{\delta} = \pi$. A *case* is a pair (X, s) with X a fact situation and s a side; a *case base* $\mathscr{C}$ is a finite set of cases. The notation and terminology of plaintiff, π, and defendant, δ, is standard in the literature. However, throughout this work we will also use the RM (and its extensions that we discuss in Chapters 3,4, and 5) to model scenarios that do not necessarily involve a plaintiff and defendant. For instance, in our running example of a recidivism risk assessment (2.1.1) the decision represents a judgment by a court of whether a person is at high or low risk of recidivism, and not a decision for a plaintiff or a defendant. For this reason we may also use the neutral denotations 0 and 1 to indicate the outcomes δ and π, respectively. We note that these are also the labels often used in the setting of binary

classification—a task commonly encountered in machine learning.

Our method of representing cases differs slightly in form compared to that used by Horty (2011), who models a fact situation X as a subset $X \subseteq F$ of the factors. However, formally speaking, there is no difference between these two approaches—it is a well-known result in set theory that the powerset $\mathscr{P}(F)$ of a set F is in bijection to the set $\{\mathbf{t}, \mathbf{f}\}^F$ of all functions with signature $F \to \{\mathbf{t}, \mathbf{f}\}$. In other words, whether we define a case in terms of a function $X : F \to \{\mathbf{t}, \mathbf{f}\}$ or a subset $X \subseteq F$ does not formally make any difference. Of course, semantically we could argue that "assigning false" ($X(p) = \mathbf{f}$) to a factor carries a different meaning than when it "is absent" from a fact situation ($p \notin X$). We will return to this discussion about undefined factors in Section 2.4 below.

2.3 Constraint

We can now define how a decision on a new case is constrained based on a given case base of past decisions. For this, we define the notion of forcing a decision in fact situation given a case base.

The idea behind the RM is that a decision of a fact situation X for a side s constitutes a balancing of the pro-s factors in X against the con-s factors in X. The support that factors provide for an outcome is defeasible and unquantified, which makes it difficult to weigh sets of pros against sets of cons. However, once a set of pros was deemed to outweigh a set of cons, any superset of the set of pros should also outweigh any subset of the set of cons. This intuition is formalized by the following definition.

Definition 2.1. Let X be a fact situation and $\mathscr{C}$ a case base, then *the decision of X for π is forced by $\mathscr{C}$*, denoted $\mathscr{C}, X \vDash \pi$, if and only if there is a case $(Y, \pi) \in \mathscr{C}$ such that:

- for all $p \in \text{Pro}$: if $Y \vDash p$ then $X \vDash p$, and
- for all $p \in \text{Con}$: if $X \vDash p$ then $Y \vDash p$.

Likewise, *the decision of X for δ is forced by $\mathscr{C}$*, denoted $\mathscr{C}, X \vDash \delta$, iff there is a case $(Y, \delta) \in \mathscr{C}$ such that:

- for all $p \in \text{Con}$: if $Y \vDash p$ then $X \vDash p$, and
- for all $p \in \text{Pro}$: if $X \vDash p$ then $Y \vDash p$.

Example 2.2. We demonstrate Definition 2.1 through an example based on the factors in (2.1.1). Let $(Y, 1)$ be a case with $Y \vDash \{\textbf{Record}, \neg\textbf{Male}, \neg\textbf{Education}, \textbf{Married}, \neg\textbf{Age}\}$, and where 1 represents a judgment of high recidivism risk. In other words, the case $(Y, 1)$ represents a decision that the person described as fact situation Y was deemed to pose a high recidivism risk by some decision-maker. When does this precedent constrain the risk assessment of a new fact situation X? Unfolding the definition we find:

$\{(Y, 1)\}, X \vDash 1$

iff • for all $p \in \{\textbf{Record}, \textbf{Male}\}$: if $Y \vDash p$ then $X \vDash p$, and

  • for all $p \in \{\textbf{Education}, \textbf{Married}, \textbf{Age}\}$: if $X \vDash p$ then $Y \vDash p$

iff $X \vDash \{\textbf{Record}, \neg\textbf{Education}, \neg\textbf{Age}\}$.

Indeed, when X matches the truth status of all the factors that were indicative of a recidivism risk in Y, then X should also be judged to be at a high risk. The truth status in X of the other factors is relevant—for example, with respect to the comparison with Y it does not matter whether X is married or not, because if X is not married then this will only further strengthen the case for a "high risk" assessment. However, if X fails to match the status of, say, **Record**, then it may be argued that X has less support for a "high risk assessment"—compared to Y—and thus the precedent case $(Y, 1)$ should not constrain X.

Note that the RM could in principle be used as a classifier which assigns outcomes to a new fact situation X: assign outcome 1 if $\mathscr{C}, X \models 1$, and assign outcome 0 if $\mathscr{C}, X \models 0$ (we will expand on this idea in Section 3.10.3). However, the intended purpose of the RM is not to weigh pros and cons against each other, but rather to normatively prescribe an a fortiori principle to some such weighing method. For example, given the factors depicted in (2.1.1), the intended purpose of the RM is not to classify a given defendant as low or high risk. Instead, it prescribes to a decision-maker, such as a court, what it means to make recidivism risk assignments in accordance with the precedent and the a fortiori principle. It is on this basis that the RM can be used in the context of AI: by interpreting features as factors, and training data as precedent cases, the decisions of an AI system can be compared to the constraint induced by the RM. Such a comparison will be the primary focus of Chapter 7 in Part II of this work.

2.4 Partial fact situations

The knowledge representation used by the RM requires that for any particular fact situation we specify, for each factor, whether it applies in the situation or not. This is not always an appropriate assumption, because in practice it may be unknown what the truth value of a factor is. It may be argued that, for the subset representation, the meaning of the absence of a factor from the subset is that it is unknown whether it applies or not. However, this interpretation has the same problem of not being sufficiently fine-grained—if, for example, the **Married** factor of (2.1.1) is absent from a subset, does this represent that the defendant is unmarried, or that it is not known whether the defendant is married?

The need for a more fine-grained representation will become all the more pressing once we move from a plain set of factors to a factor hierarchy in Chapter 4, and so we will consider a modification to the knowledge representation of the RM that accounts for unassigned factors. We do so by considering *partial* fact situations, in the sense of partial functions on sets. A partial function of a set A to a set B is a function $f : C \to B$ for some subset $C \subseteq A$. Given a partial function on a set A we write $\mathrm{dom}(f) \subseteq A$ for the subset of A on which f is defined. Applying this definition to valuations yields the definition of a partial fact situation X: a valuation that is only defined on some subset $G \subseteq F$ of factors. Note that, according to this definition, a situation X which is defined on all factors (so $\mathrm{dom}(X) = F$) is also considered a partial fact situation. To differentiate these from strictly partial fact situations we will call such fact situations *complete*. We maintain the same notation regarding applicability of factors in partial fact situations: we write $X \models p$ when $X(p) = \mathbf{t}$, and $X \models \neg p$ when $X(p) = \mathbf{f}$. In addition, we write $X \models ?p$ to denote the expression $p \notin \mathrm{dom}(X)$, so $X \models ?p$ holds if and only if X is undefined on p. Note that for partial fact situations the equivalence $X \models \neg p \leftrightarrow X \not\models p$ no longer holds. In turn, we can

now consider cases (X, s) where X is a partial fact situation, which represents a decision made by the court on the basis of incomplete information. We will say a case is partial, or complete, if its underlying fact situation is.

The use of partial fact situations and cases raises the question if, and if so how, Definition 2.1 of constraint should be altered. Suppose Y is a partial fact situation which was decided for side s, so (Y, s) is a partial case, when does the decision (Y, s) induce constraint on a partial focus fact situation X? Our philosophy when it comes to constraint induced by a partial case can be summarized as follows: if the status of a factor in the precedent case added strength to the side for which the case was decided, then the status of that factor should be matched in the focus fact situation in order for constraint to apply. We note three consequences of this view in particular:

- Factors on which the precedent case is undefined are inconsequential to the notion of constraint. We interpret a decision of Y for s as meaning that the factors in $\mathrm{dom}(Y)$ were sufficient to arrive at a decision for s, and so only the factors in $\mathrm{dom}(Y)$ should have an influence on the constraint induced by (Y, s).
- If p is a pro-s factor such that $Y \models p$, then p added strength to the side s and so it should also apply in the focus fact situation in order for it to be said to have equal-or-greater support for side s. In particular, if X is undecided on p then (Y, s) should not induce constraint on X.
- Similarly, if p is a con-s factor such that $Y \models \neg p$, then the status of p added support for s, and so it should be matched by X. In particular, if X is undecided on p then no constraint should be induced.

In principle, Definition 2.1 can be interpreted without problems when the focus fact situation X and the precedent case (Y, s) it involves are partial. However, in light of the aforementioned considerations, the resulting notion of constraint does not align with our intuitive notion of how constraint should work. The misalignment arises when we consider a con-s factor p, as illustrated in Table 2.1. The condition in Definition 2.1 for such a factor states that, in order for (Y, s) to induce constraint, the implication "if $X \models p$ then $Y \models p$" should hold. In particular, when $X \models p$ and $Y \models ?p$ then this implication does not hold, whereas intuitively we do not want to exclude constraint in such a scenario. Similarly, when $X \models ?p$ and $Y \models \neg p$ then constraint may still possibly be induced, while intuitively we want to exclude it. The rest of the possible scenarios do align with our intuition.

The solution we thus propose is to instead use the implication "if $Y \models \neg p$ then $X \models \neg p$" for the con factors. As shown in Table 2.1 this solution does align with our intuition. In sum, this results in the following definition of constraint for partial fact situations.[2]

Definition 2.3. Let X be a partial fact situation and $\mathscr{C}$ a case base of partial cases, then *the decision of X for π is forced by $\mathscr{C}$*, denoted $\mathscr{C}, X \models \pi$, if and only if there is a case $(Y, \pi) \in \mathscr{C}$ such that:

- for all $p \in \mathrm{Pro}$: if $Y \models p$ then $X \models p$, and
- for all $p \in \mathrm{Con}$: if $Y \models \neg p$ then $X \models \neg p$.

[2]This is an updated version of the definition used by van Woerkom et al. (2023a, 2023b), which contains the implication "if $X \models p$ then $Y \models p$" for the con factors of the precedent case.

Table 2.1: An overview of the differences for two possible prerequisites for a partial precedent case (Y, s) to induce constraint on a partial fact situation X. Here p is an arbitrary con-s factor, and every row represents a possible value assignment by the fact situations X and Y.

$Y(p)$	$X(p)$	if $X \vDash p$ then $Y \vDash p$	if $Y \vDash \neg p$ then $X \vDash \neg p$
t	**t**	✓	✓
t	**f**	✓	✓
t	?	✓	✓
f	**t**	×	×
f	**f**	✓	✓
f	?	✓	×
?	**t**	×	✓
?	**f**	✓	✓
?	?	✓	✓

Likewise, *the decision of X for δ is forced by $\mathscr{C}$*, denoted $\mathscr{C}, X \vDash \delta$ iff there is a case $(Y, \delta) \in \mathscr{C}$ such that:

- for all $p \in$ Pro: if $Y \vDash \neg p$ then $X \vDash \neg p$, and
- for all $p \in$ Con: if $Y \vDash p$ then $X \vDash p$.

Note that for complete fact situations the implication for con factors is just the contrapositive of the one used in Definition 2.1; e.g., for the decision of X for π we have:

"if $X \vDash p$ then $Y \vDash p$" iff "if $Y \nvDash p$ then $X \nvDash p$" iff "if $Y \vDash \neg p$ then $X \vDash \neg p$".

This means that for complete fact situations Definitions 2.1 and 2.3 coincide, which allows us to use the same notation for both definitions.

Remark 2.4. An alternative version of Definition 2.3 is obtained by using the implication "if $X \vDash p$ then $Y \vDash p$" for the con-s factors. For the sake of comparison we will refer to this version as option (2) of the definition, and the current version—using the implication "if $Y \vDash \neg p$ then $X \vDash \neg p$"—as option (1).

Example 2.5. We reconsider Example 2.2 with respect to Definition 2.3. Suppose it is not known whether Y is married or not, so $Y \vDash$ **?Married**, but that Y was nevertheless deemed to be a high recidivism risk, so we have a precedent case $(Y, 1)$. Should this case be able to induce constraint on a new fact situation X satisfying $X \vDash$ **Married**? According to Definition 2.3 the answer is yes—since $Y \nvDash \neg$**Married**, there is no further requirement placed on X with regards to the **Married** factor. In this regard, the situation is identical to the scenario in which Y *was* in fact married, i.e. it is as though we assume that $Y \vDash$ **Married**. Similarly, suppose that the age of X is unknown, whilst it is known that Y is younger than 21; so $X \vDash$ **?Age** and $Y \vDash \neg$**Age**. Again we ask: should $(Y, 1)$ be able to constrain a decision for X? This time the answer is no—since $Y \vDash \neg$**Age** is a reason to decide Y for 1, the fact that $X \vDash$ **?Age** means that X does not match this support for a high-risk judgment and so constraint is excluded. In this regard, the situation is identical to the scenario in which X

was not over the age of 21, i.e. it is as though we assume that $X \vDash \textbf{Age}$. These observations with regards to assumptions about the truth values of the undefined factors in the precedent and focus fact situations are part of a general pattern, to which we will return in Section 2.6.

2.5 Consistency

We now briefly turn to an important notion in the work of Horty (2011, 2019), which is that of consistency: decisions are consistent with a case base if and only if they are in agreement with the constraint it induces. First we give Horty's notion of consistency for the RM.

Definition 2.6. A case base $\mathscr{C}$ is *strongly inconsistent* if it contains a case (X, π) such that $\mathscr{C}, X \vDash \delta$, or if it contains a case (Y, δ) such that $\mathscr{C}, Y \vDash \pi$; otherwise it is *weakly consistent*.

Example 2.7. We illustrate this notion by reconsidering Example 2.2. Suppose Z is a fact situation with $Z \vDash \{\textbf{Record}, \textbf{Male}, \neg\textbf{Education}, \neg\textbf{Married}, \neg\textbf{Age}\}$. This means Z is in all aspects indicative of high recidivism risk. If, in spite of this, Z was deemed to be at a low risk of recidivism, we get a case base $\{(Y,1),(Z,0)\}$. We can then compute that $\{(Y,1),(Z,0)\}, Z \vDash 1$ because $Z \vDash \{\textbf{Record}, \neg\textbf{Education}, \neg\textbf{Age}\}$. This means $\{(Y,1),(Z,0)\}$ is strongly inconsistent in the sense of Definition 2.6: Z was deemed to be at a low risk, even though Y was deemed to be at a high risk and the factors in Z make a stronger case for a high risk assessment.

We have added the qualifiers "strong" and "weak" for the sake of comparison with a second possible definition of consistency, which is defined as follows.

Definition 2.8. A case base $\mathscr{C}$ is *weakly inconsistent* if there exists a fact situation X such that $\mathscr{C}, X \vDash \pi$ and $\mathscr{C}, X \vDash \delta$; otherwise, it is *strongly consistent*.

Note the difference between Definitions 2.6 and 2.8: the former quantifies only over cases within the case base $\mathscr{C}$, while the latter quantifies over all possible fact situations. The use of the strength qualifiers in these definitions is justified by the following proposition.

Proposition 2.9. *Strong inconsistency implies weak inconsistency.*

Proof. Suppose $(X, \pi) \in \mathscr{C}$ and $\mathscr{C}, X \vDash \delta$. Any case forces its own outcome so $\mathscr{C}, X \vDash \pi$, which means we have found a fact situation X satisfying the desired requirements. □

The contrapositive of Proposition 2.9 states that strong consistency implies weak consistency.

So far, no distinction has been made in the literature between Definitions 2.6 and 2.8. Presumably, this is because for complete fact situations they coincide—as we now show.

Proposition 2.10. *Let $\mathscr{C}$ be a case base consisting of complete cases, then $\mathscr{C}$ is strongly inconsistent iff it is weakly inconsistent.*

Proof. By Proposition 2.9 it remains to show the direction from right to left. If $\mathscr{C}$ is weakly inconsistent then this means there is a fact situation X such that $\mathscr{C}, X \vDash \pi$ and $\mathscr{C}, X \vDash \delta$. So, there exists a case $(Y, \pi) \in \mathscr{C}$ forcing the decision of X for π, and a case $(Z, \delta) \in \mathscr{C}$ forcing it for δ. We now show that $\mathscr{C}, Y \vDash \delta$, so that $\mathscr{C}$ is indeed strongly inconsistent. We do this by proving that

- for all $p \in \text{Pro}$: $Z \models \neg p$ implies $Y \models \neg p$, and
- for all $p \in \text{Con}$: $Z \models p$ implies $Y \models p$.

Let $p \in \text{Pro}$ such that $Z \models \neg p$. Assume for the sake of contradiction that $Y \not\models \neg p$. Then, as Y is complete, we have $Y \models p$. Since (Y, π) forces the decision of X for π this means that $X \models p$. However, since (Z, δ) forces it for δ, we have that $Z \models \neg p$ implies $X \models \neg p$, which contradicts $X \models p$, and so $Y \models \neg p$. The second implication can be proven using the same reasoning, so that indeed $\mathscr{C}, Y \models \delta$, meaning $\mathscr{C}$ is strongly inconsistent. □

The proof of Proposition 2.10 relies essentially on the assumption that $\mathscr{C}$ only contains complete fact situations; indeed, case bases that contain partial cases can be both weakly inconsistent and not strongly inconsistent.

Proposition 2.11. *Weak consistency does not imply strong consistency.*

Proof. We exhibit a case base that is both weakly consistent and weakly inconsistent. Consider a factor partition F given by $\text{Pro} = \{p, q\}$ and $\text{Con} = \{r, s\}$. Let X, Y, Z be fact situations for F such that $X \models \{p, q, r, s\}$, $Y \models \{p, ?q, r, ?s\}$, and $Z \models \{?p, q, ?r, s\}$. Now, we claim the case base $\mathscr{C} = \{(Y, \pi), (Z, \delta)\}$ satisfies the desired properties. Firstly, it is weakly consistent because $\mathscr{C}, Y \not\models \delta$ and $\mathscr{C}, Z \not\models \pi$. For example, the only way $\mathscr{C}, Y \models \delta$ could hold is if (Z, δ) forces the decision of Y for δ, in order for which we should have $Y \models s$; but Y is undecided on s. However, $\mathscr{C}$ is also weakly inconsistent, as $X \models \{p, s\}$ entails both $\mathscr{C}, X \models \pi$ and $\mathscr{C}, X \models \delta$. □

The counterexample in Proposition 2.11 may seem synthetic. In a real setting, this would have the form that a court would weigh the pro and con factors p and r in favor of the plaintiff (in absence of q and s) while weighing q and s for the defendant (in absence of p and r). This need not introduce any inconsistency with earlier decisions, but it introduces an "inconsistency waiting to happen." It would be interesting to see a real-life example of a case base that is both weakly consistent and weakly inconsistent. However, the data analyses we will perform in Part II will only use complete case bases and so the difference between the two notions of consistency discussed in this section will not play a role there.

2.6 Monotonicity

Since the notion of constraint can be thought of as a type of entailment relation we can consider whether it is *monotonic*: can deductions be invalidated by the addition of information? This is a property which characterizes the defeasibility of inferences, and plays an important role in the study of reasoning and argumentation (Strasser & Antonelli, 2019). Classical logic is well-known to be monotonic, which is formally captured by the implication that if $\Phi \vdash \phi$ and $\Phi \subseteq \Phi'$ then $\Phi' \vdash \phi$, for any formula ϕ and sets of formulas Φ, Φ'. This property runs contrary to everyday reasoning, and so the monotonicity of classical logic has spurred much research on nonmonotonic logic and defeasible reasoning in general (Koons, 2017; Strasser & Antonelli, 2019).

One relevant adaptation of the statement of monotonicity for the RM is the question whether the forcing of a decision of a fact situation for a particular side can be invalidated by an expansion of the case base. It is not hard to see that the RM is indeed monotonic in this sense, as the following proposition shows.

Proposition 2.12. *If $\mathscr{C}, X \models s$ and $\mathscr{C} \subseteq \mathscr{D}$ then $\mathscr{D}, X \models s$.*

Proof. To show $\mathscr{D}, X \models s$ we need $(Y, s) \in \mathscr{D}$ satisfying the appropriate conditions of Definition 2.3. Since $\mathscr{C}, X \models s$ there is such a $(Y, s) \in \mathscr{C}$, and therefore by $\mathscr{C} \subseteq \mathscr{D}$ we have $(Y, s) \in \mathscr{D}$ as desired. □

Example 2.13. We have seen an example of Proposition 2.12 in Example 2.7: $\{(Y, 1)\}, Z \models 1$ because $Z \models \{\textbf{Record}, \neg\textbf{Education}, \neg\textbf{Age}\}$, and the addition of $(Z, 0)$ to the case base $\{(Y, 1)\}$ does nothing to interfere with this derivation, so we also have $\{(Y, 1), (Z, 0)\}, Z \models 1$.

However, the use of partial fact situations introduces a second form of monotonicity for consideration: can deductions be invalidated by adding information to the focus fact situation? We define the addition of information in this case using the notion of function extension.

Definition 2.14. Given a function $f : A \to B$ and a subset $C \subseteq A$ we define the *restriction* $f\restriction_C : C \to B$ of f to C by $f\restriction_C(x) = f(x)$ for $x \in C$.

Definition 2.15. A function f is an *extension* of a function g, denoted by $g \subseteq f$, if $\mathrm{dom}(g) \subseteq \mathrm{dom}(f)$ and $f\restriction_{\mathrm{dom}(g)} = g$.

Since we consider fact situations to be just functions, these concepts are directly applicable to fact situations. In this context, a fact situation Y extends a fact situation X if they agree on all factors in $\mathrm{dom}(X)$, but where Y is possibly defined on more factors than X. So, we can regard an extension Y of a fact situation X as containing more information than X, and ask whether the notion of constraint is monotonic with respect to the addition of this information. Again, this is indeed the case.

Proposition 2.16. *Let X and Y be fact situations such that $X \subseteq Y$, then $\mathscr{C}, X \models s$ implies $\mathscr{C}, Y \models s$ for $s \in \{\pi, \delta\}$.*

Proof. We consider the case where $s = \pi$; the other case is similar. Suppose $\mathscr{C}, X \models \pi$, then there is $(Z, \pi) \in \mathscr{C}$ such that for all $p \in \mathrm{Pro}$: $Z \models p$ implies $X \models p$; and for all $p \in \mathrm{Con}$: $Z \models \neg p$ implies $X \models \neg p$. We show that the same implications hold for Y, so that $\mathscr{C}, Y \models \pi$ as desired. Note that since, $X \subseteq Y$, $X \models p$ implies $Y \models p$, and hence $Z \models p$ implies $X \models p$ implies $Y \models p$, for any $p \in \mathrm{Pro}$; the case for $p \in \mathrm{Con}$ is analogous. □

Remark 2.17. The phrasing of Proposition 2.16 relies on the use of partial fact situations—if a fact situation is defined on all values, meaning it assigns true or false to each factor, then there is no way in which additional information can be added to it. It may be argued that there is a difference here with respect to the representation of fact situations used by Horty (2011), where a fact situation X is given as a subset $X \subseteq F$ (cf. the discussion in Section 2.2). We could then ask whether $\mathscr{C}, X \models s$ and $X \subseteq Y$ implies $\mathscr{C}, Y \models s$, and indeed this implication does not hold. However, Horty (2011) does not specify the meaning of the absence of a factor from a fact situation, so $p \notin X$ could mean either that p explicitly does not hold in X, or that it is unknown whether p applies in X. In the former case, the addition of p to X does not constitute an addition, but rather a modification of information to X—the 'truth value' of p is changed from false to true. This would not align with the usual interpretation of monotonicity in logic; for example, $\Phi \cup \{\neg p\} \vdash \phi$ does not imply $\Phi \cup \{p\} \vdash \phi$ in general. Therefore, we feel this further distinction made by partiality of

fact situations is required to discuss monotonicity in the addition of information to fact situations.

Remark 2.18. It is interesting to note that for option (2) of the definition of constraint (see Remark 2.4), the notion of constraint is still monotonic in the case base, but *non*monotonic with respect to the addition of information to fact situations. This can be understood by looking at the row in Table 2.1 corresponding to the scenario where $X(p)$ is undefined, and $Y \models \neg p$. With option (2), constraint is not excluded in this scenario. Extending the fact situation X to Z (so $X \subseteq Z$) with $Z \models p$ would mean constraint is excluded, as illustrated by the fourth row in the table. This pattern, where options (1) and (2) give different results with respect to monotonicity of constraint, continues in the hierarchical version of the result model which we introduce in the next section (cf. Remark 4.15).

As an application of monotonicity, we will now formalize the observation made in Example 2.5 with regards to our treatment of undefined factors in Definition 2.3 of constraint, which can be phrased informally as follows. Given an outcome s and a precedent case (Y, s), the pro-s factors that are undefined in Y are treated as if they were false, and the con-s factors that are undefined in Y are treated as if they were true. Conversely, in a focus fact situation X the pro-s factors that are undefined in X are treated as if they were true, and the con-s factors are treated as if they were false. That this is true can be partially read off from Table 2.1: the rows in which $Y(p)$ is undefined are identical to the rows in which $Y(p)$ is false (with respect to the condition "if $Y \models \neg p$ then $X \models \neg p$"), and the same holds for $X(p)$. We will now phrase and prove this formally.

Let X be a fact situation, $s \in \{\pi, \delta\}$ an outcome, and define a fact situation $\underline{X}_s$ by:

$$\underline{X}_s(p) = \begin{cases} \mathbf{t} & \text{if } X(p) = \mathbf{t} \text{ or if } p \in \mathrm{Con}(s) \setminus \mathrm{dom}(X), \\ \mathbf{f} & \text{if } X(p) = \mathbf{f} \text{ or if } p \in \mathrm{Pro}(s) \setminus \mathrm{dom}(X). \end{cases} \tag{2.6.1}$$

Note that $X \subseteq \underline{X}_s$ and that $\underline{X}_s$ is complete, meaning $\mathrm{dom}(\underline{X}_s) = F$. We start with a lemma.

Lemma 2.19. *For $p \in \mathrm{Pro}(s)$, if $\underline{X}_s \models p$ then $X \models p$; for $p \in \mathrm{Con}(s)$, if $\underline{X}_s \models \neg p$ then $X \models \neg p$.*

Proof. Let $p \in \mathrm{Pro}(s)$ and suppose $\underline{X}_s \models p$. By (2.6.1) this means that either $X \models p$ or $p \in \mathrm{Con}(s) \setminus \mathrm{dom}(X)$. This latter option contradicts the assumption that $p \in \mathrm{Pro}(s)$, and so indeed $X \models p$. The reasoning for the second implication follows the same pattern. □

Let X and Y be some fact situations. We have the following propositions.

Proposition 2.20. *$\{(Y, s)\}, X \models s$ if and only if $\{(\underline{Y}_s, s)\}, X \models s$.*

Proof. For the left-to-right direction, we assume that $\{(Y, s)\}, X \models s$. Now, note that by Lemma 2.19 we have that for all $p \in \mathrm{Pro}(s)$, $\underline{Y}_s \models p$ implies $Y \models p$, and so $X \models p$ by the assumption that $\{(Y, s)\}, X \models s$. In the same vein Lemma 2.19 shows that for all $p \in \mathrm{Con}(s)$, $\underline{Y}_s \models \neg p$ implies $X \models \neg p$ and so indeed $\{(\underline{Y}_s, s)\}, X \models s$. The right-to-left direction follows readily from $Y \subseteq \underline{Y}_s$; for example if $p \in \mathrm{Pro}(s)$, then $Y \models p$ implies $\underline{Y}_s \models p$ and so $X \models p$ by assumption. □

Proposition 2.21. *$\{(Y, s)\}, X \models s$ if and only if $\{(Y, s)\}, \underline{X}_s \models s$.*

Proof. Left-to-right is an instance of Proposition 2.16. For the other direction, suppose that $\{(Y, s)\}, \underline{X}_s \models s$. Now if $p \in \mathrm{Pro}(s)$ and $Y \models p$, then $\underline{X}_s \models p$ by assumption and so $X \models p$ by Lemma 2.19. The case for $p \in \mathrm{Con}(s)$ is similar, and so we indeed have $\{(Y, s)\}, X \models s$. □

Lastly, we show that Propositions 2.20 and 2.21 also hold for case bases. To this end, let $\mathscr{C}$ be a case base, and define $\underline{\mathscr{C}} = \{(\underline{Y}_s, s) \mid (Y, s) \in \mathscr{C}\}$.

Theorem 2.22. *$\mathscr{C}, X \models s$ if and only if $\underline{\mathscr{C}}, \underline{X}_s \models s$.*

Proof. For left-to-right, suppose that $\mathscr{C}, X \models s$; then there is some $(Y, s) \in \mathscr{C}$ such that $\{(Y, s)\}, X \models s$. By Propositions 2.20 and 2.21 we thus have $\{(\underline{Y}_s, s)\}, \underline{X}_s \models s$, and so since $\{(\underline{Y}_s, s)\} \subseteq \underline{\mathscr{C}}$ we get $\underline{\mathscr{C}}, \underline{X}_s \models s$ from Proposition 2.12. The other direction is analogous. □

Theorem 2.22 shows that any instance of partial constraint can be understood as an instance of complete constraint, in which pro-s factors are taken to be false, and undefined con-s factors are taken to be true. In this sense, our approach to partial constraint can be seen as assigning *default* truth values to undefined factors, relative to an outcome s. Such default values have played a role in the study of nonmonotonic logics (Strasser & Antonelli, 2019, Section 3.4), but a further comparison between our definitions and default values in logic is beyond the scope of this work.

Example 2.23. We turn once more to the set of recidivism factors in 2.1.1. Suppose Y was a fact situation with $Y \models \{\textbf{?Record}, \textbf{Male}, \neg\textbf{Education}, \textbf{?Married}, \neg\textbf{Age}\}$. Furthermore, suppose Y was deemed to pose a recidivism risk, so that we have a case $(Y, 1)$. Proposition 2.20 tells us Y induces the same constraint as its completion $(\underline{Y}_1, 1)$. Applying the definition in (2.6.1) we see that $\underline{Y}_1 \models \{\neg\textbf{Record}, \textbf{Married}\}$. In other words, our proposed definition of constraint treats $(Y, 1)$ as if Y assigned true to the **Married** factor, and false to the **Record** factor. The same holds for a focus fact situation X satisfying, for example, $X \models \{\textbf{?Record}, \neg\textbf{Male}, \textbf{Education}, \textbf{?Married}, \textbf{Age}\}$; Proposition 2.21 states that X is forced for outcome 1 by a precedent case if and only if $\underline{X}_1$ is forced by that precedent case, and $\underline{X}_1 \models \{\neg\textbf{Record}, \textbf{Married}\}$. Note, however, that this situation is reversed with respect to constraint for outcome 0: we then have $\underline{X}_0 \models \{\textbf{Record}, \neg\textbf{Married}\}$.

Remark 2.24. For option (2) of Definition 2.3 a similar result holds, but then both pro-s and con-s are interpreted as being false, rather than just the pro-s factors.

2.7 The reason model

In addition to the result model Horty (2011) introduced a second model, called the *reason model*. It expands on the result model by supplementing the definition of a case with a *reason* for the decision, in the form of a subset of the factors pro the outcome of the case. In this thesis the focus is on the result model and not on the reason model, because the former is more readily applied to AI-related purposes. Nevertheless, the reason model has continued to receive attention in the literature, and in particular in the context of computable normative reasoning (Bench-Capon, 2024; Canavotto & Horty, 2022, 2023a, 2023b). In this section, as the last of our considerations on the topic of factor-based contraint, we investigate the relation between the result and reason models. We recall the definitions of the reason model, and show that its associated notion of consistency is interdefinable with

that of the result model. By relating the two models, we hope that our findings regarding the result model in the forthcoming chapters may be of use to work on the reason model.

For this discussion we will use the same knowledge representation as the one discussed in Section 2.2, except that we will now use the subset representation of fact situations (see the discussion in Section 2.2). In other words, a fact situation is now a subset $X \subseteq F$ of the factors, meaning that a factor is deemed to apply in X if and only if it is an element of X. Given a fact situation $X \subseteq F$ and an outcome s we write $X(s) = X \cap \text{Pro}(s)$ for the set of pro-s factors that apply in X.

Definition 2.25. A *reason* for an outcome s is a subset $U \subseteq \text{Pro}(s)$. A reason *applies* in a fact situation X when $U \subseteq X$.

Definition 2.26. A *result case* is a pair (X, s) consisting of a fact situation X and an outcome s. A *reason case* is a triple (X, U, s) such that (X, s) is a result case and U a reason for s that applies in X. A *result (reason) case base* is a finite set of result (reason) cases.

If confusion is unlikely to arise then we may simply speak of a case (base) rather than of a result or reason case (base).

Intuitively, a result case (X, s) represents a decision that the factors in $X(s)$ outweigh the factors in $X(\bar{s})$. A reason case (X, U, s) contains an additional reason U, which represents that the pro-s factors in U were already sufficient to outweigh the con-s factors in X. For the sake of comparison we restate the definition of constraint and consistency for the subset representation of cases.

Definition 2.27. Let $\mathscr{C}$ be a case base for the result model, then the outcome of a fact situation X is *forced* for outcome s by $\mathscr{C}$, denoted $\mathscr{C}, X \vDash s$, when there is a case $(Y, s) \in \mathscr{C}$ such that $Y(s) \subseteq X(s)$ and $X(\bar{s}) \subseteq Y(\bar{s})$. A case base for the result model is *inconsistent* if it contains a case $(X, s) \in \mathscr{C}$ such that $\mathscr{C}, X \vDash \bar{s}$, and *consistent* otherwise.

A case base $\mathscr{C}$ is inconsistent when it contains a case (X, s) that was decided for side s, while also containing a precedent $(Y, \bar{s})$ that forces a decision of X for the other side $\bar{s}$. Note that this witnessing precedent case will, by definition, have its own outcome forced for s by the case (X, s). This shows that inconsistency can be quantified in terms of inconsistent pairs, as shown in the following proposition.

Lemma 2.28. *A case base $\mathscr{C}$ for the result model is inconsistent iff it contains a pair of cases $(X, s), (Y, \bar{s}) \in \mathscr{C}$ such that $X(s) \subseteq Y(s)$ and $Y(\bar{s}) \subseteq X(\bar{s})$.*

The inconsistency of a reason-model case base is defined somewhat differently, at least in form. It relies on a preference relation on reasons, induced by a case.

Definition 2.29. Any case $c = (X, U, s)$ for the reason model induces a *preference relation* $<_c$, which is defined on reasons V, W by $V <_c W$ if $V \subseteq X(\bar{s})$ and $U \subseteq W$.

Definition 2.30. Any case base $\mathscr{C}$ for the result model induces a *preference relation* $<_{\mathscr{C}}$ on reasons V, W, defined by $V <_{\mathscr{C}} W$ when $V <_c W$ for some case $c \in \mathscr{C}$.

Definition 2.31. A case base $\mathscr{C}$ for the reason model is *consistent* if its induced preference relation $<_{\mathscr{C}}$ is asymmetric, and *inconsistent* otherwise.

Indeed, at first sight Definitions 2.27 and 2.31 do not have much in common, with one relying on a notion of outcome forcing, and the other on asymmetry of an induced preference relation on reasons. However, the following lemma shows that the notion of inconsistency of the reason model is closely related to that of the result model.

Lemma 2.32. *A case base $\mathscr{C}$ for the reason model is inconsistent iff it contains two cases $(X, U, s), (Y, V, \bar{s}) \in \mathscr{C}$ such that $U \subseteq Y(s)$ and $V \subseteq X(\bar{s})$.*

Proof. We begin by proving left-to-right; let W, Z be reasons such that $W <_{\mathscr{C}} Z$ and $Z <_{\mathscr{C}} W$. The assumption $W <_{\mathscr{C}} Z$ tells us that there is a case (X, U, s) with $W \subseteq X(\bar{s})$ and $U \subseteq Z$; likewise, $Z <_{\mathscr{C}} W$ means there is a case $(Y, V, \bar{s})$ such that $Z \subseteq Y(s)$ and $V \subseteq W$. This means $U \subseteq Z \subseteq Y(s)$ and $V \subseteq W \subseteq X(\bar{s})$ as desired. For right-to-left, we note that $U \subseteq Y(s)$ and $V \subseteq V$ means $U <_{(Y,V,\bar{s})} V$ and so $U <_{\mathscr{C}} V$. Similarly, we can conclude $V <_{\mathscr{C}} U$ from $V <_{(X,U,s)} U$, which means the case base $\mathscr{C}$ is indeed inconsistent. □

Lemmas 2.28 and 2.32 show us that both notions of consistency are about pairs of cases that are in a certain relation, with the only difference being that the reasons of the reason model take the place of the complete set of factors favoring the case outcome. This observation can further be made precise by interdefining both notions, as follows: let

$$f : (X, U, s) \mapsto (X(\bar{s}) \cup U, s), \tag{2.7.1}$$

this operation makes a result-model case from a reason-model case (X, U, s) by removing all the 'redundant' pro-s factors from its fact situation X, i.e. it removes the factors in $X(s) \setminus U$. Given a case base $\mathscr{C}$ for the reason model we thus get a case base $f[\mathscr{C}] = \{f(c) \mid c \in \mathscr{C}\}$ for the result model.

Proposition 2.33. *A case base $\mathscr{C}$ for the reason model is inconsistent iff $f[\mathscr{C}]$ is.*

Proof. By definition, $f[\mathscr{C}]$ is inconsistent iff it contains a case $(Y(s) \cup V, \bar{s}) \in f[\mathscr{C}]$, for some $(Y, V, s) \in \mathscr{C}$, such that $f[\mathscr{C}], Y(s) \cup V \vDash s$. This in turn means $f[\mathscr{C}]$ contains a case $(X(\bar{s}) \cup U, s)$, for $(X, U, s) \in \mathscr{C}$, such that $U \subseteq Y(s)$ and $V \subseteq X(\bar{s})$. This condition is, by Lemma 2.32, equivalent to the inconsistency of $\mathscr{C}$. □

The above shows that reason-model consistency can be defined in terms of result-model consistency. Of course, the converse can also be done. To show this, we define a function $g : (X, s) \mapsto (X, X(s), s)$ which turns a result-model case (X, s) into a reason-model case $(X, X(s), s)$ by taking all pro-s factors $X(s)$ in X as the reason for the decision for s; again we define $g[\mathscr{C}] = \{g(c) \mid c \in \mathscr{C}\}$.

Proposition 2.34. *A case base $\mathscr{C}$ for the result model is inconsistent iff $g[\mathscr{C}]$ is.*

Proof. By Lemma 2.32, $g[\mathscr{C}]$ is inconsistent iff it contains $(X, X(s), s), (Y, Y(\bar{s}), \bar{s})$, for some $(X, s), (Y, \bar{s}) \in \mathscr{C}$, such that $X(s) \subseteq Y(s)$ and $Y(\bar{s}) \subseteq X(\bar{s})$, which coincides with the definition of inconsistency for $\mathscr{C}$. □

Example 2.35. We consider the factors in (2.1.1) to exemplify the correspondence between the result and reasons models. Suppose that a fact situation Y with $Y \vDash \{\textbf{Male}, \textbf{Education}, \textbf{Married}, \textbf{Age}\}$ was deemed a low recidivism risk, so that we have a case $(Y, 0)$. The result-model interpretation of this decision is that the pro-0 factors

{**Education**, **Married**, **Age**} outweigh the con-0 factor **Male**. According to Definition 2.3 of result-model constraint, the precedent case $(Y, 0)$ constrains any subsequent fact situation that assigns true to the pro-0 factors {**Education**, **Married**, **Age**} and false to the con-0 factor **Record** to the outcome 0.

The reason-model representation of cases also allows for this case to include a subset of the pro-0 factors which were already sufficient to outweigh the con-0 factor **Male**. Suppose this reason was the subset {**Education**, **Age**}. Reason-model constraint dictates that new decisions should be made that are consistent with the precedent, in the sense of Definition 2.27. Therefore, the case $(Y, \{\textbf{Education}, \textbf{Age}\}, 0)$ constrains any subsequent fact situation that assigns true to the pro-0 factors {**Education**, **Age**} and false to the con-0 factor **Record** to the outcome 0.

Applying the function f in (2.7.1) to this situation we have

$$\begin{aligned} &f(\{\textbf{Male}, \textbf{Education}, \textbf{Married}, \textbf{Age}\}, \{\textbf{Education}, \textbf{Married}\}, 0) \\ &= (\{\textbf{Male}\} \cup \{\textbf{Education}, \textbf{Age}\}, 0) \\ &= (\{\textbf{Male}, \textbf{Education}, \textbf{Age}\}, 0). \end{aligned}$$

In other words, f deletes the pro-0 factor **Married** from Y. Proposition 2.33 tells us that the reason-model case $(\{\textbf{Male}, \textbf{Education}, \textbf{Married}, \textbf{Age}\}, \{\textbf{Education}, \textbf{Married}\}, 0)$ induces the same constraint as the result-model case $(\{\textbf{Male}, \textbf{Education}, \textbf{Age}\}, 0)$. This means that the reason model can be thought of as constituting a specific way of using the result model: when a fact situation X has been decided for outcome s based on a reason U, simply record the case as $(U \cup X(\bar{s}), s)$ rather than as (X, s). The resulting result-model case base will induce exactly the same constraint as its reason-model counterpart.

2.8 Conclusion: Moving from factors to dimensions

In this chapter we considered the RM, which formally describes what it means to make decisions in accordance with an a fortiori principle. We recalled its notion of constraint as formulated by Horty (2004), and adapted the model to operate on the basis of incomplete information. We then discussed the associated notion of consistency, and showed that this form of reasoning is monotonic in both the addition of cases and in the addition of information to fact situations. Lastly, we compared the result model to its close relative—the reason model. We showed that their associated notions of consistency are interdefinable, in an attempt to bring work on these models closer together.

The RM operates on the basis of a knowledge representation using factors: binary fact patterns that may or may not apply and are of significance to case outcomes, in that they tend to favor one side or the other. Since its introduction, Horty (2019) has expanded his model to operate on the basis of *dimensions*, where a dimension is understood as "an ordered set of legally significant values, where the ordering among values reflects the extent to which that value favors one side or the other." This adaptation is particularly relevant for AI applications, because the features of machine learning data are almost always non-binary.

In the next chapter, we will focus on this dimensional extension of the result model. We recall its definitions, and adapt it to operate on partial fact situations in the same way we did for the RM. Additionally, we view it through order-theoretic and logical perspectives, which will help us build a software implementation in Part II.

Chapter 3

Modeling Dimensional Constraint

BUILDING ON EARLIER WORK, Horty (2019) presented an adaptation of the result model (RM) that models a fortiori reasoning with dimensional information. We will refer to this as the dimensional result model (DRM). In this chapter, we discuss this model, and extend it to operate on the basis of partial fact situations, in the same way we did for the RM.[1]

We start in Section 3.1 by adapting our running example from Section 2.1 to incorporate dimensional information. We then give Horty's formalization of dimensions as a knowledge representation framework in Section 3.2, and his accompanying notion of constraint in Section 3.3. Then, in Section 3.4, we modify the knowledge representation to allow fact situations to be partial, and give a corresponding notion of constraint. We refer to the resulting model as the "dimensional result model" (DRM). In Section 3.5 we discuss the concept of case base consistency, and introduce a closely related notion of case base completeness. Then, in Section 3.6, we show that the DRM (like the RM) is monotonic in the addition of cases to the case base as well as in the addition of information to the focus fact situation. In Section 3.7 we show that the DRM is a conservative extension of the RM. As the last of our considerations on this topic, and as a step towards our data analyses in Part II of this thesis, we relate the DRM to order theory and many-sorted logic in Sections 3.9 and 3.10, respectively. We then end with some concluding remarks in Section 3.11.

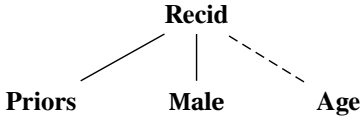
3.1 An example of dimensions

A factor can be thought of as an abstract binary property of situations, which either applies to a particular situation or not (regardless of whether this status is known to the court). A *dimension*, on the other hand, is a property of situations which can take a value from some set of possible values. In this sense, a factor is a special case of a dimension—one which

[1]The material on the presentation of the dimension-based model, its adaptation to partial fact situations, the comparison to the factor-based version of the model, and the monotonicity property, stem from (van Woerkom et al., 2025). The discussion on consistency and completeness, landmark cases, and the relation of the model to order theory and logic, were published in (van Woerkom et al., 2022a, 2024a).

takes a value from a two-element set. Usage of this terminology in the field of AI & law dates back to CATO's predecessor HYPO (Ashley, 1991).

For an example of dimensions, we return to our running example from the previous chapter. Consider the following illustration of some dimensions affecting an assessment of recidivism risk:



(3.1.1)

Previously, the **Age** factor represented whether the defendant was over the age of 21. Viewed as a dimension, **Age** can take any value above 0. Similarly, we replace the **Record** factor with a dimension **Priors**, indicating the number of previous convictions. The **Male** factor remains as it was, but is now regarded as a binary dimension.

Generally we do not say directly of a dimension whether it favors one of the two outcomes of a case. Instead, we require the dimension to come with a relation expressing the relative preference the values have for the final judgment. Research on recidivism has pointed out that in general older people tend to recidivate less, and so for the **Age** dimension we can say that the value 30 is less indicative than the value 21.

In (3.1.1) above we have again used solid and dotted links to indicate whether higher values of the dimension are suggestive of high or low risk of recidivism. Dimensions with two values, such as the **Male** dimension, generally correspond to factors as in the RM. The polarity of the links chosen here is in line with research on the topic of recidivism risk, see for example the data analysis by Angelino et al. (2018).

3.2 Knowledge representation

We now describe the knowledge representation of the DRM as used by Horty (2019). A *dimension* is a nonempty set; we denote dimensions by lower case letters d, e, f, etc. The domain is modeled by a finite set of dimensions D. A *fact situation* X is a choice function on D, i.e. a function $X : D \to \bigcup D$ such that $X(d) \in d$ for every $d \in D$.

Remark 3.1. Note that some care should be taken to avoid dimensions "collapsing" in the set-theoretic foundations of the theory. For example, in our running example in the criminal justice domain, we could have two dimensions $\textbf{Priors}_{\text{misd}}$ and $\textbf{Priors}_{\text{fel}}$, which respectively model the number of prior misdemeanor offenses and the number of prior felony offenses. With respect to influence on a high/low judgment of recidivism risk, these would both be modeled as the set of natural numbers ordered by the less-than relation, and therefore would become equal as sets—so $\textbf{Priors}_{\text{misd}} = \textbf{Priors}_{\text{fel}}$. To avoid this we should ensure that they are treated as distinct sets, even if they have the same underlying set of values. This can be achieved in a variety of ways, such as by using tagged sets or indexed sets to distinguish between the two dimensions. For the sake of simplicity we will henceforth leave such distinctions implicit.

Cases are again decided for one of the two sides π or δ. As before (cf. Section 2.2) we will also use the numbers 0 and 1 to denote the two possible case outcomes. We again assume that specific values of dimensions have a preference for either of these sides, but

this is now modeled by a binary relation on the dimension. More specifically, we assume there is for each dimension $d \in D$ a *preference* relation $\preceq$ on d, which we require to be a partial order.

Definition 3.2. A *partial order* $\leq$ on a set P is a relation satisfying the properties:

(1) $a \leq a$ for all $a \in P$;

(2) if $a \leq b$ and $b \leq c$ then $a \leq c$ for all $a, b, c \in P$;

(3) if $a \leq b$ and $b \leq a$ then $a = b$ for all $a, b \in P$.

These properties are respectively called *reflexivity*, *transitivity*, and *antisymmetry*. We say that a partial order is *total*, or *linear*, if for all $a, b \in P$ we have that $a \leq b$ or $b \leq a$.

Given values $v, w \in d$ such that $v \preceq w$, we say w prefers outcome π relative to v, and v prefers outcome δ relative to w. We often want to compare preference towards an arbitrary outcome s, so to do this we define for any dimension $(d, \preceq)$ the notation $\preceq_s = \preceq$ if $s = \pi$ and $\preceq_s = \succeq$ if $s = \delta$. Note that by definition we have $\preceq_s = \succeq_{\bar{s}}$ (where $\bar{s}$ denotes, as before, the outcome opposite to s).

3.3 Constraint

The notion of constraint for the DRM can now be stated succinctly as follows.

Definition 3.3. The decision of a fact situation X is *forced* for π by a case base $\mathscr{C}$, denoted $\mathscr{C}, X \vDash \pi$, if there is a case $(Y, \pi) \in \mathscr{C}$ with $Y(d) \preceq X(d)$ for all $d \in D$. Similarly, the decision is *forced* for δ, denoted $\mathscr{C}, X \vDash \delta$, if there is $(Y, \delta) \in \mathscr{C}$ such that $X(d) \preceq Y(d)$ for all $d \in D$.

Example 3.4. We consider an example to demonstrate the behavior of Definition 3.7, based on the set of dimensions illustrated in (3.1.1). These dimensions are given by the sets:

$$\begin{aligned} \textbf{Priors} &= \{0, 1, 2, \ldots\}, \\ \textbf{Male} &= \{0, 1\}, \\ \textbf{Age} &= \{18, 19, 20, \ldots\}. \end{aligned}$$

The associated orders are the usual less-than relation $\leq$ on the natural numbers for the **Priors** and **Male** dimensions, while for **Age** we take its inverse $\geq$. Compare this to the factors of the previous chapter in Section 2.1. Rather than having a binary criminal records indicator **Record**, we now have a dimension **Priors** specifying the number of previous convictions. In the same vein **Age** can now be specified as a number. The **Male** factor is still present, but now in the form of a binary, linearly-ordered dimension.

Suppose that based on these dimensions a judgment is made about a defendant being at high (π) or low (δ) risk of recidivism. Let (Y, π) be a case with a fact situation Y given by $Y(\textbf{Priors}) = 10$, $Y(\textbf{Male}) = 0$, and $Y(\textbf{Age}) = 20$. What constraint does the case base

$\{(Y,\pi)\}$ induce on a focus fact situation X? Applying Definition 3.3 we find:

$\{(Y,\pi)\}, X \vDash \pi$
iff for all $d \in \mathscr{D}$: $Y(d) \preceq X(d)$
iff $Y(\textbf{Priors}) \preceq X(\textbf{Priors})$, $Y(\textbf{Age}) \preceq X(\textbf{Age})$, and $Y(\textbf{Male}) \preceq X(\textbf{Male})$
iff $10 \leq X(\textbf{Priors})$, $20 \geq X(\textbf{Age})$, and $0 \leq X(\textbf{Male})$
iff $10 \leq X(\textbf{Priors})$ and $20 \geq X(\textbf{Age})$.

In other words, any defendant which is 20 or younger and has 10 or more prior offenses is constrained to be judged at a high risk of recidivism by the earlier decision (Y,π). The value that X assigns to the **Male** dimension is inconsequential to this judgment, because whatever value it assigns will be above 0 in the order.

Horty (2019, Definition 12) used a shorthand to phrase Definition 3.3, called the *strength order* on fact situation.

Definition 3.5. Given fact situations X and Y we say Y is *at least as strong* as X for an outcome s, denoted $X \preceq_s Y$, if it is at least as strong for s on every dimension d:

$$X \preceq_s Y \quad \text{if and only if} \quad X(d) \preceq_s Y(d) \text{ for all } d \in D.$$

Using this notion, we see that $\mathscr{C}, X \vDash s$ if and only if $\mathscr{C}$ contains some case (Y,s) such that $Y \preceq_s X$. We have opted to phrase constraint in Definition 3.3 without reference to the strength order to facilitate a comparison to the definitions of constraint of the other versions of the result model that we discuss. Nevertheless, we mention the strength order here because it will feature prominently in the forthcoming Sections 3.9 and 3.10.

Remark 3.6. Note that the DRM (as well as the RM) contains an independence assumption between dimensions: if $X(d) \preceq Y(d)$ holds for fact situations X and Y, then Y is considered stronger for the plaintiff along dimension d, regardless of its values in other dimensions. This need not always hold in practice—Prakken and Sartor (1998) give the following example: "even if rain and heat are individually reasons not to go jogging, then the combination of these two factors might very well be instead a reason to go jogging." In situations where this assumption is violated the result model may incorrectly impose constraint.

The factors of the knowledge representation for the RM can be modeled in the DRM as two-element dimensions. More specifically, given a factor $p \in F$ for some set of factors F, we can construct a dimension $d_p = \{\mathbf{f}_p, \mathbf{t}_p\}$ which contains exactly two elements corresponding to whether p applies or not. If $p \in \text{Pro}(\pi)$ then we define a relation $\prec$ on d_p by $\mathbf{f}_p \prec \mathbf{t}_p$, and if $p \in \text{Pro}(\delta)$ then we define $\prec$ by $\mathbf{t}_p \prec \mathbf{f}_p$. This procedure gives us a method to turn an instance of the RM into an instance of the DRM, and we will show in Section 3.7 that the notion of constraint in the DRM behaves on these translated instances just as the notion of constraint of the RM behaves on the untranslated instances.

3.4 Partial fact situations

As in the previous chapters, we consider a modification of the DRM by allowing fact situations to be partial. A partial dimensional fact situation is a partial choice function

on the set of dimensions, i.e. a function $X : E \to \bigcup D$ such that $X(d) \in d$ for every $d \in E$, where E is some subset $E \subseteq D$. We denote this domain E of X by $\mathrm{dom}(X)$.

Again this raises the question of whether, and if so how, Definition 3.3 of constraint should be modified to account for this. In Section 2.4 we outlined our position regarding constraint induced by partial cases, and it can be applied in the same way to constraint for the DRM: if the value of the precedent case in a dimension added strength to the side for which it was decided, then the focus fact situation should have equal-or-greater strength along that dimension for that side.

Consider, for example, a precedent case (Y, π) and a focus fact situation X; what conditions need to be met in order for (Y, π) to constrain the decision of X to s? If d is a dimension such that $Y(d)$ is undefined, then the value of Y on this dimension did not add strength to the side of the plaintiff and so no restrictions should be placed on $X(d)$. If, instead, d is a dimension on which $Y(d)$ is defined, then it is difficult to quantify whether $Y(d)$ added strength for the plaintiff, and so we should require that $Y(d) \preceq X(d)$. In particular, if $X(d)$ is undefined then X does not have equal-or-greater strength along this dimension, and so no constraint should be induced. There is, however, one edge case to take into account: when $Y(d)$ is the least element of d—meaning it satisfies $Y(d) \preceq v$ for every $v \in d$—then any value of $X(d)$ will match or exceed the strength for the side of the plaintiff in that dimension. Therefore, we should allow (Y, π) to induce constraint for X in this scenario. This is similar to the situation in the third row of Table 2.1: if $p \in \mathrm{Con}$ and $Y \vDash p$, then regardless of whether p applies in the focus fact situation X or not, it will match the strength of Y for the plaintiff for that factor, and so constraint is not excluded.

To capture these considerations in the definition of constraint we introduce the following shorthands:[2]

$$\mathrm{supp}(Y) = \{d \in \mathrm{dom}(Y) \mid Y(d) \text{ is not the least element of } d\}, \tag{3.4.1}$$

$$\mathrm{opp}(Y) = \{d \in \mathrm{dom}(Y) \mid Y(d) \text{ is not the greatest element of } d\}. \tag{3.4.2}$$

We now define constraint as follows, which is similar to the approach taken by Prakken (2021, Section 7).

Definition 3.7. The decision of a partial fact situation X is *forced* for π by a case base $\mathscr{C}$, denoted $\mathscr{C}, X \vDash \pi$, if there is a partial case $(Y, \pi) \in \mathscr{C}$ such that for all $d \in \mathrm{supp}(Y)$: $Y(d) \preceq X(d)$. Similarly, the decision is *forced* for δ, denoted $\mathscr{C}, X \vDash \delta$, if there is $(Y, \delta) \in \mathscr{C}$ such that for all $d \in \mathrm{opp}(Y)$: $X(d) \preceq Y(d)$.

As in Chapter 2 on the result model without dimensions (Definitions 2.1 and 2.3), Definitions 3.3 and 3.7 coincide for complete fact situations. In what follows we focus on partial fact situations and leave the treatment of complete fact situations implicit.

Example 3.8. We reconsider Example 3.4 with respect to Definition 3.7. Let Y be as

[2] This notation is shorthand for "support" and "opposition." Function support is a set-theoretic notion, where the support $\mathrm{supp}(f)$ of a function $f : X \to \mathbb{R}$ from a set X to the real numbers $\mathbb{R}$ is defined as the subset of X that is not mapped to 0 by f; i.e. $\mathrm{supp}(f) = \{x \in X \mid f(x) \neq 0\}$.

before; unfolding the definition of constraint we find:

$$
\begin{aligned}
&\{(Y,\pi)\}, X \vDash \pi \\
&\text{iff for all } d \in \operatorname{supp}(Y):\ Y(d) \preceq X(d) \\
&\text{iff } Y(\textbf{Priors}) \preceq X(\textbf{Priors}) \text{ and } Y(\textbf{Age}) \preceq X(\textbf{Age}) \\
&\text{iff } 10 \leq X(\textbf{Priors}) \text{ and } 20 \geq X(\textbf{Age}).
\end{aligned}
$$

The answer is the same as before, because $Y(\textbf{Male})$ is the least element of the **Male** dimension, and so $\textbf{Male} \notin \operatorname{supp}(Y)$. This means that if X is undefined on this dimension, Y can still induce constraint on X.

3.5 Consistency and completeness

As we did for the RM in Section 2.5, we now turn to the notion of consistency, which arises as a result of the definition of constraint.

Definition 3.9. A case (X, s) is said to be *inconsistent* with respect to a case base $\mathscr{C}$ when $\mathscr{C}, X \vDash \bar{s}$, and *consistent* otherwise. A case base is said to be *weakly consistent* when all of its cases are consistent, and *strongly inconsistent* otherwise.

Definition 3.10. A case base $\mathscr{C}$ is *weakly inconsistent* if there exists a fact situation X such that $\mathscr{C}, X \vDash \pi$ and $\mathscr{C}, X \vDash \delta$, and *strongly consistent* otherwise.

Remark 3.11. Note that if a case base has only complete fact situations, then the presence of an inconsistent case with outcome s implies the presence of an inconsistent case with outcome $\bar{s}$. To see why, consider an inconsistent case $(X, s) \in \mathscr{C}$; so $\mathscr{C}, X \vDash \bar{s}$. This means there is a case $(Y, \bar{s}) \in \mathscr{C}$ such that $Y \preceq_{\bar{s}} F$. But then, by definition, $X \preceq_s Y$ and so $\mathscr{C}, Y \vDash s$; which is to say that $(Y, \bar{s})$ is inconsistent. In practice, this means that in order to check whether such a case base is consistent, it suffices to check whether all of its cases with a specific outcome s are consistent. This may save work, for example when the distribution of outcomes is heavily skewed towards one of the two outcomes.

Consistency, as thus defined, is a binary property: a case base is either consistent or it is not. It can be made a quantitative property by considering the relative frequency of consistent cases in the case base, and we will do so in our experiments to come in Part II.

Definition 3.10 of strong consistency states that there is no fact situation that has both outcomes forced. This phrasing has a natural counterpart which—to the best of our knowledge—has not yet appeared in the literature, and which is defined as follows.

Definition 3.12. A case base is *complete* when every fact situation has an outcome forced.

Remark 3.13. The terminology we propose in Definition 3.12 is inspired by the notion of completeness of a set of formulas in logic: a set of formulas T is *consistent* if there is no formula ϕ such that both $T \vDash \phi$ and $T \vDash \neg\phi$, and it is *complete* if for all formulas ϕ either $T \vDash \phi$ or $T \vDash \neg\phi$ (Bradley & Manna, 2007, Section 3.1). Compare this with our definitions here: a case base $\mathscr{C}$ is *consistent* if there is no fact situation F such that both $\mathscr{C}, F \vDash 0$ and $\mathscr{C}, F \vDash 1$, and it is *complete* if for all fact situations F either $\mathscr{C}, F \vDash 0$ or $\mathscr{C}, F \vDash 1$. The similarity is somewhat superficial, though, as there are important differences between these

Table 3.1: An example case base for the **Age** and **Priors** dimensions, which is neither consistent nor complete. See Figure 3.1 for a graphical representation. The second, third, and fifth row correspond respectively to the fact situations X, Y, and Z from Example 3.14.

	Age	**Priors**	**Label**
	30	1	0
X	35	5	0
	45	4	0
	30	2	1
	35	7	1
Y	40	3	1
Z	45	7	–

notions; for instance, an inconsistent set of logical formulas is necessarily complete by the *ex falso* principle, while for case bases this implication does not hold.

Example 3.14. An example case base for our recidivism example can be found in Table 3.1. This case base is neither consistent nor complete. It is not consistent because the case $(X, 0)$ with $X(\textbf{Age}) = 35$ and $X(\textbf{Priors}) = 5$ has its outcome forced for 1 by the case $(Y, 1)$ with $Y(\textbf{Age}) = 40$ and $Y(\textbf{Priors}) = 3$. It is not complete because there are (infinitely many) fact situations that do not have their outcome forced for either 0 or 1, such as the listed fact situation Z with $Z(\textbf{Age}) = 45$ and $Z(\textbf{Priors}) = 7$.

A case base might become complete, or inconsistent, through the addition of new cases. Conversely, a case base can be made incomplete, or consistent, through the removal of cases. Any set of dimensions D trivially admits a sound case base; namely the empty case base $\emptyset$. It would also trivially admit a complete case base if not for our requirement that case bases are finite: simply decide all fact situations for outcome 0, or 1, or any mix thereof. Since the choice of dimensions D may give rise to an infinite number of fact situations, this may be impossible, and in fact it is impossible for our running example—as we now show.

Proposition 3.15. *There is no complete case base for the* **Age** *and* **Priors** *dimensions.*

Proof. Let $\mathscr{C}$ be any case base for the **Age** and **Priors** dimensions; we prove the proposition by constructing a fact situation X which does not have its outcome forced by $\mathscr{C}$:

$$X(\textbf{Age}) = 1 + \max_{(Y,1)\in\mathscr{C}} Y(\textbf{Age}), \qquad X(\textbf{Priors}) = 1 + \max_{(Y,0)\in\mathscr{C}} Y(\textbf{Priors}).$$

This fact situation X exists because the case base $\mathscr{C}$ is finite. The claim is that X does not have its outcome forced by $\mathscr{C}$. Suppose, to the contrary, that X were forced for 0; then there is a case $(Z, 0) \in \mathscr{C}$ such that $X \preceq Z$. This means that $X(\textbf{Priors}) \preceq Z(\textbf{Priors})$, and so $X(\textbf{Priors}) = 1 + \max_{(Y,0)\in\mathscr{C}} Y(\textbf{Priors}) \preceq Z(\textbf{Priors})$ (note the order used here). This implies that $\max_{(Y,0)\in\mathscr{C}} Y(\textbf{Priors}) < Z(\textbf{Priors})$, contradicting $(Z, 0) \in \mathscr{C}$. If X had its outcome forced for 1 then a similar contradiction occurs with the definition of $X(\textbf{Age})$. □

This proposition demonstrates that there are sets of dimensions D which—from the onset—do not admit any complete case base, because of the requirement that case bases are finite. We maintain this requirement because real-world datasets are necessarily finite, and because it allows us in general to construct logical formulas describing the forcing behavior of the case base, which we will make use of in Section 3.10.

3.6 Monotonicity

Like the factor-based models, the DRM is monotonic in the case base.

Proposition 3.16. *If $\mathscr{C} \subseteq \mathscr{D}$ then $\mathscr{C}, X \models s$ implies $\mathscr{D}, X \models s$ for $s \in \{\pi, \delta\}$.*

Proof. Suppose $s = \pi$, and let X be a fact situation such that $\mathscr{C}, X \models \pi$, then this means there is a case $(Y, \pi) \in \mathscr{C}$ satisfying the required conditions of Definitions 3.7, and so since $\mathscr{C} \subseteq \mathscr{D}$ we get $(Y, \pi) \in \mathscr{D}$ and $\mathscr{D}, X \models \pi$ as desired. The case for $s = \delta$ is similar. □

Fact situations are again modelled by functions, so we can define the addition of information to fact situations in terms of function extension in the sense of Definition 2.15.

Proposition 3.17. *If $X \subseteq Y$ then $\mathscr{C}, X \models s$ implies $\mathscr{C}, Y \models s$ for $s \in \{\pi, \delta\}$.*

Proof. Suppose $s = \pi$. Since $\mathscr{C}, X \models \pi$ there is $(Z, \pi) \in \mathscr{C}$ such that $Z(d) \preceq X(d) = Y(d)$ for all $d \in \mathrm{supp}(Z)$, and so $\mathscr{C}, Y \models \pi$ as desired. The case for $s = \delta$ is similar. □

3.7 Relation to the result model

We will now show that the DRM is a conservative extension of the RM, in the sense that when the DRM is restricted to binary dimensions with linear orders, it reduces to the RM. To show this, we construct a translation f that maps instances of the RM to instances of the DRM, and prove that this translation respects Definitions 2.3 and 3.7 of constraint.

Definition 3.18. A set of dimensions D is *binary* if each $d \in D$ has cardinality 2 and is ordered linearly.

Let $F = \mathrm{Pro} \cup \mathrm{Con}$ be a factor partition, for each $p \in F$ we define a binary dimension $d_p = \{\mathbf{t}_p, \mathbf{f}_p\}$. The associated order for d_p is the reflexive closure of $\mathbf{f}_p \prec \mathbf{t}_p$ if $p \in \mathrm{Pro}$, and that of $\mathbf{t}_p \prec \mathbf{f}_p$ if $p \in \mathrm{Con}$. We write $f[F] = \{d_p \mid p \in F\}$ for the set of dimensions corresponding to the factor partition F. Next, we extend f to operate on fact situations, cases, and case bases. Given a fact situation X for the RM (with respect to the factor partition F) we translate it to a fact situation $f(X)$ for the DRM (with respect to the set of dimensions $f[F]$) by defining $f(X)$ on d_p for $p \in \mathrm{dom}(X)$:

$$f(X)(d_p) = \begin{cases} \mathbf{t}_p & \text{if } X \models p, \\ \mathbf{f}_p & \text{if } X \models \neg p. \end{cases}$$

Similarly, a case (X, s) is translated to a case $(f(X), s)$, and given a case base $\mathscr{C}$ for F we let $f[\mathscr{C}]$ denote $\{(f(X), s) \mid (X, s) \in \mathscr{C}\}$.

The translation f preserves and reflects constraint, in the following sense.

Theorem 3.19. *Given a case base $\mathscr{C}$ for a factor partition $F = \mathrm{Pro} \cup \mathrm{Con}$ and a focus fact situation X we have*

$$\mathscr{C}, X \models \pi \text{ iff } f[\mathscr{C}], f(X) \models \pi \quad \text{and} \quad \mathscr{C}, X \models \delta \text{ iff } f[\mathscr{C}], f(X) \models \delta.$$

Proof. We consider the first equivalence. Spelling out Definition 3.7 we get

$f[\mathscr{C}], f(X) \models \pi$
iff there is $(Y, \pi) \in f[\mathscr{C}]$ such that for all $d_p \in \mathrm{supp}(Y)$: $Y(d_p) \preceq f(X)(d_p)$
iff there is $(Y, \pi) \in \mathscr{C}$ such that for all $d_p \in \mathrm{supp}(f(Y))$: $f(Y)(d_p) \preceq f(X)(d_p)$
iff there is $(Y, \pi) \in \mathscr{C}$ such that for all $p \in \mathrm{dom}(Y)$:
if $f(Y)(d_p)$ is not the least element of d_p then $f(Y)(d_p) \preceq f(X)(d_p)$.

Now either $p \in \mathrm{Pro}$ or $p \in \mathrm{Con}$. In the former case, d_p is ordered by $\mathbf{f}_p \prec \mathbf{t}_p$ and so the least element of d_p is $\mathbf{f}_p$. The latter case is dual. We can thus continue with:

iff there is $(Y, \pi) \in \mathscr{C}$ such that:
- for all $p \in \mathrm{dom}(Y) \cap \mathrm{Pro}$: if $f(Y)(d_p) \neq \mathbf{f}_p$ then $f(Y)(d_p) \preceq f(X)(d_p)$, and
- for all $p \in \mathrm{dom}(Y) \cap \mathrm{Con}$: if $f(Y)(d_p) \neq \mathbf{t}_p$ then $f(Y)(d_p) \preceq f(X)(d_p)$.

If $p \in \mathrm{dom}(Y) \cap \mathrm{Pro}$ then $f(Y)$ is defined on d_p and so $f(Y)(d_p) \neq \mathbf{f}_p$ implies $f(Y)(d_p) = \mathbf{t}_p$, and furthermore $f(Y)(d_p) = \mathbf{t}_p \preceq f(X)(d_p)$ reduces to $f(X)(d_p) = \mathbf{t}_p$. Applying similar reasoning to the case for $p \in \mathrm{dom}(Y) \cap \mathrm{Con}$ we obtain:

iff there is $(Y, \pi) \in \mathscr{C}$ such that:
- for all $p \in \mathrm{dom}(Y) \cap \mathrm{Pro}$: if $f(Y)(d_p) = \mathbf{t}_p$ then $f(X)(d_p) = \mathbf{t}_p$, and
- for all $p \in \mathrm{dom}(Y) \cap \mathrm{Con}$: if $f(Y)(d_p) = \mathbf{f}_p$ then $f(X)(d_p) = \mathbf{f}_p$

iff there is $(Y, \pi) \in \mathscr{C}$ such that:
- for all $p \in \mathrm{dom}(Y) \cap \mathrm{Pro}$: if $Y(p) = \mathbf{t}$ then $X(p) = \mathbf{t}$, and
- for all $p \in \mathrm{dom}(Y) \cap \mathrm{Con}$: if $Y(p) = \mathbf{f}$ then $X(p) = \mathbf{f}$.

Intersecting with $\mathrm{dom}(Y)$ is redundant because for undefined values $Y(p)$ the implication over which is being quantified holds trivially. Furthermore, $Y(p) = \mathbf{t}$ just means the same as $Y \models p$, and likewise for $Y(p) = \mathbf{f}$ and $Y \models \neg p$, so we have:

iff there is $(Y, \pi) \in \mathscr{C}$ such that:
- for all $p \in \mathrm{Pro}$: if $Y \models p$ then $X \models p$, and
- for all $p \in \mathrm{Con}$: if $Y \models \neg p$ then $X \models \neg p$

iff $\mathscr{C}, X \models \pi$.

The derivation of $\mathscr{C}, X \models \delta$ iff $f[\mathscr{C}], f(X) \models \delta$ is very similar, so we omit it. □

3.8 Landmark cases

Of particular interest with respect to the forcing relation are what we call landmark cases. The motivating idea is that when a case has its outcome forced by another, it is—by

transitivity of the strength order—rendered superfluous as a precedent. As such, the most salient cases are those that do not have their outcome forced by another case.

Definition 3.20. A case $(X, s) \in \mathscr{C}$ is called a *landmark* case if $\mathscr{C} \setminus \{(X, s)\}, X \not\models s$. Cases that are not landmarks are called *regular*. We let $\mathscr{L} \subseteq \mathscr{C}$ denote the subset of $\mathscr{C}$ containing just its landmark cases.

Intuitively speaking, a case is a landmark when the decision of its fact situation is not forced for its outcome by the rest of the case base. Do note, however, that a case (X, s) can be a landmark while $\mathscr{C} \setminus \{(X, s)\}, X \models \bar{s}$.

The relevance of landmarks is described by the following proposition.

Proposition 3.21. *For a case base $\mathscr{C}$, a fact situation X, and an outcome s, we have*

$$\mathscr{C}, X \models s \iff \mathscr{L}, X \models s.$$

The direction from right to left is just an instance of monotonicity, but the other direction is somewhat more difficult to justify. We defer a proof until Section 3.9.3, as we will develop some notation in the meantime that will ease this task.

Remark 3.22. Note that Definition 3.20 applies equally well to the RM. We have introduced it here because it will play a role in the coming two sections.

3.9 An order-theoretic perspective

The mathematical tools used in Horty's model of a fortiori reasoning have been studied more generally, as part of a branch of mathematics known as order theory: the study of binary relations on sets that correspond intuitively to the notion of order (as in Definition 3.2). In this section, we recall some notions from order theory and relate them to Horty's model. We do this because they help clarify the formal aspect of Horty's model, and because we will make use of them in Section 3.10 where we relate the model to (many-sorted) logic. See the work by Davey and Priestley (2002) for a detailed introduction to order theory and its connection with logic.

In order not to overcomplicate things we will restrict our attention to complete fact situations—i.e. those X for which $\mathrm{dom}(X) = D$—throughout this section and the one thereafter.

3.9.1 The product order and its up- and down-sets

Given a set P of sets, the *product* of P, denoted by $\prod P$, is the set containing all choice functions on P;

$$\prod P = \{f : P \to \bigcup P \mid f(A) \in A \text{ for all } A \in P\}.$$

If every set $A \in P$ comes with a partial order $\leq_A$, then P can itself be partially ordered. In fact, this can be done in multiple ways, but we will use what is called the *coordinatewise order* or *product order* $\leq_\prod$, which is defined for $f, g \in \prod P$ by $f \leq_\prod g$ if and only if $f(A) \leq_A g(A)$ for all $A \in P$ (Davey & Priestley, 2002, p. 18).

We have seen a particular instance of this construction in Section 3.2; given a set of dimensions D we let $\mathscr{X} = \prod D$ denote the set of fact situations, and write $\preceq$ for the product order on $\mathscr{X}$, which in the theory of precedential constraint is known as the strength order.

A case base is a finite subset $\mathscr{C} \subseteq \mathscr{X} \times \{0,1\}$, but we can also think of $\mathscr{C}$ as comprising two designated subsets of $\mathscr{X}$; one $\mathscr{C}_0 \subseteq \mathscr{X}$ containing the fact situations of cases with outcome 0, and one $\mathscr{C}_1 \subseteq \mathscr{X}$ with those that received outcome 1. Given a case base $\mathscr{C}$, we identify these subsets with the notation $\mathscr{C}_s = \{X \in \mathscr{X} \mid (X, s) \in \mathscr{C}\}$.

A concept from order theory that we will use extensively is that of *up-sets* and *down-sets*. Given an ordered set $(P, \leq)$ and a subset $A \subseteq P$, we define its up- and down-sets $\uparrow A, \downarrow A$ by

$$\uparrow A = \{b \in P \mid a \leq b \text{ for some } a \in A\}, \qquad \downarrow A = \{b \in P \mid b \leq a \text{ for some } a \in A\}.$$

If A is a singleton $\{a\}$ we may write $\uparrow a$ instead of $\uparrow\{a\}$; note that $\uparrow A = \bigcup_{a \in A} \uparrow a$. These operations satisfy the closure operation conditions (Davey & Priestley, 2002, Chapter 7):

Lemma 3.23. *Let $(P, \leq)$ be a partially ordered set. The upset operation $\uparrow$ on P is a* closure operation*, meaning it satisfies the following properties for all subsets $A, B \subseteq P$:*

- $A \subseteq \uparrow A$*;*
- *if $A \subseteq B$ then $\uparrow A \subseteq \uparrow B$;*
- $\uparrow\uparrow A = \uparrow A$*.*

The same holds for the down-set operator $\downarrow$.

Working with fact situations we are also interested in the opposite of the product order, for which we use the notation $\preceq_s = \preceq$ if $s = 1$ and $\preceq_s = \succeq$ if $s = 0$. We will do the same for the up- and down-set notation: $\uparrow_s = \uparrow$ if $s = 1$ and $\uparrow_s = \downarrow$ if $s = 0$.

3.9.2 Forcing as up- and down-set membership

This concept of up- and down-sets is useful because it is closely related to the definition of case base forcing. The following is just a simple rephrasing of definitions, but we state it explicitly because we will use it frequently.

Lemma 3.24. *$\mathscr{C}, X \vDash s$ is equivalent to $X \in \uparrow_s \mathscr{C}_s$.*

Proof. We have $\mathscr{C}, X \vDash s \iff Y \preceq_s X$ for some $Y \in \mathscr{C}_s \iff X \in \uparrow_s \mathscr{C}_s$. □

Example 3.25. We consider again the case base of Example 3.14, listed in Table 3.1. The dimensionality of $\mathscr{X} = \mathbb{N} \times \mathbb{N}$ is low and so we can visualize it—see Figure 3.1. The up-sets of the cases are also shown, which we will call *forcing cones*. For instance, for the case $(X, 1)$ with $X(\textbf{Age}) = 30$ and $X(\textbf{Priors}) = 2$, any fact situation Y with $Y(\textbf{Age}) \leq 20$ and $Y(\textbf{Priors}) \geq 2$ will have greater strength for outcome 1 than X, and so will be forced for side 1.

The visualization in Figure 3.1 shows that concepts of interest can be phrased in terms of the forcing cones; for instance, we can see landmarks as cases that are not within a forcing cone of a case with its own outcome. Furthermore, inconsistency corresponds to

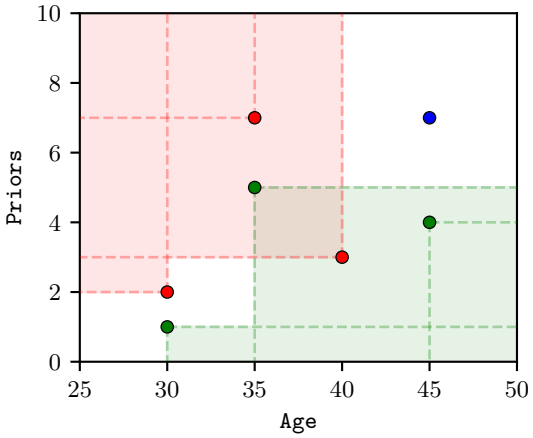


Figure 3.1: An illustration of the case base in Table 3.1. Green dots are cases with outcome 0, red dots are cases with outcome 1, and the shaded regions indicate their up- and down-sets in the strength order, which we call forcing cones. The blue dot is a counterexample to completeness.

overlapping cones of cases with opposite outcomes, and completeness corresponds to areas that are not covered by any cones. If there are no overlapping forcing cones of different outcomes, then the case base is consistent; likewise, if the whole space of fact situations is covered by the forcing cones, then the case base is complete. This is stated formally by the following lemma.

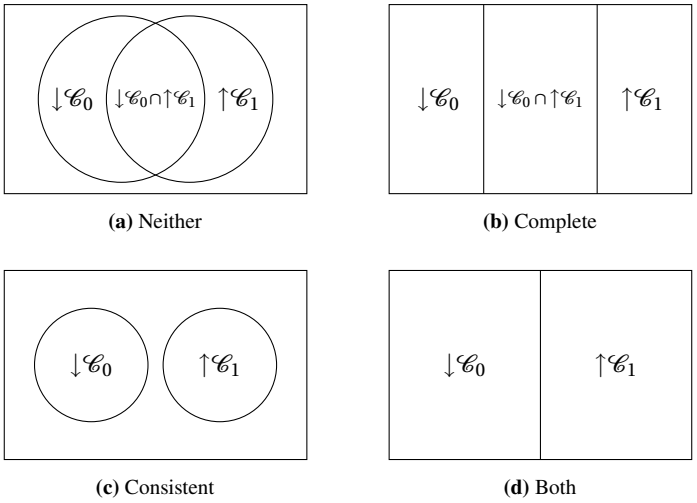
Lemma 3.26. *A case base $\mathscr{C}$ is consistent iff $\downarrow\mathscr{C}_0 \cap \uparrow\mathscr{C}_1 = \emptyset$, and complete iff $\downarrow\mathscr{C}_0 \cup \uparrow\mathscr{C}_1 = \mathscr{X}$.*

Remark 3.27. As mentioned in Remark 3.11, for consistency it also suffices to check either of the equations $\mathscr{C}_0 \cap \uparrow\mathscr{C}_1 = \emptyset$ or $\mathscr{C}_1 \cap \downarrow\mathscr{C}_0 = \emptyset$.

The visualization of Figure 3.1 is possible because our example has only two dimensions—with more than two such a visualization becomes impractical. However, in the general case we can still usefully visualize the forcing cones using Euler diagrams. For example, Lemma 3.26 relates consistency and completeness to the sets $\downarrow\mathscr{C}_0 \cap \uparrow\mathscr{C}_1$ and $\mathscr{X} \setminus (\downarrow\mathscr{C}_0 \cup \uparrow\mathscr{C}_1)$ being empty. So, the four possible situations with regards to the status of consistency and completeness of a case base can be visualized using Euler diagrams; see Figure 3.2. We will make use of such visualizations for our data analysis in Chapter 7.

3.9.3 Landmarks as minimal and maximal elements

Another useful concept from order theory is that of minimal and maximal elements. Given a partially ordered set $(P, \leq)$ and a subset $A \subseteq P$, we say an element $a \in A$ is *minimal in A* if there is no $b \in A$ such that $b < a$. We denote the set of minimal elements of A by $\min A$. Dually, we say $a \in A$ is *maximal in A* if there is no $b \in A$ such that $a < b$. We denote the set of maximal elements of A by $\max A$. Again, we define some notion to account for the two sides: $\min_s = \min$ if $s = 1$ and $\min_s = \max$ if $s = 0$.



(a) Neither **(b)** Complete

(c) Consistent **(d)** Both

Figure 3.2: Euler diagram representations of the consistency and completeness properties.

This notion makes it easier to understand the set of landmarks $\mathscr{L}$ of a case base $\mathscr{C}$:

$$\begin{aligned} X \in \mathscr{L}_s &\iff (X, s) \in \mathscr{L} \\ &\iff \mathscr{C} \setminus \{(X, s)\}, X \not\models s \\ &\iff Y \not\prec_s X \text{ for all } (Y, s) \in \mathscr{C} \\ &\iff X \in \min_s \mathscr{C}_s. \end{aligned}$$

This means that $\mathscr{L}_s = \min_s \mathscr{C}_s$, or more explicitly, that $\mathscr{L}_0 = \max \mathscr{C}_0$ and $\mathscr{L}_1 = \min \mathscr{C}_1$. This is a useful fact in combination with the following (well-known) lemma.

Lemma 3.28. *If $(P, \leq)$ is a partially ordered set then any finite $A \subseteq P$ satisfies $\uparrow A = \uparrow \min A$.*

Proof. We prove the inclusions separately, using the properties listed in Lemma 3.23.

- ($\subseteq$) First we note that since A is finite, there is for every $a_1 \in A$ a finite descending chain $a_1 > a_2 > \cdots > a_n$ within A, so $a_1 \geq a_n$ by transitivity, for some $a_n \in \min A$. This is to say that $A \subseteq \uparrow \min A$, and therefore $\uparrow A \subseteq \uparrow\uparrow \min A = \uparrow \min A$ as desired.
- ($\supseteq$) It follows from $\min A \subseteq A$ that $\uparrow \min A \subseteq \uparrow A$. □

Corollary 3.29. *For any case base $\mathscr{C}$ we have $\uparrow_s \mathscr{C}_s = \uparrow_s \mathscr{L}_s$.*

Proof. Immediate from Lemma 3.28 and the fact that $\mathscr{L}_s = \min_s \mathscr{C}_s$. □

Proposition 3.21 now also follows immediately from Corollary 3.29 and Lemma 3.24.

Proof of Proposition 3.21. $\mathscr{C}, X \models s \iff X \in \uparrow_s \mathscr{C}_s \iff X \in \uparrow_s \mathscr{L}_s \iff \mathscr{L}, X \models s$. □

Intuitively, what Proposition 3.21 tells us is that when a fact situation is forced by a case base then it is also forced by a landmark case of that case base. This means that in order to get an understanding of the behavior of the strength order of a case base, it suffices to consider the landmark cases. This can be a very useful reduction in practice—we will see some datasets that contain thousands of cases but only some handfuls of landmarks.

3.10 A logical perspective on a fortiori reasoning

In this section, we phrase the DRM from the point of view of logic. To do this we use many-sorted logic, so we begin in Section 3.10.1 by describing this general framework, and then proceed in Section 3.10.2 to demonstrate that the DRM can be phrased as an instance of this framework. There are many works in the literature giving such descriptions; see for example the work by de Moura and Bjørner (2009) and Manzano and Aranda (2022).

We use many-sorted logic, as opposed to e.g. Liu et al. (2022) who use modal logic, as it is the type of logic used in contemporary SMT solvers. This means that the rephrasing in logic allows us to use the machinery of SMT solvers to reason about specific instances of the DRM. We describe how this can be done in Chapter 7, and use this implementation in Part II to evaluate the DRM on several datasets.

3.10.1 Many-sorted logic

Many-sorted logic is very similar to unsorted logic—it revolves around questions of satisfiability of formulas built from the familiar logical connectives as well as function and relation symbols from some signature. The difference is that these symbols have some *sort*, which can be thought of as a datatype in programming. Examples of such sorts are those corresponding to integers, rationals, or boolean values.

More formally, we assume there is a finite set of sorts $S = \{s_1, \dots, s_n\}$.[3] A *signature* Σ over S is a set of *function* and *predicate* symbols, together with a map $\mathrm{ar} : \Sigma \to S^+$ where S^+ is the set of n-tuples of elements of S for all $n \geq 1$. The map ar associates each element of Σ with an *arity*. For a function symbol $f \in \Sigma$ we write $f : s_1 \times \dots \times s_n \to s$ instead of $\mathrm{ar}(f) = (s_1, \dots, s_n, s)$. If $n = 0$ we say f is a *constant*. Similarly, for a relation symbol $R \in \Sigma$ we write $R : s_1 \times \dots \times s_n$ instead of $\mathrm{ar}(R) = (s_1, \dots, s_n)$. In addition, we assume there is a set of *variables* X, each of which is associated with a sort; we write $x : s$ to denote that $x \in X$ is of sort $s \in S$.

Any signature Σ over a set of sorts S induces a set T^Σ of *terms*, each of which again has a specific sort. We write $t : s$ to mean that $t \in T^\Sigma$ is of sort $s \in S$. These terms are inductively defined as the smallest set t^Σ such that:

- any variable $x : s$ is a term of sort s;
- given a function symbol $f : s_1 \times \dots \times s_n \to s$ of Σ and terms $t_1 : s_1, \dots, t_n : s_n$ there is a term $f(t_1, \dots, t_n) : s$ of sort s.

Formulas of many-sorted logic are built using the terms T^Σ together with the usual logical symbols $\top, \bot, \wedge, \vee, \neg, \to, \leftrightarrow$, and with an equality symbol $\doteq_s : s \times s$ for every sort $s \in S$. In practice, we will usually omit the subscript s and let the context determine which of the equality symbols is used. Using these we build the set L^Σ_{at} of *atomic* Σ*-formulas* with:

- $\bot, \top$;
- $t_1 \doteq_s t_2$ for every sort $s \in S$ and terms $t_1 : s,\ t_2 : s$;
- $R(t_1, \dots, t_n)$ for every relation symbol $R : s_1 \times \dots \times s_n$ and terms $t_1 : s_1, \dots, t_n : s_n$.

[3] The notation s_i for sorts clashes with the notation for case outcomes—we let the context disambiguate.

From the atomic Σ-formulas we now inductively build the full set of Σ-formulas L^Σ as the smallest set such that:

- $L^\Sigma_{\text{at}} \subseteq L^\Sigma$;
- if $\phi \in L^\Sigma$ then $\neg\phi \in L^\Sigma$;
- if $\phi, \psi \in L^\Sigma$ then $\phi \wedge \psi, \phi \vee \psi, \phi \to \psi, \phi \leftrightarrow \psi \in L^\Sigma$;
- if $\phi \in L^\Sigma$ and $x : s$ is a variable then $(\forall x : s)\phi, (\exists x : s)\phi \in L^\Sigma$.

Furthermore, given an indexed set of formulas $\{\phi_i \mid i \in I\} \subseteq L^\Sigma$ for a finite set I we define:

$$\bigwedge_{i\in I} \phi_i = \phi_{i_1} \wedge \cdots \wedge \phi_{i_n}, \qquad \bigvee_{i\in I} \phi_i = \phi_{i_1} \vee \cdots \vee \phi_{i_n}.$$

The formulas in L^Σ can now be interpreted in *structures* assigning meaning to the sorts and symbols. More specifically, a Σ-*structure* $\mathbb{A} = ((A_s)_{s\in S}, I)$ associates to each sort $s \in S$ a set A_s; to each function symbol $f : s_1 \times \ldots \times s_n \to s \in \Sigma$ a function $f_I : A_{s_1} \times \ldots \times A_{s_n} \to A_s$; and to each relation symbol $R : s_1 \times \ldots \times s_n \in \Sigma$ a relation $R_I \subseteq A_{s_1} \times \ldots \times A_{s_n}$. An *assignment* α is an assignment of meaning to the variables, according to their sort. More specifically, an assignment α is a function on X such that $\alpha(x : s) \in A_s$ for $x : s \in X$. Such an assignment can be extended to operate on the set of terms T^Σ by recursively defining $\alpha(f(t_1, \ldots, t_n) : s) = f_I(\alpha(t_1), \ldots, \alpha(t_n))$.

Given a Σ-structure $\mathbb{A}$ and an assignment α we can now define what it means for a formula $\phi \in L^\Sigma$ to be *true* in $\mathbb{A}$. We do so by induction on the complexity of formulas:

$$\begin{aligned}
\mathbb{A}, \alpha &\vDash \bot \text{ is never true,}\\
\mathbb{A}, \alpha &\vDash \top \text{ is always true,}\\
\mathbb{A}, \alpha \vDash t_1 \doteq t_2 &\iff \alpha(t_1) = \alpha(t_2),\\
\mathbb{A}, \alpha \vDash R(t_1, \ldots, t_n) &\iff R_I(\alpha(t_1), \ldots, \alpha(t_n)),\\
\mathbb{A}, \alpha \vDash \neg\phi &\iff \mathbb{A}, \alpha \nvDash \phi,\\
\mathbb{A}, \alpha \vDash \phi \wedge \psi &\iff \mathbb{A}, \alpha \vDash \phi \text{ and } \mathbb{A}, \alpha \vDash \psi,\\
\mathbb{A}, \alpha \vDash \phi \vee \psi &\iff \mathbb{A}, \alpha \vDash \phi \text{ or } \mathbb{A}, \alpha \vDash \psi,\\
\mathbb{A}, \alpha \vDash \phi \to \psi &\iff \text{if } \mathbb{A}, \alpha \vDash \phi \text{ then } \mathbb{A}, \alpha \vDash \psi,\\
\mathbb{A}, \alpha \vDash \phi \leftrightarrow \psi &\iff \mathbb{A}, \alpha \vDash \phi \text{ if and only if } \mathbb{A}, \alpha \vDash \psi.
\end{aligned}$$

Given some set $T \subseteq L^\Sigma$, often called a *theory*, we write $\mathbb{A}, \alpha \vDash T$ if $\mathbb{A}, \alpha \vDash \phi$ for every $\phi \in T$.

A notion of central importance in logic is satisfiability. A formula ϕ is said to be *satisfiable* if there exists a model $\mathbb{A}$ and an assignment α such that $\mathbb{A}, \alpha \vDash \phi$, and *unsatisfiable* otherwise. Computer scientists (as opposed to model theorists) are often interested in a more restricted notion of satisfiability, which fixes the structure $\mathbb{A}$. For this reason, we may also say a formula is $\mathbb{A}$-*satisfiable* if there is some assignment α such that $\mathbb{A}, \alpha \vDash \phi$. Often we are particularly interested in the satisfiability of a formula ϕ relative to some set of background restrictions given by a theory T. To this end, we say a formula is $\mathbb{A}$-satisfiable *modulo a theory* T, denoted $\mathbb{A}, \alpha \vDash_T \phi$, if $\mathbb{A}, \alpha \vDash T \cup \{\phi\}$.

In the remainder of this work, we will be working with satisfiability for some fixed structure $\mathbb{A}$, modulo some background theory T. Therefore, in order not to clutter notation, we will simply speak of satisfiability when we mean $\mathbb{A}$-satisfiability modulo T, and write

$\mathbb{A}, \alpha \vDash \phi$ when we mean $\mathbb{A}, \alpha \vDash_T \phi$. The particular structure $\mathbb{A}$ and background theory T relative to which we are referring to will either be irrelevant or clear from the context.

We conclude this section with some notions closely related to satisfiability. Two formulas $\phi, \psi \in L^\Sigma$ are said to be *equivalent*, denoted $\phi \equiv \psi$, when for all assignments α we have $\mathbb{A}, \alpha \vDash \phi$ if and only if $\mathbb{A}, \alpha \vDash \psi$. A formula ϕ is *valid* if $\mathbb{A}, \alpha \vDash \phi$ for all assignments α. We define the *semantics* $[\![\phi]\!]$ of a formula ϕ as the set of all satisfying assignments α; i.e. $[\![\phi]\!] = \{\alpha \mid \mathbb{A}, \alpha \vDash \phi\}$. These notions are all related, as the following equivalences show:

$$\begin{aligned} \phi \text{ is valid} &\iff \neg\phi \text{ is unsatisfiable} \\ &\iff \phi \equiv \top, \\ \phi \text{ is unsatisfiable} &\iff \phi \equiv \bot, \\ \phi \leftrightarrow \psi \text{ is valid} &\iff \phi \equiv \psi \\ &\iff [\![\phi]\!] = [\![\psi]\!]. \end{aligned}$$

3.10.2 A logical formulation of the a fortiori model

We now show how precedential constraint can be framed in terms of many-sorted logic. For a given set of dimensions D, we take D as the set of sorts and define a signature $\Sigma(D)$:

$$\Sigma(D) = \{c_v \mid v \in d \in D\} \cup \{\sqsubseteq_d \mid d \in D\}.$$

In other words, we introduce for every dimension $d \in D$ the following set of symbols: a constant c_v with $c_v : d$ for every value $v \in d$, and a relation symbol $\sqsubseteq_d$ with $\sqsubseteq_d : d \times d$ for the dimension order of d. For the variables we take a set *Var* with precisely one variable x_d with $x_d : d$ for each dimension d; so $\mathit{Var} = \{x_d \mid d \in D\}$.

We now fix a structure $\mathbb{D} = (D, I)$ for this signature: the domains are given by the set of dimensions D, and the interpretation I simply interprets the symbols according to their intended meaning: $I(c_v) = v$ and $I(\sqsubseteq_d) = \preceq_d$. Due to this fixed interpretation—and to avoid notational clutter—we may henceforth, by abuse of notation, write v where we mean c_v and $\preceq$ where we mean $\sqsubseteq_d$. So for instance, for $v, w \in d$ we will simply write $v \preceq w$ where, strictly speaking, we should write $c_w \sqsubseteq_d c_v$. We will also use the subscript s notation as we did before.

Now, an assignment α of this language is a function on *Var* that maps a variable to a value of the type of that variable, i.e. $\alpha(x_d) \in d$. Since the variables correspond one-to-one with the dimensions, this means that an assignment for this language is essentially the same thing as a fact situation. Therefore, we will henceforth treat the two as interchangeable and use $X, Y, Z, \ldots$ as variables to denote assignments.

Lastly, we need a background theory T relative to which we phrase satisfiability. For instance, T should specify precisely how the elements of the dimensions are related to each other in their respective orders or other axioms related to the dimensions. For instance, for the **Age** dimension $(\mathbb{N}, \geq)$ we need a theory T that includes the theory of the natural numbers, in order to interpret, e.g., the constants that will appear in the formulas. Bradley and Manna (2007, Section 3) and Bjørner and Nachmanson (2020) give many examples of such theories.

Example 3.30. We consider the structure and language for our running example with dimensions $(\textbf{Age}, \preceq_{\textbf{Age}}) = (\mathbb{N}, \geq)$ and $(\textbf{Priors}, \preceq_{\textbf{Priors}}) = (\mathbb{N}, \leq)$. For some fact situation X

we can now form a formula $x_{\mathbf{Age}} \sqsubseteq_{\mathbf{Age}} c_{X(\mathbf{Age})} \wedge c_{X(\mathbf{Priors})} \sqsubseteq_{\mathbf{Priors}} x_{\mathbf{Priors}}$. What does it mean for this formula to be satisfiable? Let Y be any assignment, then

$$\begin{aligned}
&\mathbb{D}, Y \vDash x_{\mathbf{Age}} \sqsubseteq_{\mathbf{Age}} c_{X(\mathbf{Age})} \wedge c_{X(\mathbf{Priors})} \sqsubseteq_{\mathbf{Priors}} x_{\mathbf{Priors}} \\
&\iff \mathbb{D}, Y \vDash x_{\mathbf{Age}} \sqsubseteq_{\mathbf{Age}} c_{X(\mathbf{Age})} \text{ and } \mathbb{D}, Y \vDash c_{X(\mathbf{Priors})} \sqsubseteq_{\mathbf{Priors}} x_{\mathbf{Priors}} \\
&\iff I(\sqsubseteq_{\mathbf{Age}})(Y(x_{\mathbf{Age}}), c_{X(\mathbf{Age})}) \text{ and } I(\sqsubseteq_{\mathbf{Priors}})(c_{X(\mathbf{Priors})}, x_{\mathbf{Priors}}) \\
&\iff Y(x_{\mathbf{Age}}) \geq X(\mathbf{Age}) \text{ and } X(\mathbf{Priors}) \leq Y(x_{\mathbf{Priors}}) \\
&\iff X \preceq Y.
\end{aligned}$$

So, an assignment Y satisfies this formula if and only if $X \preceq Y$. In other words,

$$[\![x_{\mathbf{Age}} \sqsubseteq_{\mathbf{Age}} c_{X(\mathbf{Age})} \wedge c_{X(\mathbf{Priors})} \sqsubseteq_{\mathbf{Priors}} x_{\mathbf{Priors}}]\!] = \uparrow X.$$

Example 3.30 shows that the semantics of formulas in $L^{\Sigma(D)}$ can correspond to subsets of interest. In general, since assignments correspond to fact situations, the semantics function associates each formula $\phi \in L^{\Sigma(D)}$ to some subset $[\![\phi]\!] \subseteq \mathscr{X}$ of satisfying assignments.

Lemma 3.31. *The following equations hold for the semantics function* $[\![-]\!] : L^{\Sigma(D)} \to P(\mathscr{X})$*:*

$$\begin{aligned}
[\![\bot]\!] &= \emptyset, \\
[\![\top]\!] &= \mathscr{X}, \\
[\![v \doteq x_d]\!] &= \{X \in \mathscr{X} \mid v = X(d)\} \\
[\![v \preceq_s x_d]\!] &= \{X \in \mathscr{X} \mid v \preceq_s X(d)\}, \\
[\![\neg\phi]\!] &= \mathscr{X} \setminus [\![\phi]\!], \\
[\![\phi \wedge \psi]\!] &= [\![\phi]\!] \cap [\![\psi]\!], \\
[\![\phi \vee \psi]\!] &= [\![\phi]\!] \cup [\![\psi]\!].
\end{aligned}$$

Proof. By a routine induction on the complexity of formulas in $L^{\Sigma(D)}$. □

Using the semantics function $[\![-]\!]$ we can now generalize Example 3.30 and show that questions related to the a fortiori model can be phrased in terms of formula satisfiability in the structure $\mathbb{D}$. In particular, we will relate the notion of forcing, landmarks, consistency, and completeness to satisfiability in $\mathbb{D}$.

To start, let X be a fact situation; we define a formula $\phi_s(X) \in L^{\Sigma(D)}$ which states that the variable fact situation x is at least as strong for side s as X:

$$\phi_s(X) = \bigwedge_{d \in D} X(d) \preceq_s x_d. \tag{3.10.1}$$

Using the equations of Lemma 3.31 we can now easily derive that $[\![\phi_s(X)]\!] = \uparrow_s X$:

$$\begin{aligned}
[\![\phi_s(X)]\!] &= \Big[\!\Big[\bigwedge_{d \in D} X(d) \preceq_s x_d\Big]\!\Big] \\
&= \bigcap_{d \in D} [\![X(d) \preceq_s x_d]\!] \\
&= \bigcap_{d \in D} \{Y \in \mathscr{X} \mid X(d) \preceq_s Y(d)\} \\
&= \{Y \in \mathscr{X} \mid X \preceq_s Y\} \\
&= \uparrow_s X.
\end{aligned}$$

Next, using $\phi_s(X)$, we define a formula $\Phi_s(\mathscr{C}) \in L^{\Sigma(D)}$ which states that the variable fact situation x has its outcome forced for s by the case base $\mathscr{C}$:

$$\Phi_s(\mathscr{C}) = \bigvee_{X \in \mathscr{C}_s} \phi_s(X). \tag{3.10.2}$$

Using the semantics of $\phi_s(X)$, it is easy to see that $[\![\Phi_s(\mathscr{C})]\!] = \uparrow_s \mathscr{C}_s$:

$$[\![\Phi_s(\mathscr{C})]\!] = \left[\!\!\left[\bigvee_{X \in \mathscr{C}_s} \phi_s(X) \right]\!\!\right] = \bigcup_{X \in \mathscr{C}_s} [\![\phi_s(X)]\!] = \bigcup_{X \in \mathscr{C}_s} \uparrow_s X = \uparrow_s \mathscr{C}_s.$$

Using this we can relate the notion of satisfiability of $\Phi_s(\mathscr{C})$ and the forcing relation induced by the case base $\mathscr{C}, X \vDash s$. This result formally establishes the connection between the a fortiori model and the reformulation we present in many-sorted logic.

Proposition 3.32. *$\mathbb{D}, X \vDash \Phi_s(\mathscr{C})$ if and only if $\mathscr{C}, X \vDash s$.*

Proof. Combining the previous results we get:

$$\mathbb{D}, X \vDash \Phi_s \iff X \in [\![\Phi_s(\mathscr{C})]\!] \iff X \in \uparrow_s \mathscr{C}_s \iff \mathscr{C}, X \vDash s. \qquad \square$$

To make claims about a particular fact situation we need a way to fix the interpretation. There is no symbol in our language for directly equating a fact situation X to the variable fact situation x (i.e. $X \doteq x$ is not a valid formula), but we can define a formula $X \stackrel{\circ}{=} x$ amounting to the same:

$$X \stackrel{\circ}{=} x = \bigwedge_{d \in D} X(d) \doteq x_d. \tag{3.10.3}$$

Again, we can use the equations of Lemma 3.31 to show that this has the intended semantics:

$$\begin{aligned} [\![X \stackrel{\circ}{=} x]\!] &= \left[\!\!\left[\bigwedge_{d \in D} X(d) \doteq x_d \right]\!\!\right] \\ &= \bigcap_{d \in D} [\![X(d) \doteq x_d]\!] \\ &= \bigcap_{d \in D} \{Y \in \mathscr{X} \mid X(d) = Y(d)\} \\ &= \{X\}. \end{aligned}$$

This formula can now be used to make claims relating to a specific fact situation X. For example, for two fact situations $X, Y \in \mathscr{X}$ the question of whether $X \preceq Y$ corresponds to the existence of a satisfying assignment for the formula $x \stackrel{\circ}{=} Y \wedge \phi_s(X)$, as

$$[\![x \stackrel{\circ}{=} Y \wedge \phi_s(X)]\!] = [\![x \stackrel{\circ}{=} Y]\!] \cap [\![\phi_s(X)]\!] = \{Y\} \cap \uparrow_s X.$$

In other words, we have that $x \stackrel{\circ}{=} Y \wedge \phi_s(X)$ is satisfiable if and only if $\{Y\} \cap \uparrow_s X$ is nonempty, which is another way of saying that $X \preceq_s Y$. In a similar way, we can check whether $\mathscr{C}, X \vDash s$ for some X by checking satisfiability of the formula $x \stackrel{\circ}{=} X \wedge \Phi_s(\mathscr{C})$.

Next, we consider how to phrase whether $(X, s) \in \mathscr{C}$ is a landmark case. Let

$$\lambda_s(X) = X \stackrel{\circ}{=} x \wedge \neg \Phi_s(\mathscr{C} \setminus \{(X, s)\}). \tag{3.10.4}$$

This formula states that the variable fact situation x is equal to X, and that $\mathscr{C} \setminus \{(X, s)\}$ does not force x for s. Again we can use the equations of Lemma 3.31 to show that this formula has the intended semantics.

Lemma 3.33. *A case (X, s) is a landmark iff $\lambda_s(X)$ is satisfiable.*

Proof. As in the previous results, we simply apply the equations for the semantics function:

$$\begin{aligned}
[\![\lambda_s(X)]\!] &= \left[\!\!\left[X \stackrel{\circ}{=} x \wedge \neg\Phi_s(\mathscr{C} \setminus \{(X, s)\}) \right]\!\!\right] \\
&= [\![X \stackrel{\circ}{=} x]\!] \cap [\![\neg\Phi_s(\mathscr{C} \setminus \{(X, s)\})]\!] \\
&= \{X\} \cap (\mathscr{X} \setminus [\![\Phi_s(\mathscr{C} \setminus \{(X, s)\})]\!]) \\
&= \{X\} \setminus \uparrow_s(\mathscr{C}_s \setminus \{X\}) \\
&= \begin{cases} \{X\} & \text{if } X \notin \uparrow_s(\mathscr{C}_s \setminus \{X\}), \\ \emptyset & \text{otherwise.} \end{cases}
\end{aligned}$$

□

Remark 3.34. Note that Corollary 3.29 tells us $\Phi_s(\mathscr{C})$ and $\Phi_s(\mathscr{L})$ are logically equivalent:

$$\Phi_s(\mathscr{C}) \equiv \Phi_s(\mathscr{L}) \iff [\![\Phi_s(\mathscr{C})]\!] = [\![\Phi_s(\mathscr{L})]\!] \iff \uparrow_s \mathscr{C}_s = \uparrow_s \mathscr{L}_s.$$

This means we can freely interchange these formulas, which can be computationally advantageous if there are significantly fewer landmarks than regular cases. Of course, this does incur the overhead of computing the set of landmarks $\mathscr{L}$, which may itself be resource intensive. In the remainder of this work we may write Φ_s instead of $\Phi_s(\mathscr{C})$ or $\Phi_s(\mathscr{L})$.

Lastly, we mention that case base consistency and completeness are now easily phrased using the logical language, as the following proposition shows.

Proposition 3.35. *$\mathscr{C}$ consistent iff $\Phi_0 \wedge \Phi_1$ is unsatisfiable, and complete iff $\Phi_0 \vee \Phi_1$ is valid.*

Proof. We apply the semantics function of Lemma 3.31 and then appeal to Lemma 3.26:

$$\begin{aligned}
&\Phi_0 \wedge \Phi_1 \text{ is unsat} \iff \Phi_0 \wedge \Phi_1 \equiv \bot \iff [\![\Phi_0 \wedge \Phi_1]\!] = [\![\bot]\!] \iff \downarrow\mathscr{C}_0 \cap \uparrow\mathscr{C}_1 = \emptyset, \\
&\Phi_0 \vee \Phi_1 \text{ is valid} \iff \Phi_0 \vee \Phi_1 \equiv \top \iff [\![\Phi_0 \vee \Phi_1]\!] = [\![\top]\!] \iff \downarrow\mathscr{C}_0 \cup \uparrow\mathscr{C}_1 = \mathscr{X}.
\end{aligned}$$

□

3.10.3 A case base as a binary classifier

As the last of our theoretical considerations we investigate the relation between a case base and the concept of a classifier from machine learning; i.e. an algorithm that sorts a set of input data into one or more classes. A case base, together with the notion of forcing of Definition 3.5, can be considered as a classifier that can assign 0 or 1 to a new fact situation. This is also the view adopted in the work by Liu et al. (2022) and Odekerken et al. (2023b). In fact, the a fortiori model has been implemented in a human-in-the-loop decision support system for web shop classification at the Dutch National Police Force (Odekerken & Bex, 2020). It is therefore of interest to further examine the theoretical relation between the a fortiori model and binary classifiers in general.

Formally a binary classifier on a set A is a function $f : A \to \{0, 1\}$. The set A contains the input data, and each element $a \in A$ is assigned a label $f(a)$ which is either 0 or 1. Set-theoretically speaking, a function $f : A \to B$ with domain A and codomain B is a set of ordered pairs $\{(a, b) \in A \times B \mid f(a) = b\}$. In other words, f is a relation $f \subseteq A \times B$. However, not every relation between A and B is a function. In order for a relation $R \subseteq A \times B$ to qualify as a function it should satisfy the following criteria.

Definition 3.36. A relation $R \subseteq A \times B$ between sets A and B is *well-defined* if $R(a,b)$ and $R(a,b')$ implies $b = b'$, and *total* if for every $a \in A$ there is some $b \in B$ such that $R(a,b)$. When R is both well-defined and total we say it is *functional*, and write $R : A \to B$.

A relation $R \subseteq A \times B$ is functional if it associates each element in A to precisely one element of B. Given a case base $\mathscr{C}$ we define a relation $c \subseteq \mathscr{X} \times \{0,1\}$ by $c = \{(X,s) \mid \mathscr{C}, X \vDash s\}$, so c is the forcing relation between facts and sides for a given case base $\mathscr{C}$. The question now is under what conditions c is a function $c : \mathscr{X} \to \{0,1\}$, i.e. when is c a binary classifier? Spelling out the condition of being well-defined of Definition 3.36 for the relation c, we have that c is well-defined iff for a fact situation X, and outcomes s and t, we have that $\mathscr{C}, X \vDash s$ and $\mathscr{C}, X \vDash t$ implies $s = t$. In other words, c is well-defined exactly when the case base is consistent. Similarly, to say that c is total is just to say that $\mathscr{C}$ is complete.

We have discussed several equivalent formulations of case base consistency and completeness, corresponding to the different views of the a fortiori models discussed in the preceding sections, and we summarize them in the following proposition.

Proposition 3.37. *The following are equivalent statements about consistency of $\mathscr{C}$:*

(1) *$\mathscr{C}$ is consistent;*

(2) *There is no fact situation X such that $\mathscr{C}, X \vDash 0$ and $\mathscr{C}, X \vDash 1$;*

(3) $\downarrow\mathscr{C}_0 \cap \uparrow\mathscr{C}_1 = \emptyset$*;*

(4) $\Phi_0 \wedge \Phi_1$ *is unsatisfiable;*

(5) *The classify relation c is well-defined.*

Dually, we have the following list of equivalent statements expressing completeness of $\mathscr{C}$:

(1) *$\mathscr{C}$ is complete;*

(2) *For every fact situation X either $\mathscr{C}, X \vDash 0$ or $\mathscr{C}, X \vDash 1$;*

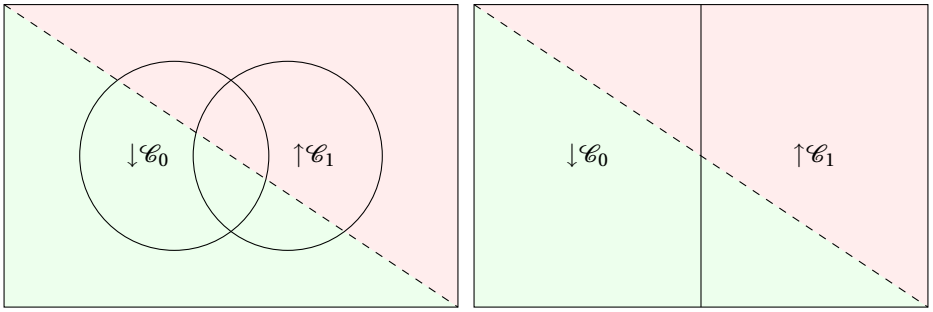
(3) $\downarrow\mathscr{C}_0 \cup \uparrow\mathscr{C}_1 = \mathscr{X}$*;*

(4) $\Phi_0 \vee \Phi_1$ *is valid;*

(5) *The classify relation c is total.*

When considering classifiers, one is often interested in classification accuracy. When the set of fact situations $\mathscr{X}$ comes with ground truth labels, it is partitioned by two sets $\mathscr{X}_0 \cup \mathscr{X}_1 = \mathscr{X}$ indicating these labels. In such a scenario, we can consider the degree to which the labels forced by the case base are in agreement with these ground truth labels. We can visualize this comparison by modifying the Euler diagram representation that we gave in Figure 3.2. We do this by indicating the subset $\mathscr{X}_0 \subseteq \mathscr{X}$ as a green shaded area, and the subset $\mathscr{X}_1 \subseteq \mathscr{X}$ as a red shaded area, divided by a dashed line; see Figure 3.3.

The Euler diagram corresponding to the general case, when the case base is neither consistent nor complete, is depicted by Figure 3.3a. If, however, the case base is a proper classifier (meaning it is consistent and complete), the picture looks as in Figure 3.3b. We can think of this Euler diagram as a confusion matrix: $\downarrow\mathscr{C}_0 \cap \mathscr{X}_0$ contains the *true negative* fact situations, $\downarrow\mathscr{C}_0 \cap \mathscr{X}_1$ the *false negative* fact situations, $\uparrow\mathscr{C}_1 \cap \mathscr{X}_1$ the *true positive* fact situations, and $\uparrow\mathscr{C}_1 \cap \mathscr{X}_0$ the *false positive* fact situations. We will return to this representation for our data analysis in Chapter 7.



(a) Neither consistent nor complete **(b)** Both consistent and complete

Figure 3.3: An adaptation of the Euler diagrams in Figure 3.2 for when the space of fact situations $\mathscr{X}$ is partitioned by $\mathscr{X}_0$, the green shaded area, and $\mathscr{X}_1$, the red shaded area, indicating ground truth labels 0 and 1, respectively.

3.11 Conclusion: Moving to hierarchical structures

In this chapter we considered the DRM: an extension of the RM that can handle dimensional data. We recalled its definitions as proposed by Horty (2019), and adapted them to partial fact situations. We then showed that the RM can be seen as a special case of the DRM. Lastly, as a step towards our data analysis in Part II of this thesis, we described the DRM in terms of order theory and many-sorted logic.

A second extension of the RM, which we refer to as the hierarchical result model (HRM), has been proposed in the literature. It expands on the knowledge representation of the RM by assuming that the factors form a hierarchy. In this context, factors do not directly favor a decision for a decision for the plaintiff or the defendant, but provide support to each other in a hierarchical fashion. Recognizing this hierarchical structure—as was first done by the developers of the CATO program (Aleven & Ashley, 1997)—raises the question how the notion of constraint should be altered. We address this, and related questions, in the next chapter.

Chapter 4

Modeling Factor-Based Hierarchical Constraint

THE GOAL OF THIS CHAPTER is to develop a version of the RM that accommodates additional hierarchical structure on the set of factors, which we do based on ideas in earlier work on formal models of precedential constraint by Roth (2003) and Roth and Verheij (2004).[1] We begin in Section 4.1 by discussing some examples of factor hierarchies, one of which is based on the running example introduced in Section 2.1, and then formalize the notion of factor hierarchy as a knowledge representation framework in Section 4.2. We then discuss some possible approaches to adapting the notion of constraint from the RM to this extended representation in Section 4.3, and subsequently present our variant—the resulting model is referred to as the "hierarchical result model" (HRM). We contrast our definition of constraint with that found in the literature through some examples, and then spend the remainder of this chapter analyzing formal properties of the HRM. In Section 4.6 we show that the HRM is monotonic in the addition of cases and in the addition of information to a focus fact situation. Then, in Section 4.7, we show that the RM is a special case of the HRM, or in other words, that the HRM is a conservative extension of the RM. Finally, we end with some concluding remarks in Section 4.8.

4.1 Examples of factor hierarchies

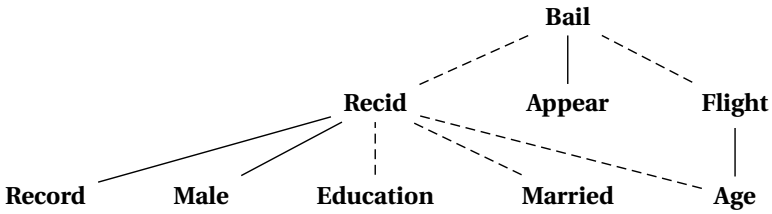
We continue our running example on the domain of criminal sentencing. Recall that in Section 2.1 we discussed some factors influencing a recidivism risk assessment. A downstream purpose of such an assessment is to determine whether a defendant should be released on bail. Bail is a sum of money that a defendant must pay to the court as a guarantee that they will appear at their trial—if the defendant does not appear, the bail is forfeited. The decision to grant bail, like recidivism risk, is influenced by several factors;

[1] All of the material in this chapter, with the exception of some of the examples, stems from van Woerkom et al. (2023a, 2025).

e.g. a defendant with a high risk of flight is less likely to be granted bail, while one with a history of appearing to court is more likely to be granted bail.

In other words, determining bail is a domain to which the result model can be applied, but this time one of the input factors—risk of recidivism—can itself be determined on the basis of a fortiori reasoning. This situation is described as a *factor hierarchy* in the AI & law literature. The hierarchical structure of factors was first recognized by Aleven (1997, 2003) in his work on the CATO program, and has since become a standard component of work in AI & law involving factor-based knowledge representations; see e.g. the work by Bruninghaus and Ashley (2003), Grabmair (2017), Prakken and Sartor (1998), and Roth (2003), among others.

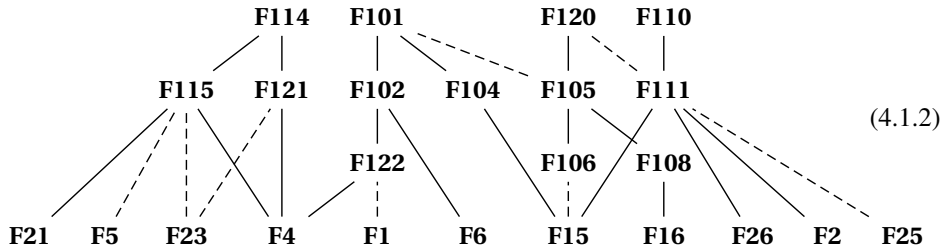
We expand the set (2.1.1) from Section 2.1 to a factor hierarchy for a bail decision:



(4.1.1)

The **Bail** node corresponds to a decision to grant bail. The links are either solid or dotted, which carries the same meaning as it did in our earlier example; for instance, the dotted line from **Recid** to **Bail** indicates that high risk of recidivism suggests bail should be denied. Two other factors are added: **Appear**, which represents that the defendant has a history of always appearing for trial, and **Flight**, which stands for a low or high risk of fleeing. This means that in this example we assume that older people are more likely to flee before trial than younger people. Whether this assumption holds in practice is debatable—it is added primarily to exemplify that factors can influence multiple higher level factors. The vertical order in the graph represents the level of abstraction: the lower the position of the factor, the less abstract it is. Lastly, we stress that the hierarchy is not intended to be an exhaustive representation of the real-world factors influencing a bail decision.

Aside from our running example we mention the prototypical example of a factor hierarchy: the hierarchy developed by Aleven (1997) on the domain of trade secrets, used by his CATO program. A fragment is depicted below (Aleven & Ashley, 1997, Figure 4):



(4.1.2)

Aleven maintains the following terminology regarding this hierarchy. The factors at the bottom of the hierarchy are at the lowest level of abstraction and are called "base-level factors." For example, the factor **F1** represents that the "plaintiff disclosed its

product information in negotiations with the defendant." The base-level factors link to the "Intermediate Legal Concerns," which are more abstract factors representing normative legal knowledge of the domain. For example, the aforementioned **F1** negatively influences the intermediate legal concern **F122**, which states that the "plaintiff took efforts to maintain the secrecy of its information in its dealings with the defendant." In turn, this factor positively influences the more general intermediate legal concern **F102**, which states that the "plaintiff took efforts to maintain the secrecy of its information." Finally, at the top of the hierarchy are the "Legal Issues," which correspond to the main issues that courts use to explain their decisions (Aleven, 1997, Section 3.2.2). For example, the intermediate legal concern **F102** positively affects the legal issue **F101**, which states that the "plaintiff's information is a trade secret." See Aleven (1997, Appendices 1 and 2) for a complete overview of the factors in the CATO hierarchy and their meaning.

We now proceed to introduce the general framework of knowledge representation that we will use to formally represent factor hierarchies, and for which we will subsequently define a notion of constraint.

4.2 Knowledge representation

We now expand the knowledge representation used by the RM, discussed in Section 2.2, to one incorporating hierarchical structure on the set F of factors. Horty (2011, Section 7) discusses a possible extension of his result and reason models of precedential constraint to account for hierarchical structure. He remarks that in such an extension, the outcomes π and δ can be taken as special factors, and a precedent case as "a linked set of precedent constituents, beginning with a characterization of the initial fact situation as a set of base-level factors, proceeding through a series of higher-level legal concepts, and eventually arriving at a decision [...] for the plaintiff or defendant." This is very similar to the approach taken by Roth (2003), who represents cases as sets of factors and links between them (indicating the hierarchical structure), and in which the case outcome is considered to be just one of the factors.

We will use a slightly different approach in this work. Firstly, we will not assume that the set of factors necessarily contains a factor corresponding directly to the case outcome. This is to accommodate the point of view, as recently advocated by e.g. Bench-Capon and Atkinson (2021), that precedent cases serve to determine the applicability of legal *issues*—the legal questions that need to be answered in order to determine the outcome of the case. Proponents of this view argue that case outcomes are determined as a logical function of the legal issues of the domain, and not directly on the basis of precedent; rather, it is the applicability of the issues that is determined by precedent. In our approach, we assume that the factor hierarchy of the domain culminates in one or more factors, which may either correspond to the issues of the domain, or (if there is just one) directly to the case outcome. For the sake of discussion we will henceforth refer to the top-level elements of the hierarchy simply as issues, even in the case where there is a single top-level element corresponding directly to the case outcome.

Secondly, we will not consider the hierarchical structure to be part of individual cases—as in the representation used by Roth—but rather as a separate structure which the set of factors is endowed with. More specifically, we model this structure using a relation

on the set F of factors.

Definition 4.1. A *factor hierarchy* $(F, \mathsf{Pro}, \mathsf{Con})$ consists of a finite set of factors F with two binary relations Pro and Con on F, such that $\mathsf{Pro} \cap \mathsf{Con} = \emptyset$, and such that the transitive closure of their union $\mathsf{H} = \mathsf{Pro} \cup \mathsf{Con}$ is irreflexive.

In other words, a factor hierarchy consists of a set of factors as in the RM, which is given hierarchical structure by a relation H; given factors $p, q \in F$ the relation $\mathsf{H}(p, q)$ means p is directly below q in the hierarchy. The hierarchical structure may not contain loops, and any link $\mathsf{H}(p, q)$ represents either that the presence of p lends defeasible support to the presence of q—if $\mathsf{Pro}(p, q)$—or that it lends defeasible opposition to the presence of q—if $\mathsf{Con}(p, q)$—but never both, as $\mathsf{Pro} \cap \mathsf{Con} = \emptyset$.

All the usual terminology employed in the literature on factor hierarchies can now be expressed in terms of the structure $(F, \mathsf{Pro}, \mathsf{Con})$. An H-minimal factor is called *base-level*, and we write B for the set of all base-level factors in the hierarchy. A factor that is not base-level is called *abstract*, and the set of abstract factors is denoted by A, and so F is partitioned by $F = A \cup B$. Factors are assumed to support or oppose each other in hierarchical fashion, as indicated by the relations Pro and Con. When $\mathsf{Pro}(p, q)$ holds we say p is a pro-q factor, and when $\mathsf{Con}(p, q)$ we say p is a con-q factor. Note that any subset $G \subseteq F$ of the set of factors, together with H restricted to G, yields a factor hierarchy $(G, \mathsf{Pro}, \mathsf{Con})$.

In the RM, a fact situation is made a case by pairing it with an outcome. In Definition 4.1 of a factor hierarchy, the distinction between factors and issues, or case outcomes, is dropped. This means that if we define a "fact situation" X as a valuation $X : F \to \{\mathbf{f}, \mathbf{t}\}$ then X will necessarily assign a truth value to the issues of the domain. This definition is therefore more akin to the notion of a case than to the notion of a fact situation. As such, in order to be able to speak of a fact situation, we should allow these valuations to be partial in the sense introduced in the preceding chapters (see Sections 2.4 and 3.4).

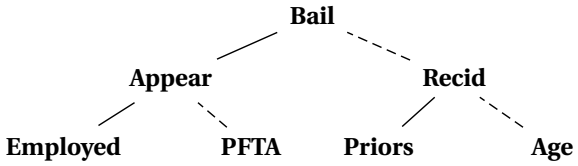
More specifically, a fact situation X is a valuation of a subset of F, i.e., a partial fact situation as before, and again we write $\mathrm{dom}(X)$ for the subset of F on which X is defined. We employ the same notation with regards to the truth values of the factors in X, so for example $X \vDash p$ means $X(p) = \mathbf{t}$. An important conceptual difference with the RM is that p may now be an abstract factor, such as an issue. These do not receive their truth values "as a matter of fact," but are assigned a truth value by a decision-maker as a result of weighing the pro- and con-p factors. In other words, if $X \vDash p$ then this means it was *decided* that p applies in X—possibly on the basis of precedential reasoning.

Example 4.2. Consider the factor hierarchy depicted in (4.1.1). We can extend the case $(Y, 1)$ from Example 3.4 with $Y \vDash \{\textbf{Record}, \neg\textbf{Male}, \neg\textbf{Education}, \textbf{Married}, \neg\textbf{Age}\}$ to a fact situation for this domain. The outcome 1 of the case $(Y, 1)$, which represents a judgment that Y poses a recidivism risk, now means that Y assigns true to the **Recid** factor, so $Y \vDash \textbf{Recid}$. Suppose, furthermore, that $Y \vDash \{\textbf{Appear}, \neg\textbf{Flight}, \textbf{Bail}\}$, which means that Y has a history of appearing to court, is judged to be a low flight risk, and (thus) was granted bail. Notice the difference between the base-level factors, which correspond to the "plain facts" of the situation, and the abstract factors, which correspond to judgments by a decision-maker. Whether Y has had a high school education is a fact about Y which is pre-determined, while the question of whether Y poses a flight risk is something to be decided presently, by weighing the pro- and con-**Flight** factors in the hierarchy.

In this knowledge representation we also allow cases to be partial. The reason for this is twofold. Firstly, pragmatically, we want to allow the possibility for cases to be undefined on factors because in practice a court may not make a decision about every factor of the domain. Secondly, formally, no further requirements on cases are needed for our definition of constraint (cf. Section 4.3). This generality allows users of the HRM to further augment the definition of a case, if needed. For example, in practice it could be natural to require that a case—by definition—is decided on at least one of the issues of the factor hierarchy. Furthermore, as a second example of a natural requirement of cases, we could ask that for each abstract factor p which is assigned true in the case, at least one pro-p factor q is also assigned true in the case. However, we stress once more that neither of these requirements is necessary for defining a notion of constraint in our framework, and so we omit them.

4.3 Approaches to hierarchical constraint

The introduction of the hierarchical structure poses the question how it should influence the notion of constraint, and we will now consider some approaches. As an example to guide this discussion we will consider the following modification of the hierarchy depicted in (4.1.1). The **PFTA** factor corresponds to whether the defendant has previously failed to appear to court.



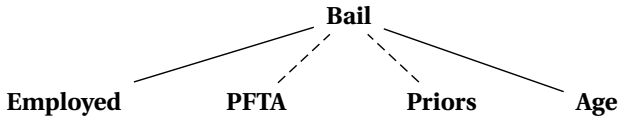
(4.3.1)

The first question we consider is whether constraint should work only on the top level factors in the hierarchy, or also on the abstract factors. For example, in the hierarchy depicted by (4.3.1), should previous decisions about the recidivism risk factor **Recid** constrain future decisions about this factor? Or should constraint be limited to function only on the ultimate decision about granting bail, represented by the **Bail** factor? Bench-Capon (2023) argues that the abstract factors appearing in factor hierarchies in the literature do not form part of the law in the domain they are describing; therefore, there is no suggestion that judges would acknowledge them. In this view, decisions about abstract factors should not induce constraint, at least in the legal setting. In related work on hierarchical constraint, Canavotto and Horty (2023a) argue in response to Bench-Capon's objections that formal work on hierarchical constraint describes normative reasoning more generally, and have wider potential applicability than just to the domain of law. Indeed, the notion of hierarchical constraint put forth by Canavotto and Horty (2023b, Definition 9) functions on the abstract factors, in addition to the issues of a factor hierarchy. We agree with this more general view, and will define constraint to operate on abstract factors.

The second question we consider is whether the status of the abstract factors of a precedent case should play a role in the constraint that is induced, or whether it should be only the status of the base-level factors that influence whether constraint is applied. The definition of constraint developed by Roth (2003, Section 3.6.6) takes this second stance, and a similar form of constraint is studied by Canavotto and Horty (2023b, Section 5).

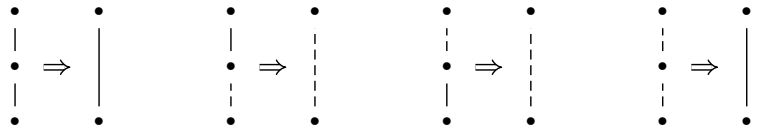
Following terminology of Canavotto and Horty (2023a, 2023b), we will refer to this as *flat* constraint. This involves successively adding connections to the hierarchy from lower level factors to higher level ones, in order to ultimately be able to say whether the base-level factors support or oppose abstract factors that they are not directly connected to in the hierarchy. This then gives rise to an instance of the regular RM, so that its notion of constraint can be applied as before.

To exemplify these approaches we consider the factor hierarchy depicted in (4.3.1). Since the base-level factors in this hierarchy do not directly connect to the bail decision issue, we can not apply the RM to constrain bail decisions on the basis of truth values of the base-level factors. However, there are (by definition) paths in the hierarchy from the base-level factors leading up to the issues (in this case the bail factor) through the intermediate abstract factors. The idea of flat constraint is to successively add missing links on the basis of the polarity of these intermediate links. For instance, in (4.3.1), the base-level factors do not have links directly to the **Bail** issue, but do have links to the intermediate abstract factors **Appear** and **Recid**. Since **Employed** supports the abstract factor **Appear**, which in turn supports the **Bail** issue, we can see that the presence of the **Employed** factor also supports a decision to grant bail. Therefore, we could extend the hierarchy to include a direct link of support from the **Employed** factor to the **Bail** factor. Similarly, the **Priors** factor is a pro-**Recid** factor, which is itself a con-**Bail** factor. So, we can conclude that prior offenses indirectly suggest bail should not be granted, and this can be encoded as a negative link from **Priors** to **Bail**. In the end, we obtain:



(4.3.2)

This method of successively adding links from the base-level factors to the issues can be described by the following four patterns:



(4.3.3)

When the factor hierarchy is more than two levels high, such as the one in (4.1.2), the newly added links can then be recursively subjected to the same rules until only direct links between base-level factors and the issues remain. The rules in (4.3.3) correspond to the clauses of Definition 2 of Canavotto and Horty (2023b) when translated to our setting. Roth's (2003) Definition 6 works similarly, but only includes the two patterns on the left.

This flattened hierarchy in (4.3.2) constitutes a factor partition in the sense considered in Section 2.2, and so the notion of constraint of the RM can be applied to it. We thus see that, in the resulting notion of constraint for factor hierarchies, the status of the intermediate factors in the constraining precedent case becomes inconsequential.

An alternative to flat constraint is what is called *hierarchical* constraint by Canavotto and Horty (2023a, 2023b). In this approach, the constraint is induced directly on the basis of the status of the abstract factors in the precedent case. The idea underlying this approach

is that it is the status of the factors that is consequential (with respect to connected higher level factors), rather than the reason for their status (resulting from weighing connecting lower level factors). For example, looking again at (4.3.1), this approach argues that when it comes to a decision about granting bail, what matters is whether someone poses a recidivism risk or not. Once it has been established that someone poses a recidivism risk, it becomes irrelevant whether this was because they were young, or because they had many prior offenses.

In the subsequent section we implement hierarchical constraint by use of recursion. Each abstract factor p in the hierarchy, together with its direct subordinate pro- and con-p factors, is taken as an instance of the regular RM. Constraint is then applied recursively, each step involving the citation of a precedent case. This allows for the notion of constraint to involve several precedent cases, successively building up constraint on the more abstract levels of the hierarchy. To illustrate, we again consider the factor hierarchy depicted in (4.3.1). Suppose there is a precedent case X in which it was determined that a defendant over the age of 21 with prior convictions was deemed at high risk of recidivism. In other words, the **Priors** factor outweighs the **Age** factor when it comes to determining risk of recidivism. Furthermore, suppose there is a second precedent case Y in which it was determined that a high risk of recidivism is reason enough to deny bail, even when the defendant has a high chance of appearing to court. This means it was determined that the **Recid** factor outweighs the **Appear** factor. In the presence of these two precedent cases, any fact situation containing the **Priors** factors should therefore be forced for the outcome of having bail denied: first X constrains the decision for a high recidivism risk judgment, and then, on the basis of this high recidivism risk, Y constrains a decision for bail denial.

To define constraint we write $\mathsf{Pro}(p)$ for the pre-image of p under the relation Pro, so $\mathsf{Pro}(p) = \{q \in F \mid \mathsf{Pro}(q, p)\}$, and similarly $\mathsf{Con}(p)$ for its pre-image under Con. We now define two relations $\mathscr{C}, X \vDash p$ and $\mathscr{C}, X \vDash \neg p$ of constraint for the HRM by mutual recursion.

Definition 4.3. Let $p \in F$, X a fact situation and $\mathscr{C}$ a case base, then *the decision of X for p is forced by $\mathscr{C}$*, denoted $\mathscr{C}, X \vDash p$, if and only if either

- $X \vDash p$, or
- $p \in A$ and there is a case $Y \in \mathscr{C}$ with $Y \vDash p$ and
  - for all $q \in \mathsf{Pro}(p)$: if $Y \vDash q$ then $\mathscr{C}, X \vDash q$, and
  - for all $q \in \mathsf{Con}(p)$: if $Y \vDash \neg q$ then $\mathscr{C}, X \vDash \neg q$.

Likewise, *the decision of X for $\neg p$ is forced by $\mathscr{C}$*, denoted $\mathscr{C}, X \vDash \neg p$ iff either

- $X \vDash \neg p$, or
- $p \in A$ and there is a case $Y \in \mathscr{C}$ with $Y \vDash \neg p$ and
  - for all $q \in \mathsf{Pro}(p)$: if $Y \vDash \neg q$ then $\mathscr{C}, X \vDash \neg q$, and
  - for all $q \in \mathsf{Con}(p)$: if $Y \vDash q$ then $\mathscr{C}, X \vDash q$.

Remark 4.4. Definition 4.3 of constraint differs from that proposed by van Woerkom et al. (2023a), who use the implication "if $\mathscr{C}, X \vDash q$ then $Y \vDash q$" for the con-p factors q, whereas Definition 4.3 uses the implication "if $Y \vDash \neg q$ then $\mathscr{C}, X \vDash \neg q$." This is entirely analogous to the differences discussed in Section 2.4 and corresponds to the two options (1) and (2) discussed in Remark 2.4.

Remark 4.5. Note that, according to this definition of constraint, in order for a fact situation X to induce constraint on an abstract factor p we should have $p \in \mathrm{dom}(X)$, meaning X should be defined on p. Therefore, in order to serve as a precedent in the usual sense of the word, a fact situation X should be defined on at least one abstract factor.

We note two simple consequences of this definition. Firstly, any factor $p \in F$ that is deemed to hold in a fact situation X, so $X \models p$, continues to hold in the presence of a case base $\mathscr{C}$.

Lemma 4.6. *If $X \models p$ then $\mathscr{C}, X \models p$. Likewise, $X \models \neg p$ implies $\mathscr{C}, X \models \neg p$.*

For base-level factors—i.e. those in $B = F \setminus A$—the converse of Lemma 4.6 holds too.

Lemma 4.7. *If $p \in B$ then $\mathscr{C}, X \models p$ implies $X \models p$, and $\mathscr{C}, X \models \neg p$ implies $X \models \neg p$.*

Proof. As $p \in B$ we have $p \notin A$, and so spelling out Definition 4.3 gives:

$\mathscr{C}, X \models p$
iff • $X \models p$, or
  • $p \in A$ and there is $Y \in \mathscr{C}$ with $Y \models p$ and
    – for all $q \in \mathsf{Pro}(p)$: if $W \models q$ then $\mathscr{C}, X \models q$, and
    – for all $q \in \mathsf{Con}(p)$: if $W \models \neg q$ then $\mathscr{C}, X \models \neg q$
iff $X \models p$. □

Example 4.8. To illustrate Definition 4.3 we work through an example based on the factor hierarchy depicted in (4.1.1), which is formally specified by $(F, \mathsf{Pro}, \mathsf{Con})$ where

$$\begin{aligned} F &= \{\textbf{Bail}, \textbf{Recid}, \textbf{Appear}, \textbf{Flight}, \textbf{Record}, \textbf{Male}, \textbf{Education}, \textbf{Married}, \textbf{Age}\}, \\ \mathsf{Pro} &= \{(\textbf{Record}, \textbf{Recid}), (\textbf{Male}, \textbf{Recid}), (\textbf{Appear}, \textbf{Bail}), (\textbf{Age}, \textbf{Flight})\}, \\ \mathsf{Con} &= \{(\textbf{Education}, \textbf{Recid}), (\textbf{Married}, \textbf{Recid}), (\textbf{Age}, \textbf{Recid}), (\textbf{Recid}, \textbf{Bail}), \\ &\quad (\textbf{Flight}, \textbf{Bail})\}. \end{aligned}$$

Let X and Y be two fact situations for this hierarchy, defined by

$$\begin{aligned} X &\models \{\textbf{?Bail}, \textbf{?Recid}, \textbf{Appear}, \textbf{Flight}, \textbf{Record}, \neg\textbf{Male}, \textbf{Education}, \neg\textbf{Married}, \textbf{Age}\}, \\ Y &\models \{\textbf{Bail}, \neg\textbf{Recid}, \textbf{Appear}, \textbf{Flight}, \neg\textbf{Record}, \textbf{Male}, \textbf{Education}, \textbf{Married}, \neg\textbf{Age}\}. \end{aligned}$$

We consider whether the decision to grant bail to the defendant in fact situation Y induces constraint for the decision to grant bail to the defendant in the focus fact situation X:

$\{Y\}, X \models \textbf{Bail}$
iff • $X \models \textbf{Bail}$, or
  • $\textbf{Bail} \in A$ and there is $W \in \{Y\}$ with $W \models \textbf{Bail}$ and
    – for all $q \in \mathsf{Pro}(\textbf{Bail})$: if $W \models q$ then $\{Y\}, X \models q$, and
    – for all $q \in \mathsf{Con}(\textbf{Bail})$: if $W \models \neg q$ then $\{Y\}, X \models \neg q$
iff $\{Y\}, X \models \{\textbf{Appear}, \neg\textbf{Recid}\}$.

As $X \models$ **Appear**, and so $\{Y\}, X \models$ **Appear**, this simplifies to $\{Y\}, X \models \neg$**Recid**, but

$\{Y\}, X \models \neg$**Recid**
iff • $X \models \neg$**Recid**, or
  • **Recid** $\in A$ and there is $W \in \{Y\}$ with $W \models \neg$**Recid** and
    – for all $q \in \mathsf{Pro}$(**Recid**): if $W \models \neg q$ then $\{Y\}, X \models \neg q$, and
    – for all $q \in \mathsf{Con}$(**Recid**): if $W \models q$ then $\{Y\}, X \models q$
iff $\{Y\}, X \models \{\neg$**Record**, **Married**, **Education**$\}$,
iff $X \models \{\neg$**Record**, **Married**, **Education**$\}$.

The last equivalence follows from Lemma 4.7. By definition $X \models$ **Record** and so $X \not\models \neg$**Record** and thus $\{Y\}, X \not\models$ **Bail**, meaning there is no constraint induced by Y.

The case Y does not constrain X in this example, because in Y it was decided that the defendant was not at risk of recidivism while X is undecided on this factor. Furthermore, Y does not force a low risk of recidivism assessment in X. However, due to the recursive call in Definition 4.3, a second precedent case may be used to force a decision in X for $\neg$**Recid**. To illustrate this, we now supplement the case base $\{Y\}$ with a case Z such that

$$Z \models \{\neg\textbf{Recid}, \textbf{Record}, \neg\textbf{Male}, \textbf{Education}, \neg\textbf{Married}, \neg\textbf{Age}\}.$$

This new case base $\{Y, Z\}$ does constrain X on the recidivism factor:

$\{Y, Z\}, X \models \neg$**Recid**
iff • $X \models \neg$**Recid**, or
  • **Recid** $\in A$ and there is $W \in \{Y, Z\}$ with $W \models \neg$**Recid** and
    – for all $q \in \mathsf{Pro}$(**Recid**): if $W \models \neg q$ then $\{Y, Z\}, X \models \neg q$, and
    – for all $q \in \mathsf{Con}$(**Recid**): if $W \models q$ then $\{Y, Z\}, X \models q$
iff $\{Y, Z\}, X \models \{\neg$**Male**, **Education**$\}$,
iff $X \models \{\neg$**Male**, **Education**$\}$.

We thus find that $\{Y, Z\}, X \models \neg$**Recid**. So, as $X \models$ **Appear** we have $\{Y, Z\}, X \models$ **Appear** by Lemma 4.6 and similarly $\{Y, Z\}, X \models$ **Flight** which means that $\{Y, Z\}, X \models$ **Bail** by the earlier line of reasoning.

Note that we have not specified the values of Z on the other factors influencing the decision to grant bail. This was done intentionally—to demonstrate that the values of Z on those factors is irrelevant with respect to its role in constraining the decision of X for $\neg$**Recid**. In fact, the line of reasoning of this example would still work just as well if $Z \models \neg$**Bail**, which means that a case can contribute to forcing an outcome which it itself was not decided for: we would have both $Z \models \neg$**Bail** and $\{Y, Z\}, X \models$ **Bail** (while $\{Y\}, X \not\models$ **Bail**).

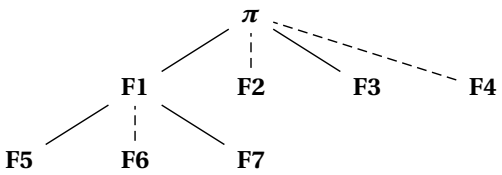
Lastly, this example has served to demonstrate that more than one precedent case can be involved in forcing an outcome in a focus fact situation—a feature which results from our use of recursion in Definition 4.3. In contrast, the flat version of constraint, which we discussed in Section 4.3, necessarily involves only a single precedent case.

Table 4.1: A description of some factors influencing a decision of whether a dismissal can be voided, as used by Roth and Verheij (2004). The factors are based on the Dutch Civil Code. See (4.4.1) for a representation of the hierarchical structure between these factors.

Factor	Description
$\boldsymbol{\pi}$	The dismissal can be voided.
F1	The dismissed person has always behaved like a good employee.
F2	The dismissed person committed a serious act of violence.
F3	The working atmosphere has not been affected by the dismissal.
F4	The employee has a criminal record.
F5	The dismissed person always arrived on time for work.
F6	The dismissed person once insulted a superior.
F7	The dismissed person was always dressed properly.

4.4 Discussion of related research

Example 4.9. Next we compare our notion of constraint to that used by Roth and Verheij (2004, Section 3), by examining their example on the domain of Dutch dismissal law. The associated hierarchy is shown below, and a description of the factors is given in Table 4.1:



(4.4.1)

The case outcome corresponds to whether or not the dismissal can be voided, and relevant factors are considered such as whether the dismissed person has always behaved like a good employee (**F1**). We consider a precedent Y and a focus fact situation X satisfying $Y \models \{\boldsymbol{\pi}, \mathbf{F2}, \neg\mathbf{F3}, \mathbf{F4}, \mathbf{F5}, \mathbf{F6}, \neg\mathbf{F7}\}$ and $X \models \{?\boldsymbol{\pi}, ?\mathbf{F1}, \mathbf{F2}, \mathbf{F3}, \neg\mathbf{F4}, \mathbf{F5}, \neg\mathbf{F6}, \mathbf{F7}\}$. Unfolding Definition 4.3 we see that $\{Y\}, X \models \boldsymbol{\pi}$ if and only if $Y \models \mathbf{F1}$:

$\{Y\}, X \models \boldsymbol{\pi}$ iff • $X \models \boldsymbol{\pi}$, or
- • $\boldsymbol{\pi} \in A$ and there is $Z \in \{Y\}$ with $Z \models \boldsymbol{\pi}$ and
  - for all $p \in \mathsf{Pro}(\boldsymbol{\pi})$: if $Z \models p$ then $\{Y\}, X \models p$, and
  - for all $p \in \mathsf{Con}(\boldsymbol{\pi})$: if $Z \models \neg p$ then $\{Y\}, X \models \neg p$

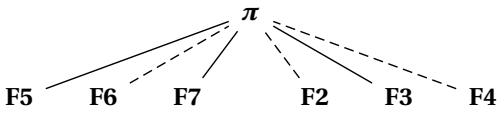
iff $\{Y\}, X \models \mathbf{F1}$

iff • $X \models \mathbf{F1}$, or
- • $\mathbf{F1} \in A$ and there is $Z \in \{Y\}$ with $Z \models \mathbf{F1}$ and
  - for all $p \in \mathsf{Pro}(\mathbf{F1})$: if $Z \models p$ then $\{Y\}, X \models p$, and
  - for all $p \in \mathsf{Con}(\mathbf{F1})$: if $Z \models \neg p$ then $\{Y\}, X \models \neg p$

iff $Y \models \mathbf{F1}$.

In other words, whether **F1** applies in Y or not entirely determines whether Y can constrain the decision of X for $\boldsymbol{\pi}$.

We contrast this with the framework of Roth and Verheij. Recall from Section 4.3 that they use the flat version of constraint. More specifically, applying the two leftmost transformations of (4.3.3) to the dismissal hierarchy in (4.4.1), we obtain the following set of factors directly favoring or opposing dismissal voidance:



Now we can simply apply the RM, i.e. Definition 2.3 of constraint, and find:

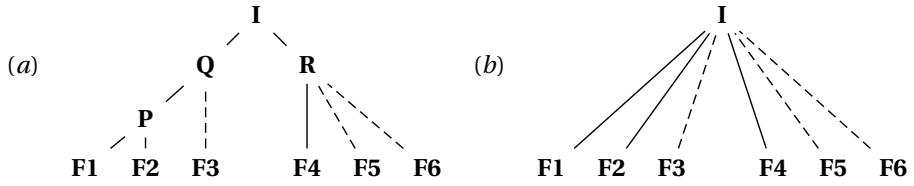
$\{Y\}, X \models \pi$
iff • for all $p \in \mathsf{Pro}(\boldsymbol{\pi})$: if $Y \models p$ then $X \models p$, and
• for all $p \in \mathsf{Con}(\boldsymbol{\pi})$: if $Y \models \neg p$ then $X \models \neg p$
iff $X \models \mathbf{F5}$.

As we assumed that **F5** applies in X, this means that according to flat constraint we indeed have $\{Y\}, X \models \pi$—regardless of whether **F1** was deemed to apply in Y or not. In the words of Roth and Verheij (2004, adapted to our notation): "there is more dialectical support for conclusion $\boldsymbol{\pi}$ in X than in Y, so that the conclusion $\boldsymbol{\pi}$ can follow in X as well. Note that for concluding to the outcome that there is more dialectical support for $\boldsymbol{\pi}$ in X, it does not matter how the conflict with regard to the intermediate **F1** is to be resolved."

In the previous example, we saw that there is an essential difference between flat and hierarchical constraint on an abstract factor p: for flat constraint, the truth values of the abstract factors below p are irrelevant with respect to the constraint induced by precedent on p, while for hierarchical constraint these truth values play an essential role. The differences between flat and hierarchical constraint are treated extensively by Canavotto and Horty (2023b, Section 5) in the context of the reason model, and Bench-Capon (2024) argues that flat constraint is more faithful to legal practice than the hierarchical variant. The following example illustrates his point of view.

Example 4.10. Bench-Capon (2024, Section 3) has argued that flat constraint is a more appropriate notion than hierarchical constraint. His argument is guided by an example on a *treat* domain, in which parents use several factors to decide whether their children should be given treats or not.[2]

The factor hierarchy of this domain is given by (*a*) below:



An overview of the meaning of the factors can be found in Table 4.2. Bench-Capon

[2]This non-legal example is also used by Canavotto and Horty (2023a), continuing a tradition of Twining and Miers (2010).

Table 4.2: The factors of the treat domain, in which parents decide whether or not to give their child ice cream, used by Bench-Capon (2024) and Canavotto and Horty (2023a).

Factor	Description
I	Deserved ice cream.
P	Tidied room.
Q	Behaved well at home.
R	Behaved well at school.
F1	Folded clothes.
F2	Made the bed.
F3	Threw toys.
F4	Handed in homework.
F5	Was inattentive in class.
F6	Interrupted the teacher.

considers several cases for this domain, one of which is called *MaxMonday*, which we will denote by X, given by $X \vDash \{\mathbf{F1}, \neg\mathbf{F2}, \neg\mathbf{F3}, \neg\mathbf{F4}, \mathbf{F5}, \neg\mathbf{F6}, \mathbf{P}, \mathbf{Q}, \neg\mathbf{R}, \mathbf{I}\}$.[3] As we can see, in the case X the child was given ice cream, based on the presence of the factors **F1** and **F5**, and the absence of the rest of the base-level factors. The parents deemed that the child tidied their room, (and thus) behaved well at home, but did not behave well at school because they were inattentive in class.

Bench-Capon then discusses two possible interpretations of the decisions by the parents regarding X. The first is the hierarchical interpretation that the factor **Q** of good behaviour at home outweighs the factor $\neg\mathbf{R}$ of bad behaviour at school. Bench-Capon argues that this is a quite sweeping judgment, and that it is inappropriate to generalize this single decision by the parents to mean the parents find that good behaviour at home always outweighs bad behaviour at school. He favors a second interpretation, which is that the base-level factor **F1** outweighs **F5** (granted that no other base-level factors apply).

It is our interpretation of the factor hierarchy (*a*) above that the decision in *MaxMonday* indeed represents a decision that **Q** outweighs $\neg\mathbf{R}$. This is based on our interpretation of the meaning of a "factor hierarchy:" what is relevant with respect to constraint is *whether* a subordinate factor applies and not *why* it applies. This can be seen in the more realistic hierarchy of (4.1.1)—with respect to a **Bail** decision, it is relevant to know if the recidivism risk factor **Recid** applies or not. We regard the factor hierarchy of (4.1.1) as encoding an assumption that it is relevant for a **Bail** decision to know whether **Recid** does or does not apply, and that it is not relevant for a **Bail** decision to know why **Recid** does or does not apply. If the specific circumstances of the causes influencing the application of **Recid** were directly influential to a **Bail** decision, then this should be encoded in the hierarchy by direct links to the **Bail** issue from the base-level factors subordinate to **Recid**.

Another consideration with respect to hierarchical constraint that touches on this is whether the decision-maker wants to be consistent with respect to decisions about the application of intermediate abstract factors. Bench-Capon mentions that the parents of Max want to be consistent with respect to the issue **I** of granting treats or not, and thus opt to apply a fortiori case-based reasoning. However, he does not specify whether they also want to be consistent with respect to decisions on the intermediate facotrs **P**, **Q**, and **R**.

[3] Bench-Capon only implicitly states the truth values of the abstract factors.

Hierarchical constraint takes consistency of intermediate factors into account, whereas flat constraint does not—we will discuss this more generally in Section 4.5.

Finally, we note that the HRM can be used by proponents of either the hierarchical or flat constraint view. This is because when the HRM is applied to a flat hierarchy such as (*b*) above, it acts exactly as the RM would, as we will show in Section 4.7. In fact, the HRM can be more generally applied because the flattening procedure of (4.3.3) may produce factors that are simultaneously pro and con an issue; this is why Canavotto and Horty (2023b, Definition 3) only consider hierarchies for which this cannot occur.

4.5 Consistency

As with the RM, the definition of constraint also gives rise to a notion of (strong) consistency of case bases. The difference is that there are now multiple factors on which constraint is induced, rather on just the outcome. As a result, we have a notion of inconsistency for each of the abstract factors.

Definition 4.11. Let $p \in A$ be an abstract factor, and $\mathscr{C}$ a case base. A fact situation X is *p-inconsistent* with respect to $\mathscr{C}$ if both $\mathscr{C}, X \models p$ and $\mathscr{C}, X \models \neg p$; otherwise, X is *p-consistent*.

Example 4.12. To give an example of p-inconsistency we return to the setting of Example 4.8. Suppose we had, in addition to the fact situations X, Y, Z, a fourth fact situation W such that $W \models \{\textbf{Recid}, \textbf{Record}, \neg\textbf{Male}, \textbf{Education}, \textbf{Married}, \textbf{Age}\}$. We can compute that $\{W\}, X \models \textbf{Recid}$, and so (by Proposition 4.13 below) we have $\{W, Y, Z\}, X \models \textbf{Recid}$ and $\{W, Y, Z\}, X \models \neg\textbf{Recid}$, which means X is **Recid**-inconsistent with respect to the case base $\{W, Y, Z\}$. This need not imply that X is also inconsistent with respect to more abstract factors above **Recid**, such as in this case the **Bail** factor. If, for example, we had $Z \models \textbf{Bail}$ and $W \models \textbf{Bail}$ while leaving the truth values of the other factors the same, then X would be **Recid**-inconsistent and **Bail**-consistent.

4.6 Monotonicity

The HRM, like the RM, is monotonic in the addition of new cases.

Proposition 4.13. *Let $p \in F$, $\mathscr{C} \subseteq \mathscr{D}$ be two case bases, and X some fact situation; then $\mathscr{C}, X \models p$ implies $\mathscr{D}, X \models p$, and $\mathscr{C}, X \models \neg p$ implies $\mathscr{D}, X \models \neg p$.*

Proof. We proceed by mutual induction on the position of p in the hierarchy H. For the base case, when p is H-minimal (so $p \in B$) Lemmas 4.6 and 4.7 tell us

$$\mathscr{C}, X \models p \quad \text{iff} \quad X \models p \quad \text{iff} \quad \mathscr{D}, X \models p,$$

and similarly $\mathscr{C}, X \models \neg p$ iff $\mathscr{D}, X \models \neg p$. In the induction case, when $p \in A$, we know that for all $q \in \mathsf{Pro}(p) \cup \mathsf{Con}(p)$: $\mathscr{C}, X \models q$ implies $\mathscr{D}, X \models q$, and $\mathscr{C}, X \models \neg q$ implies $\mathscr{D}, X \models \neg q$;

therefore:

$\mathscr{D}, X \models p$
iff • $X \models p$, or
  • $p \in A$ and there is $Y \in \mathscr{D}$ with $Y \models p$ and
    – for all $q \in \mathsf{Pro}(p)$: if $Y \models q$ then $\mathscr{D}, X \models q$
    – for all $q \in \mathsf{Con}(p)$: if $Y \models \neg q$ then $\mathscr{D}, X \models \neg q$
if • $X \models p$, or (∗)
  • $p \in A$ and there is $Y \in \mathscr{C}$ with $Y \models p$ and
    – for all $q \in \mathsf{Pro}(p)$: if $Y \models q$ then $\mathscr{C}, X \models q$
    – for all $q \in \mathsf{Con}(p)$: if $Y \models \neg q$ then $\mathscr{C}, X \models \neg q$
iff $\mathscr{C}, X \models p$.

Step (∗) follows from $\mathscr{C} \subseteq \mathscr{D}$ and the induction hypothesis: if there is $Y \in \mathscr{C}$ satisfying the conditions described of (∗), then $Y \in \mathscr{D}$. Furthermore, given $q \in \mathsf{Pro}(p)$, if $Y \models q$, then $\mathscr{C}, X \models q$ and so $\mathscr{D}, X \models q$ by the induction hypothesis; the case for $q \in \mathsf{Con}(p)$ follows the same pattern. By similar reasoning we can show $\mathscr{C}, X \models \neg p$ implies $\mathscr{D}, X \models \neg p$, which completes the induction case and thus the proof. □

Just as the RM, the HRM is also monotonic with respect to the extension of fact situations. The proof of this is just as in Proposition 2.16, so we omit it.

Proposition 4.14. *If $X \subseteq Y$ then $\mathscr{C}, X \models p$ implies $\mathscr{C}, Y \models p$, and $\mathscr{C}, X \models \neg p$ implies $\mathscr{C}, Y \models \neg p$.*

Remark 4.15. We mention, as in Remark 2.18, that the differences between options (1) and (2) lead to different results regarding monotonicity. In fact, with option (2), the HRM is nonmonotonic with respect to both the addition of cases to the case base, and to addition of information to the focus fact situation, the former of which was shown by van Woerkom et al. (2023a, Proposition 4.18).

4.7 Relation to the result model

We will now show that the HRM is a conservative extension of the RM, in the sense that when the HRM is restricted to hierarchies with only one abstract factor, and case bases which are defined on this factor, it reduces to the RM. To show this, we will construct a translation f that maps instances of the RM to instances of the HRM, and prove that this translation respects Definitions 2.3 and 4.3 of constraint.

Definition 4.16. A factor hierarchy is *flat* if its set of abstract factors is a singleton.

Any factor partition $F = \text{Pro} \cup \text{Con}$ can be mapped to a flat hierarchy $f(\text{Pro}, \text{Con})$. To do this, we first introduce a new factor p_π (so $p_\pi \notin F$), and then define

$$f(\text{Pro}, \text{Con}) = (F \cup \{p_\pi\}, \text{Pro} \times \{p_\pi\}, \text{Con} \times \{p_\pi\}).$$

It is easily checked that $f(\text{Pro}, \text{Con})$ satisfies the requirements of Definition 4.1, and that the set of abstract factors of $f(\text{Pro}, \text{Con})$ is the singleton $\{p_\pi\}$. Next, we extend f to operate on fact situations, cases, and case bases. Given a fact situation X for the RM (with respect to the factor partition (Pro, Con)) we translate it to a fact situation $f(X)$ for the HRM (with respect to the factor hierarchy $f(\text{Pro}, \text{Con})$) by defining $f(X)(p) = X(p)$ if $p \in \text{dom}(X)$, and leaving it undefined otherwise. Similarly, a case (X, s) can be translated to a case $f(X, s)$ by defining $f(X, s)(p)$ for $p \in F \cup \{\pi\}$ by

$$f(X,s)(p) = \begin{cases} X(p) & \text{if } p \in \text{dom}(X), \\ \mathbf{t} & \text{if } p = p_\pi \text{ and } s = \pi, \\ \mathbf{f} & \text{if } p = p_\pi \text{ and } s = \delta. \end{cases}$$

Lastly, given a case base $\mathscr{C}$ we write $f[\mathscr{C}] = \{f(X, s) \mid (X, s) \in \mathscr{C}\}$.

The translation f preserves and reflects constraint, in the following sense.

Theorem 4.17. *Given a case base $\mathscr{C}$ for a factor partition $F = \text{Pro} \cup \text{Con}$ and a focus fact situation X we have*

$$\mathscr{C}, X \models \pi \text{ iff } f[\mathscr{C}], f(X) \models p_\pi \quad \text{and} \quad \mathscr{C}, X \models \delta \text{ iff } f[\mathscr{C}], f(X) \models \neg p_\pi.$$

Proof. We consider the first equivalence. Spelling out Definition 4.3 we get

$f[\mathscr{C}], f(X) \models p_\pi$
iff • $f(X) \models p_\pi$, or
• $p_\pi \in A$ and there is a case $Y \in f[\mathscr{C}]$ with $Y \models p_\pi$ and
– for all $q \in \mathsf{Pro}(p_\pi)$: if $Y \models q$ then $f[\mathscr{C}], f(X) \models q$, and
– for all $q \in \mathsf{Con}(p_\pi)$: if $Y \models \neg q$ then $f[\mathscr{C}], f(X) \models \neg q$.

By definition, $f(X) \not\models p_\pi$, and so only the second disjunct in this statement remains. The requirement that $p_\pi \in A = \{p_\pi\}$ holds, and so can be removed. We also know that $\mathsf{Pro}(p_\pi) = \text{Pro}$ and $\mathsf{Con}(p_\pi) = \text{Con}$, so these can be substituted. Further filling in the shape of $Y \in f[\mathscr{C}]$ as $f(Y, s)$ for some $(Y, s) \in \mathscr{C}$ we can thus continue with:

iff there is a case $(Y, s) \in \mathscr{C}$ with $f(Y, s) \models p_\pi$ and
– for all $q \in \text{Pro}$: if $f(Y, s) \models q$ then $f[\mathscr{C}], f(X) \models q$, and
– for all $q \in \text{Con}$: if $f(Y, s) \models \neg q$ then $f[\mathscr{C}], f(X) \models \neg q$.

By definition of $f(Y, s)$, the requirement that $f(Y, s) \models p_\pi$ just reduces to $s = \pi$; and, again by definition, $f(Y, s) \models q$ holds iff $Y \models q$. Since the hierarchy is flat, the factors over which the 'for all' statements are quantifying are base-level, and so the statement that $f[\mathscr{C}], f(X) \models q$ reduces to $f(X) \models q$ by Lemma 4.7, which is in turn equivalent to $X \models q$ by definition of $f(X)$. We thus continue with:

iff there is a case $(Y, \pi) \in \mathscr{C}$ such that
– for all $q \in \text{Pro}$: if $Y \models q$ then $X \models q$, and
– for all $q \in \text{Con}$: if $Y \models \neg q$ then $X \models \neg q$
iff $\mathscr{C}, X \models s$.

The proof of the second equivalence is very similar to the above, so we omit it. □

4.8 Conclusion: Moving to hierarchical structures

In this chapter we considered the hierarchical result model (HRM): an extension of the result model (RM) that operates on a knowledge representation framework based on the concept of factor hierarchies. We developed a notion of "hierarchical" constraint for this setting, and constrasted it with its "flat" counterpart which has been considered in the literature. We then showed that our notion is monotonic in both the addition of new cases and the addition of new information to a fact situation, and can be seen to extend the notion of constraint of the RM when the hierarchy under consideration is flat. Lastly, we discussed the evident adaptation of case base consistency to this version of constraint.

So far we have discussed the RM, in Chapter 2, and subsequently two of its extensions: the dimensional result model (DRM), in Chapter 3, and the HRM, in Chapter 4. Our main goal in studying and further developing these variations of the RM is to apply them to build explanation methods for data-driven decisions, such as those made by machine-learned AI systems. To do this, it is important that we use a knowledge representation framework that uses dimensions rather than factors, because the data used for machine learning system is almost always non-binary. This raises the question: can we adapt the HRM so that it operates on the basis of "dimension hierarchies," and thus unify the HRM and the DRM? The next chapter is dedicated to developing this unification.

Chapter 5

Modeling Hierarchical Dimensional Constraint

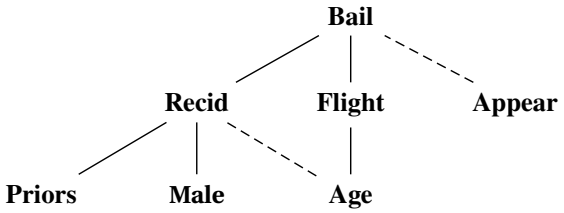
XTENDING THE KNOWLEDGE REPRESENTATIONS of Chapters 3 and 4, this chapter unifies them, along with their associated notions of constraint, by introducing a set of dimensions which has hierarchical structure—we call this a *dimension hierarchy*.[1] We start in Section 5.1 by extending the running example from Section 2.1 to incorporate both dimensional and hierarchical information. The notion of dimension hierarchy is then formalized in Section 5.2, and its associated notion of constraint is given in Section 5.3. As before, this induces a notion of consistency, which we define in Section 5.4. We refer to the resulting model as the DHRM and the rest of the sections are dedicated to investigating some of its formal properties. In particular, in Section 5.5, we show that the DHRM is monotonic in both the addition of cases to the case base and in addition to information to a focus fact situation. Then, in Section 5.6, we show that the DHRM is a conservative extension of both the HRM and the DRM, respectively. Finally, we end with some concluding remarks in Section 5.7.

5.1 An example of a dimension hierarchy

A major difference between the DRM and the DHRM is that dimensions can now be subordinate to multiple more abstract factors. This raises the question whether the dimension preference orders $\preceq$ should be specified per dimension or per link in the hierarchy. Additionally, dimensions are now subordinate to other dimensions, rather than to a (binary) case outcome or legal issue. What does it mean for a dimension to influence a more abstract dimension? We will illustrate these differences through some examples to build intuition, before continuing to the formal definitions.

[1] The formalism described in this chapter was first presented by van Woerkom et al. (2023b). The results related to monotonicity, as well as the formal comparison to the variants of the RM discussed in the preceding chapters, are from van Woerkom et al. (2025).

Consider the following modification to the factor hierarchy of (4.1.1):



(5.1.1)

In the setting of a dimension hierarchy we can consider recidivism risk as a dimension, for instance as a score ranging from 1 to 10. Additionally, **Bail** can now be considered a dimension, specifying the amount of bail in, say, USD. Note that denial of bail can still be modeled as an 'infinite' amount of bail. **Appear**, too, can be considered a dimension, indicating the relative frequency of past trial appearances by the defendant.

To begin building some intuition for an appropriate notion of constraint in this setting we illustrate a difference with the DRM, which is that dimensions now affect other dimensions instead of the case outcome directly. To this end, we consider the subgraph of (5.1.1) consisting of just the dimensions **Recid**, **Priors**, **Male**, and **Age**, and the fact situations X, Y, and Z, listed in Table 5.1. The situation Z concerns a 25-year-old female with 2 prior offenses. What recidivism risk score may be consistently assigned to Z, given the previous judgments that a 30-year-old female with 1 prior offense received score 5 (situation X), and that a 20-year-old male with 4 prior offenses received score 8 (situation Y)? Comparing the situation Z to X we see that Z is dimension-wise equal or more indicative of recidivism risk than X: Z is younger, both Z and X are female, and Z has more prior offenses. Since X received a recidivism risk score of 5, it seems sensible to require that Z would get *at least* a score of 5, but possibly higher since Z is indicative of higher risk on some dimensions. This exemplifies one of the key differences between the DHRM and the previous models—decisions are not forced exactly, but constrained to lie within an interval. Comparing Z to the situation Y we get the opposite picture; Y has received a risk score of 8, but Z is dimension-wise equal or less indicative of recidivism risk than Y. Therefore, we expect Z to receive a score of *at most* 8. In sum, $\{X, Y\}$ should produce the constraint that $5 \preceq Z(\textbf{Recid}) \preceq 8$.

We now turn our attention to the full hierarchy, depicted by graph (5.1.1), involving a downstream judgment of bail amount. In such a scenario, we can apply a recursive notion of constraint as in the HRM. Consider, again, the fact situations listed in Table 5.1. We have seen that X and Y bind the recidivism score of Z to the integer range $[5, 8]$. In addition, we now have two situations V and W for which a bail amount was determined on the basis of their recidivism risk assessment, risk of flight, and relative frequency of previous trial appearances. We omit the information that led to the recidivism score assignment for these fact situations. This is a general feature of the hierarchical models—once an abstract dimension has been given a value it can be used for downstream reasoning regardless of the reason behind the value's assignment. Defendant V was granted a bail amount of \$2,500, on the basis of a recidivism risk score of 2, a perceived low risk of flight, and an 80% appearance rate at previous trials. Defendant Z has a lower appearance rate at previous trials and is similarly perceived as unlikely to flee, but is not yet assigned a definitive recidivism risk score. However, since we know that Z should receive a risk score of at least

Table 5.1: Five example fact situations V, W, X, Y, and Z for the bail domain.

	Age	**Male**	**Priors**	**Recid**	**Flight**	**Appear**	**Bail**
X	30	0	1	5	–	–	–
Y	20	0	4	8	–	–	–
V	–	–	–	2	0	0.8	\$2,500
W	–	–	–	9	1	0.3	\$20,000
Z	25	0	2	?	0	0.5	?

5 it will in any case be higher than V's score of 2. Therefore, we would ultimately expect Z to receive a bail amount which is equal or greater than that of V—so \$2,500 $\preceq Z($**Bail**$)$. Similarly, we can deduce from the case W that, since Z should receive a risk score of at most 8, the amount of bail for Z should not exceed \$20,000—so $Z($**Bail**$) \preceq$ \$20,000. In sum, the case base $\{V, W, X, Y\}$ should produce the constraints \$2,500 $\preceq Z($**Bail**$) \preceq$ \$20,000.

This use of recursion is useful, because it allows the use of the forcing relation despite some dimensions not having been assigned an exact value. Consider, for instance, the decision support system implemented by the Dutch National Police Force (Odekerken & Bex, 2020). It is argued by Odekerken et al. (2023b) that determining the values of dimensions for a specific case can be costly, and the aforementioned use of recursion can alleviate this need for abstract factors.

5.2 Knowledge representation

We now adapt Definition 4.1 of a factor hierarchy to that of a dimension hierarchy.

Definition 5.1. A *dimension hierarchy* is a triple $(D, \mathsf{Pro}, \mathsf{Con})$, where D is a finite set of dimensions, and Pro and Con are relations on D such that $\mathsf{Pro} \cap \mathsf{Con} = \emptyset$, and such that the transitive closure of $\mathsf{H} = \mathsf{Pro} \cup \mathsf{Con}$ is irreflexive.

We maintain the same terminology as in the HRM for factors with respect to their position in the hierarchy—a dimension is *base-level* if it is H-minimal, and the set of base level dimensions is denoted by B; if a dimension is not base-level then it is *abstract*, and we write A for the set of abstract dimensions. A *fact situation* X is a choice function on a subset of D; we denote its domain by $\mathrm{dom}(X)$. Lastly, we assume each dimension $d \in D$ is assigned a partial order $\preceq$ on d.

The way we define a dimension hierarchy in this section is with a fixed order for every node of the hierarchy; outgoing positive links represent a positive correlation with respect to this order, and negative links represent a negative correlation. The reader may wonder about the situation when a dimension has an influence on multiple more abstract dimension—but with respect to different orders. Consider, for example, an **Education** dimension, with three possible values: `none`, `high school`, and `higher education`. With respect to influence on risk of recidivism, we may order these values by `higher education` $\prec$ `high school` $\prec$ `none`; but **Education** may also influence other dimensions with neither this nor its inverse order. For example, a judge may take into account the defendant's societal ties when determining flight risk. Having had a form of education could be

taken as reducing flight risk, without differentiating specifically between a high school diploma and higher education. So, with respect to flight risk we could use an ordering with `none` $\prec$ `high school` and `none` $\prec$ `higher education`, but with `high school` and `higher education` mutually incomparable.

One way to deal with this would be to associate orders with links, rather than with dimensions directly. In the previous example, this would mean that the **Education** dimension has one outgoing link with respect to the linear order `higher education` $\prec$ `high school` $\prec$ `none`, and one with respect to the order in which `high school` and `higher education` are incomparable. The difficulty with this approach is that, if a dimension does not have a single order with which it is associated, it becomes cumbersome to specify what it means for two dimensions to be positively correlated. For example, we would like to interpret a positive link of the form $d \to e$ in a hierarchy to mean that "a higher value for d tends to imply a higher value for e," but the meaning of this statement would be unclear in a scenario where there are multiple orderings associated to the dimension e. To deal with that, each link of the form $d \to e$ in the hierarchy would need to specify a dimension order on d and on e, and thus incur an additional bookkeeping cost. The resulting notion of constraint would seem to become more complex than the one we will shortly propose in Definition 5.2.

We thus opt for a simpler solution: Fix one order for every dimension, and let the dimension influence, and be influenced, only according to this fixed order. To deal with cases where a dimension has an influence with respect to multiple different orders, such as the aforementioned **Education** dimension, we require the dimension to be added multiple times to the hierarchy, with one instance for each of the orders with which it has an influence on other dimensions. We think the price of the redundancy in this representation is compensated by a simpler definition of constraint.

5.3 Constraint

As in the HRM we define $\mathsf{Pro}(d) = \{e \in D \mid \mathsf{Pro}(e,d)\}$. Using this, together with the definitions of support and opposition of (3.4.1) and (3.4.2), we now define by mutual recursion two relations $\mathscr{C} \vDash v \preceq X(d)$ and $\mathscr{C} \vDash X(d) \preceq v$, indicating constraint in the form of lower and upper bounds for $X(d)$ in a focus fact situation X.

Definition 5.2. Given a case base $\mathscr{C}$ and a value v in some dimension d, a fact situation X is *lower bounded* by v and $\mathscr{C}$, written $\mathscr{C} \vDash v \preceq X(d)$, if and only if either

- $v \preceq X(d)$, or
- $d \in A$ and there is $Y \in \mathscr{C}$ such that $v \preceq Y(d)$ and
  - for all $e \in \mathsf{Pro}(d) \cap \mathrm{supp}(Y)$: $\mathscr{C} \vDash Y(e) \preceq X(e)$, and
  - for all $e \in \mathsf{Con}(d) \cap \mathrm{opp}(Y)$: $\mathscr{C} \vDash X(e) \preceq Y(e)$.

Similarly, the *upper bound* by v, written $\mathscr{C} \vDash X(d) \preceq v$, is defined to hold iff:

- $X(d) \preceq v$, or
- $d \in A$ and there is $Y \in \mathscr{C}$ such that $Y(d) \preceq v$ and

– for all $e \in \mathsf{Pro}(d) \cap \mathrm{opp}(Y)$: $\mathscr{C} \models X(e) \preceq Y(e)$, and
– for all $e \in \mathsf{Con}(d) \cap \mathrm{supp}(Y)$: $\mathscr{C} \models Y(e) \preceq X(e)$.

The idea behind the recursive clause is that there is a precedent Y which, by the a fortiori principle, forces $X(d)$ to take a value which is at least $Y(d)$, and therefore $v \preceq X(d)$ follows by transitivity from $v \preceq Y(d) \preceq X(d)$. Note that, by definition, $\mathscr{C} \models v \preceq X(d)$ is always false when X is undefined on d.

Analogous to Lemmas 4.6 and 4.7 we have some simple consequences of Definition 5.2.

Lemma 5.3. *$v \preceq X(d)$ implies $\mathscr{C} \models v \preceq X(d)$, and $X(d) \preceq v$ implies $\mathscr{C} \models X(d) \preceq v$.*

Lemma 5.4. *If $d \in B$ then $\mathscr{C} \models v \preceq X(d)$ implies $v \preceq X(d)$ and $\mathscr{C} \models X(d) \preceq v$ implies $X(d) \preceq v$.*

Example 5.5. To verify Definition 5.2 correctly captures the intuition of the example in Section 5.1 we now consider a dimension hierarchy $(D, \mathsf{Pro}, \mathsf{Con})$ as depicted in (5.1.1), so

$$\begin{aligned}
D &= \{\textbf{Priors}, \textbf{Male}, \textbf{Age}, \textbf{Recid}, \textbf{Flight}, \textbf{Appear}, \textbf{Bail}\},\\
\textbf{Priors} &= \{0, 1, 2, \ldots\},\\
\textbf{Male} &= \{0, 1\},\\
\textbf{Age} &= \{18, 19, 20, \ldots\},\\
\textbf{Recid} &= \{1, 2, \ldots, 9, 10\},\\
\textbf{Flight} &= \{0, 1\},\\
\textbf{Appear} &= [0, 1],\\
\textbf{Bail} &= \{\$x \mid x \in \{0, 1, 2, \ldots\}\},\\
\mathsf{Pro} &= \{(\textbf{Priors}, \textbf{Recid}), (\textbf{Male}, \textbf{Recid}), (\textbf{Age}, \textbf{Flight}), (\textbf{Recid}, \textbf{Bail}),\\
&\qquad (\textbf{Flight}, \textbf{Bail})\},\\
\mathsf{Con} &= \{(\textbf{Age}, \textbf{Recid}), (\textbf{Appear}, \textbf{Bail})\}.
\end{aligned}$$

The dimension orders are all just given by the usual less-than order $\leq$. We let W, Y, and Z be as listed in Table 5.1. The question is now whether $\{W, Y\} \models Z(\textbf{Bail}) \preceq \$20{,}000$. To check this, we first verify that $\{W, Y\} \models Z(\textbf{Recid}) \preceq 9$:

$\{W, Y\} \models Z(\textbf{Recid}) \preceq 9$
if there is $T \in \{W, Y\}$ such that $T(\textbf{Recid}) \preceq 9$ and
- for all $d \in \mathsf{Pro}(\textbf{Recid}) \cap \mathrm{opp}(T)$: $\{W, Y\} \models Z(d) \preceq T(d)$, and
- for all $d \in \mathsf{Con}(\textbf{Recid}) \cap \mathrm{supp}(T)$: $\{W, Y\} \models T(d) \preceq Z(d)$.

We may now substitute $T = Y$ as $Y(\textbf{Recid}) = 8 \preceq 9$:

if • for all $d \in \mathsf{Pro}(\textbf{Recid}) \cap \mathrm{opp}(Y)$: $\{W, Y\} \models Z(d) \preceq Y(d)$, and
- for all $d \in \mathsf{Con}(\textbf{Recid}) \cap \mathrm{supp}(Y)$: $\{W, Y\} \models Y(d) \preceq Z(d)$.

Note that $\mathsf{Pro}(\textbf{Recid}) \cap \mathrm{opp}(Y) = \{\textbf{Priors}\}$ ($\textbf{Male} \notin \mathrm{opp}(Y)$ as $Y(\textbf{Male}) = 1$ is its greatest element) and $\mathsf{Con}(\textbf{Recid}) \cap \mathrm{supp}(Y) = \{\textbf{Age}\}$, so this evaluates to

if $\{W, Y\} \models Z(\textbf{Priors}) \preceq Y(\textbf{Priors})$ and $\{W, Y\} \models Y(\textbf{Age}) \preceq Z(\textbf{Age})$.

All dimensions subordinate to **Recid** are base-level so this simplifies by Lemma 5.4 to:

if $Z(\textbf{Priors}) \preceq 4$ and $20 \preceq Z(\textbf{Age})$.

Indeed, defendant Z of Table 5.1 satisfies these conditions and so $\{W, Y\} \models Z(\textbf{Recid}) \preceq 9$. Next, we proceed in the same fashion to confirm that $\{W, Y\} \models Z(\textbf{Bail}) \preceq \$20{,}000$:

$\{W, Y\} \models Z(\textbf{Bail}) \preceq \$20{,}000$
if there is $T \in \{W, Y\}$ such that $T(\textbf{Bail}) \preceq \$20{,}000$ and
- for all $d \in \mathsf{Pro}(\textbf{Bail}) \cap \mathrm{opp}(T)$: $\{W, Y\} \models Z(d) \preceq T(d)$, and
- for all $d \in \mathsf{Con}(\textbf{Bail}) \cap \mathrm{supp}(T)$: $\{W, Y\} \models T(d) \preceq Z(d)$

if • for all $d \in \mathsf{Pro}(\textbf{Bail}) \cap \mathrm{opp}(Y)$: $\{W, Y\} \models Z(d) \preceq Y(d)$, and
- for all $d \in \mathsf{Con}(\textbf{Bail}) \cap \mathrm{supp}(Y)$: $\{W, Y\} \models Y(d) \preceq Z(d)$

if $\{W, Y\} \models Z(\textbf{Recid}) \preceq 9$, and $\{W, Y\} \models 0.3 \preceq Z(\textbf{Appear})$
if $\{W, Y\} \models Z(\textbf{Recid}) \preceq 9$ and $0.3 \preceq Z(\textbf{Appear})$.

We have established that $\{W, Y\} \models Z(\textbf{Recid}) \preceq 9$, and so as $0.3 \preceq 0.5 = Z(\textbf{Appear})$ we indeed have $\{W, Y\} \models Z(\textbf{Bail}) \preceq \$20{,}000$ as desired.

Remark 5.6. A dimension hierarchy, like a factor hierarchy, can contain positive and negative links. It is worth noting that the polarity of these links can be switched by reversing the dimension order associated with the subordinate dimension. Consider, as an example, the negative link between **Age** and **Recid** in (5.1.1). The order associated with **Age** is the usual order $\leq$ on natural numbers, and so, since older people tend to recidivate less, the **Age** dimension has a negative link to the **Recid** dimension. However, if we choose to pair **Age** with the $\geq$ order on the natural numbers, then this same relation between age and recidivism would be represented by a positive link. Of course, changing the order of **Age** in this hierarchy would mean that the polarity of the link from **Age** to **Flight** would have to be flipped as well.

5.4 Consistency

As before, the notion of constraint comes with a notion of (strong) case base consistency, and just as in the HRM (Definition 4.11) this notion can be applied to any abstract factor p, because we have defined constraint as working on all abstract factors.

Definition 5.7. Let $\mathscr{C}$ be a case base for a dimension hierarchy (D, H), $d \in A$ an dimension, and X a fact situation; X is *d-inconsistent* with respect to $\mathscr{C}$ if there are values $v \prec w \in d$ such that both $\mathscr{C} \models X(d) \preceq v$ and $\mathscr{C} \models w \preceq X(d)$; otherwise X is *d-consistent*.

Example 5.8. To illustrate this definition we once again turn to our running example. Consider the hierarchy in (5.1.1), and the fact situations X and Z listed in Table 5.1. Following the line of reasoning in Example 5.5 we have that $\{X\} \models Z(\textbf{Recid}) \preceq 9$. Suppose that the decisionmaker of this domain does not follow this constraint and assigns $Z(\textbf{Recid}) = 10$, which also means that $\{X\} \models 10 \preceq Z(\textbf{Recid})$. This means that Z is is **Recid**-inconsistent with respect to $\mathscr{C}$, according to Definition 5.7 because we have values $9 \prec 10 \in \textbf{Recid}$ such

that $\{X\} \models Z(\textbf{Recid}) \preceq 9$ and $\{X\} \models 10 \preceq Z(\textbf{Recid})$. In this example Z is **Recid**-inconsistent because it assigns a value to **Recid** which violates the constraint induced by the case base. Do note that Definition 5.7 also allows fact situations to be inconsistent on dimensions on which they are undefined.

5.5 Monotonicity

Like the other models, the DHRM is monotonic in both the addition of new cases and in the addition of information to fact situations.

Proposition 5.9. *Let $\mathscr{C} \subseteq \mathscr{D}$ be case bases, d a dimension, $v \in d$ a value, and be case bases; then $\mathscr{C} \models v \preceq X(d)$ implies $\mathscr{D} \models v \preceq X(d)$ and $\mathscr{C} \models X(d) \preceq v$ implies $\mathscr{D} \models X(d) \preceq v$.*

Proof. The proof (just as that of Proposition 4.13) proceeds by induction on the position of d in the hierarchy H. The base case, when $d \in B$, follows from Lemmas 5.3 and 5.4. When $d \in A$ the induction hypothesis is that for all $e \in \mathsf{Pro}(d) \cup \mathsf{Con}(d)$ and any value $w \in e$: $\mathscr{C} \models w \preceq X(e)$ implies $\mathscr{D} \models w \preceq X(e)$ and $\mathscr{C} \models X(e) \preceq w$ implies $\mathscr{D} \models X(e) \preceq w$. Now, assume $\mathscr{C} \models v \preceq X(d)$. If this is due to $v \preceq X(d)$ then $\mathscr{D} \models v \preceq X(d)$ follows immediately. So, suppose that it is due to $Y \in \mathscr{C}$ with $v \preceq Y(d)$ with

– for all $e \in \mathsf{Pro}(d) \cap \mathrm{supp}(Y)$: $\mathscr{C} \models Y(e) \preceq X(e)$, and
– for all $e \in \mathsf{Con}(d) \cap \mathrm{opp}(Y)$: $\mathscr{C} \models X(e) \preceq Y(e)$.

Applying the induction hypothesis we can conclude that

– for all $e \in \mathsf{Pro}(d) \cap \mathrm{supp}(Y)$: $\mathscr{D} \models Y(e) \preceq X(e)$, and
– for all $e \in \mathsf{Con}(d) \cap \mathrm{opp}(Y)$: $\mathscr{D} \models X(e) \preceq Y(e)$.

which is to say that $\mathscr{D} \models v \preceq X(d)$. By the same reasoning we can show $\mathscr{C} \models X(d) \preceq v$ implies $\mathscr{D} \models X(d) \preceq v$, which completes the induction case and thus the proof. □

Proposition 5.10. *If $X \subseteq Y$ then $\mathscr{C} \models v \preceq X(d)$ implies $\mathscr{C} \models v \preceq Y(d)$, and similarly $\mathscr{C} \models X(d) \preceq v$ implies $\mathscr{C} \models Y(d) \preceq v$.*

Proof. We proceed by induction on d. In the base case we have by Lemmas 5.3 and 5.4 that if $\mathscr{C} \models v \preceq X(d)$ then $v \preceq X(d) = Y(d)$, and so $\mathscr{C} \models v \preceq Y(d)$; and similarly for $\mathscr{C} \models X(d) \preceq v$. The induction case proceeds just as in the proof of Proposition 5.9. □

5.6 Relation to the other models

In this section, we show that the DHRM is a conservative extension of the HRM in the sense that when the DHRM is restricted to dimension hierarchies with only binary dimensions, it reduces to the HRM. We then proceed in the same fashion to show that it is a conservative extension of the DRM.

Beginning with the HRM, we construct a translation f that maps instances of the HRM to instances of the DHRM, and prove that this translation respects Definitions 4.3 and 5.2 of constraint. Of course, the DHRM constrains fact situations to take values within

intervals, rather than forcing them to take specific values, but forcing specific values is easily expressed, for instance by defining $\mathscr{C} \vDash X(d) = v$ as the conjunction of $\mathscr{C} \vDash v \preceq X(d)$ and $\mathscr{C} \vDash X(d) \preceq v$. If v is a maximal element of a dimension d, meaning there is no w in d satisfying $v \prec w$, then it suffices to require that $\mathscr{C} \vDash v \preceq X(d)$. This is how we encode HRM-style forcing in the DHRM—by defining a binary dimension $d_p = \{\mathbf{f}_p, \mathbf{t}_p\}$ with $\mathbf{f}_p \prec \mathbf{t}_p$ (as in the translation in Theorem 3.19), and requiring $\mathscr{C} \vDash \mathbf{t}_p \preceq X(d_p)$, which amounts to the requirement that $X(d_p)$ takes the value $\mathbf{t}_p$.

Definition 5.11. A dimension hierarchy $(D, \mathsf{Pro}, \mathsf{Con})$ is *binary* if each $d \in D$ has cardinality 2 and is ordered linearly.

Let $(F, \mathsf{Con}, \mathsf{Pro})$ be a factor hierarchy; for any $p \in F$ we define a binary dimension $d_p = \{\mathbf{f}_p, \mathbf{t}_p\}$, linearly ordered by the reflexive closure of $\mathbf{f}_p \prec \mathbf{t}_p$. We now translate $(F, \mathsf{Con}, \mathsf{Pro})$ to a binary dimension hierarchy $f(F, \mathsf{Con}, \mathsf{Pro})$ by

$$f(F, \mathsf{Con}, \mathsf{Pro}) = (D, \mathsf{Con}', \mathsf{Pro}'),$$

where $D = \{d_p \mid p \in F\}$, $\mathsf{Pro}'(d_p, d_q)$ iff $\mathsf{Pro}(p, q)$, and similarly $\mathsf{Con}'(d_p, d_q)$ iff $\mathsf{Con}(p, q)$; so $\mathsf{Pro}'(d_p) = \{d_q \in D \mid q \in \mathsf{Pro}(p)\}$, and similarly $\mathsf{Con}'(d_p) = \{d_q \in D \mid q \in \mathsf{Con}(p)\}$. Next, we extend f to operate on fact situations and case bases. Given a fact situation X for the HRM (with respect to the factor hierarchy $(F, \mathsf{Pro}, \mathsf{Con})$) we translate it to a fact situation $f(X)$ for the DHRM (with respect to the dimension hierarchy $f(F, \mathsf{Pro}, \mathsf{Con})$):

$$f(X)(d_p) = \begin{cases} \mathbf{t}_p & \text{if } X \vDash p, \\ \mathbf{f}_p & \text{if } X \vDash \neg p, \\ \text{undefined} & \text{otherwise.} \end{cases}$$

As before, we write $f[\mathscr{C}] = \{f(X) \mid X \in \mathscr{C}\}$ for a case base $\mathscr{C}$.

The translation f preserves and reflects constraint, in the following sense.

Theorem 5.12. *Given a case base $\mathscr{C}$ for a factor hierarchy $(F, \mathsf{Pro}, \mathsf{Con})$, a focus fact situation X, and a factor $p \in F$, we have*

$$\mathscr{C}, X \vDash p \text{ iff } f[\mathscr{C}] \vDash \mathbf{t}_p \preceq f(X)(d_p) \quad \text{and} \quad \mathscr{C}, X \vDash \neg p \text{ iff } f[\mathscr{C}] \vDash f(X)(d_p) \preceq \mathbf{f}_p.$$

Proof. We begin by noting that for a partial valuation Y of F we have

$$\begin{aligned} \mathrm{dom}(f(Y)) &= \{d_q \in D \mid q \in \mathrm{dom}(Y)\}, \\ \mathrm{supp}(f(Y)) &= \{d_q \in \mathrm{dom}(f(Y)) \mid f(Y)(d_q) \text{ is not the least element of } d_q\} \\ &= \{d_q \in D \mid q \in \mathrm{dom}(Y) \text{ and } f(Y)(d_q) \neq \mathbf{f}_q\} \\ &= \{d_q \in D \mid Y \vDash q\}, \qquad (5.6.1) \\ \mathrm{opp}(f(Y)) &= \{d_q \in D \mid Y \vDash \neg q\}. \qquad (5.6.2) \end{aligned}$$

We proceed by induction on the position of p in the hierarchy H. In the base case, p is

base-level and therefore so is d_p, we thus have

$$\begin{aligned}
&\mathscr{C}, X \models p \\
&\text{iff } X \models p && \text{(Lemmas 4.6 and 4.7)} \\
&\text{iff } f(X)(d_p) = \mathbf{t}_p && \text{(def. of } f(X)\text{)} \\
&\text{iff } \mathbf{t}_p \preceq f(X)(d_p) && \text{(def. of } \preceq \text{ on } d_p\text{)} \\
&\text{iff } \mathscr{C} \models \mathbf{t}_p \preceq f(X)(d_p). && \text{(Lemmas 5.3 and 5.4)}
\end{aligned}$$

The other equivalence follows the same pattern.

In the induction case, when $p \in A$, we know that $\mathscr{C}, X \models q$ iff $f[\mathscr{C}] \models \mathbf{t}_q \preceq f(X)(d_q)$ and $\mathscr{C}, X \models \neg q$ iff $f[\mathscr{C}] \models f(X)(d_q) \preceq \mathbf{f}_q$ for every $q \in \mathsf{Pro}(p) \cup \mathsf{Con}(p)$; now,

$f[\mathscr{C}] \models \mathbf{t}_p \preceq f(X)(d_p)$
iff • $\mathbf{t}_p \preceq f(X)(d_p)$, or
- there is $Y \in f[\mathscr{C}]$ such that $\mathbf{t}_p \preceq Y(d_p)$ and
  - for all $d_q \in \mathsf{Pro}'(d_p) \cap \mathrm{supp}(Y)$: $f[\mathscr{C}] \models Y(d_q) \preceq f(X)(d_q)$,
  - for all $d_q \in \mathsf{Con}'(d_p) \cap \mathrm{opp}(Y)$: $f[\mathscr{C}] \models f(X)(d_q) \preceq Y(d_q)$

iff • $\mathbf{t}_p = f(X)(d_p)$, or
- there is $Y \in \mathscr{C}$ such that $\mathbf{t}_p \preceq f(Y)(d_p)$ and
  - for all $d_q \in \mathsf{Pro}'(d_p) \cap \mathrm{supp}(f(Y))$: $f[\mathscr{C}] \models f(Y)(d_q) \preceq f(X)(d_q)$,
  - for all $d_q \in \mathsf{Con}'(d_p) \cap \mathrm{opp}(f(Y))$: $f[\mathscr{C}] \models f(X)(d_q) \preceq f(Y)(d_q)$.

We now apply Equations (5.6.1) and (5.6.2), and fill in $f(Y)(d_q)$, to continue with:

iff • $X \models p$, or
- there is $Y \in \mathscr{C}$ such that $Y \models p$ and
  - for all $d_q \in D$ such that $q \in \mathsf{Pro}(p)$ and $Y \models q$: $f[\mathscr{C}] \models \mathbf{t}_q \preceq f(X)(d_q)$,
  - for all $d_q \in D$ such that $q \in \mathsf{Con}(p)$ and $Y \models \neg q$: $f[\mathscr{C}] \models f(X)(d_q) \preceq \mathbf{f}_q$.

This means we can apply the induction hypothesis to obtain:

iff • $X \models p$, or
- there is $Y \in \mathscr{C}$ such that $Y \models p$ and
  - for all $q \in \mathsf{Pro(p)}$: if $Y \models q$ then $\mathscr{C}, X \models q$,
  - for all $q \in \mathsf{Con}(p)$: if $Y \models \neg q$ then $\mathscr{C}, X \models \neg q$

iff $\mathscr{C}, X \models p$.

The other equivalence can be derived in the same manner, which completes the induction case and thus the proof. □

We now show that the DHRM is a conservative extension of the DRM, in the sense that when it is restricted to flat dimensions hierarchies, with only one 'outcome' issue, it reduces to the DRM. To show this, we will construct a translation f that maps instances of the DRM

to instances of the DHRM, and prove that this translation respects Definitions 3.7 and 5.2 of constraint. As in Section 5.6, we use the notion of lower bounding to a maximal element (or upper bounding to a minimal element) to play the role of forcing specific outcomes. In other words, a decision for an outcome dimension d_π is lower bounded to $\mathbf{t}$ in the translated DHRM instance if and only if the decision was forced for π in the original DRM instance.

Definition 5.13. A dimension hierarchy is *flat* if its set of abstract dimensions is a singleton, and *issue binary* if its issues are binary dimensions in the sense of Definition 3.18.

Let D be a set of dimensions. We define an 'outcome dimension' $d_\pi = \{\mathbf{t}, \mathbf{f}\}$, ordered by the reflexive closure of $\mathbf{f} \prec \mathbf{t}$, and relations Pro and Con on $D \cup \{d_\pi\}$ by $\mathsf{Pro} = D \times \{d_\pi\}$ and $\mathsf{Con} = \emptyset$, so that we have a flat, issue binary dimension hierarchy $f(D) = (D \cup \{d_\pi\}, \mathsf{Pro}, \mathsf{Con})$ (the choice of link polarity is arbitrary here, cf. Remark 5.6). Next, we extend f to operate on fact situations, cases, and case bases. Given a fact situation X for the DRM (with respect to the set of dimensions D) we translate it to a fact situation $f(X)$ for the DHRM (with respect to the dimension hierarchy $f(D)$) by defining $f(X)(d) = X(d)$ if $d \in \mathrm{dom}(X)$, and otherwise leaving $f(X)(d)$ undefined. For a case (X, s) we extend $f(X)$ to a choice function $f(X, s)$ on $D \cup \{d_\pi\}$ by:

$$f(X,s)(d) = \begin{cases} X(d) & \text{if } d \in \mathrm{dom}(X), \\ \mathbf{t} & \text{if } d = d_\pi \text{ and } s = \pi, \\ \mathbf{f} & \text{if } d = d_\pi \text{ and } s = \delta. \end{cases}$$

As before, we write $f[\mathscr{C}] = \{f(X,s) \mid (X,s) \in \mathscr{C}\}$.

The translation f preserves and reflects constraint, in the following sense.

Theorem 5.14. *Given a case base $\mathscr{C}$ for a factor partition $F = \mathrm{Pro} \cup \mathrm{Con}$ and a focus fact situation X we have $\mathscr{C}, X \models \pi$ iff $f[\mathscr{C}] \models \mathbf{t} \preceq f(X)(d_\pi)$; and similarly, $\mathscr{C}, X \models \delta$ iff $f[\mathscr{C}] \models f(X)(d_\pi) \preceq \mathbf{f}$.*

Proof. We begin by noting that for a case $(Y, \pi) \in \mathscr{C}$ we have

$$\begin{aligned} \mathrm{dom}(f(Y,\pi)) &= \mathrm{dom}(Y) \cup \{d_\pi\}, \\ \mathsf{Pro}(d_\pi) &= D, \\ \mathsf{Con}(d_\pi) &= \emptyset, \text{ and so} \\ \mathsf{Pro}(d_\pi) \cap \mathrm{supp}(f(Y,\pi)) &= \mathrm{supp}(Y), \\ \mathsf{Con}(d_\pi) \cap \mathrm{opp}(f(Y,\pi)) &= \emptyset. \end{aligned}$$

We now derive the equivalence $\mathscr{C}, X \models \pi$ iff $f[\mathscr{C}] \models \mathbf{t} \preceq f(X)(d_\pi)$:

$f[\mathscr{C}] \models \mathbf{t} \preceq f(X)(d_\pi)$
iff • $\mathbf{t} \preceq f(X)(d_\pi)$, or
- $d_\pi \in A$ and there is $Y \in f[\mathscr{C}]$ such that $\mathbf{t} \preceq Y(d_\pi)$ and
  - for all $d \in \mathsf{Pro}(d_\pi) \cap \mathrm{supp}(Y)$: $f[\mathscr{C}] \models Y(d) \preceq f(X)(d)$,
  - for all $d \in \mathsf{Con}(d_\pi) \cap \mathrm{opp}(Y)$: $f[\mathscr{C}] \models f(X)(d) \preceq Y(d)$

iff there is $(Y, s) \in \mathscr{C}$ such that $\mathbf{t} \preceq f(Y, s)(d_\pi)$ and
- for all $d \in \mathsf{Pro}(d_\pi) \cap \mathrm{supp}(f(Y, s))$: $f[\mathscr{C}] \models f(Y, s)(d) \preceq f(X)(d)$,
- for all $d \in \mathsf{Con}(d_\pi) \cap \mathrm{opp}(f(Y, s))$: $f[\mathscr{C}] \models f(X)(d) \preceq f(Y, s)(d)$

iff there is $(Y, \pi) \in \mathscr{C}$ such that
- for all $d \in \mathrm{supp}(Y)$: $f[\mathscr{C}] \models f(Y, s)(d) \preceq f(X)(d)$

iff there is $(Y, \pi) \in \mathscr{C}$ such that for all $d \in \mathrm{supp}(Y)$: $Y(d) \preceq X(d)$
iff $\mathscr{C}, X \models \pi$.

The derivation of $\mathscr{C}, X \models \delta$ iff $f[\mathscr{C}] \models f(X)(d_\pi) \preceq \mathbf{f}$ is very similar, so we omit it. □

5.7 Conclusion

In this chapter we considered the dimensional hierarchical result model (DHRM): a version of the result model (RM) that operates on a knowledge representation framework build around the concept of a dimension hierarchy. We developed a notion of hierarchical constraint for this setting, and illustrated it on our running example on criminal sentencing. As we did for the other models, we showed that this form of constraint is monotonic and induces a notion of case base consistency. Furthermore, we showed that the DHRM extends the dimensional result model (DRM) and the hierarchical result model (HRM).

This completes Part I of our work, which is primarily theoretical in nature. In Part II we will work towards applying the theory of the result model to the development of explainable AI techniques, and to the analysis of data-driven decisions.

Part II

Applications in Artificial Intelligence and Law

Tho Alisaunder say this,
Herith what he saide, ywis,
Hit is ywritein, Every thyng,
Himseolf schewith in tastyng.

—*Kyng Alisaunder*, early 14th century

AVING DEVELOPED a theory of a fortiori reasoning in the first part, we now turn to applying it to artificial intelligence and law in this second part. In particular, we investigate how it can be used to justify and analyze data-driven decisions. In Chapter 6 we review the a fortiori case-based argumentation method of explanation developed by Prakken and Ratsma (2022). We show how the theory of precedential constraint can be used in this context to develop formal notions of justification, compensation, and citability. Then, in Chapter 7, we show that the connection between the DRM and many-sorted logic, which was established in Chapter 3, can be used to implement the model of precedential constraint in the SMT solver Z3 (de Moura & Bjørner, 2008). This implementation is put to work by analyzing the consistency of several machine learning datasets. In particular, a logical analysis of several datasets with labels determined by logical formulas showcases the full capabilities of the SMT-based implementation. Lastly, in Chapter 8, we show how the implementation can be extended to compute constraint as defined by the DHRM model which was introduced in Chapter 5. This lets us analyze the internal consistency of decisions made by the COMPAS program. The results of this analysis contradict theoretical predictions, which bring to light a flaw in the dataset that has thus far received little attention in the literature. This indicates that case base consistency can be a useful measure for analyzing data-driven decisions.

Chapter 6

Explaining Data-Driven Outcomes

RAKKEN AND RATSMA (2022) developed a case-based reasoning method to explain data-driven automated decisions for binary classification, based on the theory of precedential constraint introduced by Horty (2011, 2019), which we discussed in Chapters 2 and 3.[1] This method, which we will refer to as "A Fortiori Case-Based Argumentation" (AF-CBA), is motivated by an analogy between the way in which a machine learning system draws on training data to assign a label to a new data point and the way in which a court of law draws on previously decided cases to make a decision about a new fact situation, because in both of these situations the precedent that has been set must be adhered to as closely as possible. The theory of precedential constraint, which has been developed to describe the type of a fortiori reasoning used for legal decision making on the basis of case law, can therefore be applied to analyze machine-learned decisions that are made on the basis of training data.

More specifically, the method of Prakken and Ratsma (2022) formally models the kind of dialogue in which lawyers cite precedents to argue in favor of their preferred outcome of the new fact situation. These citations, and the way in which they attack the opponent's citation, are formalized using an *abstract argumentwation framework* (Dung, 1995). A winning strategy in the grounded argument game on this framework, starting with an initial citation of a suitable precedent case, is taken as the explanation of the decision of the new fact situation.

In this chapter we examine the explanation model of Prakken and Ratsma (2022) in detail and make various suggestions and modifications for improvement. Particularly close attention is paid to the subject of *compensation*; the way in which important differences between a new fact situation and a precedent case can be compensated for by features of the focus case. We make the formal nature of this subject more explicit, and specify various desirable properties it may have.

Subsequently, we show that the model can be equivalently viewed as extending the

[1] All the material in this chapter has previously been published by van Woerkom et al. (2022b).

theory of precedential constraint with notions of *justification* and *citability* which, together with the notion of *forcing*, constitute the explanations produced by the model. This equivalent formulation only uses the simple notion of relations (in the set-theoretic sense), thus simplifying the specification of the model. The resulting view may be more broadly applied to the type of downplaying attacks seen in similar systems such as CATO (Aleven & Ashley, 1997).

We begin by summarizing the relevant aspects of the theory of precedential constraint in Section 6.1.1. In Section 6.1 we give a description of the explanation method of Prakken and Ratsma (2022). In Section 6.2 we revisit the definition of best citability, suggest some improvements, and demonstrate their potential experimentally. Then, in Section 6.3, we reconsider the compensation relation and formulate desirable properties. These considerations lead us to giving an equivalent formulation of the model just in terms of relations, which we do in Section 6.4. We conclude the chapter in Section 6.5 with some final thoughts and remarks.

6.1 A case-based reasoning explanation method

In this section we detail the workings of the dimension-based AF-CBA method of explanation developed by Prakken and Ratsma (2022), which was inspired by the work by Čyras et al. (2019). A more detailed comparison between the similarities of these, and other related works, was given by Prakken and Ratsma (2022, Section 8). The AF-CBA method is built upon Horty's (2019) DRM, and conceptually tries to mimic the arguments relating to precedent used by lawyers with respect to case law. In such discussions, precedent cases are cited by both sides as a means of arguing that the present (focus) case should be decided similarly as the precedent. Both sides may attack the other's citations, by pointing to important differences between the citation and the focus case; and they may defend themselves against such attacks, by pointing to aspects of the focus case which compensates for these differences. Each of the elements of such a discussion—case citations, pointing to differences, and compensating for differences—has its counterpart in the AF-CBA method.

A key idea underlying the approach is that a tabular dataset for binary classification can be interpreted as a case base $\mathscr{C}$. The method assumes access to the training data used by the system, and interprets each of the features in the data as a dimension. The corresponding dimension orders may be determined by knowledge engineering, statistical methods, or a combination thereof. This gives us a body of precedents $\mathscr{C}$ upon which the machine learning system bases its decisions.

Under this interpretation the machine learning system can be seen as deciding new fact situations for one of two sides. The goal is to explain a particular decision of a fact situation X for a side s, called the *focus case* (X, s). This explanation is provided in the form of a *best citable precedent* $(Y, s) \in \mathscr{C}$ together with an *explanation dialogue* in which the choice for this case (Y, s) is justified. Such an explanation is formalized as a winning strategy in the grounded argument game of a particular abstract argumentation framework, which we recall below in Definition 6.7.

Before we can apply the theory of precedential constraint, we need a method of determining the dimension orders. Prakken and Ratsma (2022) have developed a statistical method of doing so, of which a modified version is used by van Woerkom et al. (2024a).

Given a dataset that we would like to interpret as a case base, these methods apply concepts from statistics to the dataset in order to estimate the influence of the values of the dimensions on the outcomes. We will give a detailed description of these methods in the next chapter in Section 7.1.1.

Once the dimension orders are determined, the data can be interpreted as a case base, on the basis of which explanations can be generated. Any explanation dialogue should start with the citation of a best citable case. A suggestion for the definition of this notion is given by Prakken and Ratsma (2022) and we recall it in Section 6.1.2, after which we explain and motivate the presence of the arguments occurring in the argumentation framework in Sections 6.1.3 and 6.1.4. We are then ready to give the formal definition of the framework in Section 6.1.5, explain what it means to have a winning strategy in the argument game it induces, and as such what constitutes an explanation according to the model.

6.1.1 Recalling the dimension-based result model

A *dimension* d is a nonempty set together with a partial order $\preceq$ on d; we denote dimensions by lower case letters d, e, f, etc. The domain is modeled by a finite set of dimensions D. A *fact situation* X is a choice function on D, i.e. a function $X : D \to \bigcup D$ such that $X(d) \in d$ for every $d \in D$. A *case* (X, s) is a fact situation X paired with an outcome $s \in \{0, 1\}$. A *case base* is a finite set of cases.

We will frequently omit explicit reference to the dimension order $\preceq$ and instead refer to just the set d when we speak of a dimension. The order $\preceq$ of a dimension d specifies the relative preference the elements of d have towards either of two outcomes 0 and 1. More specifically, if $v \prec w$ for $v, w \in d$ this means w prefers outcome 1 relative to v, and conversely v prefers outcome 0 relative to w. Usually we want to compare preference towards an arbitrary outcome s, so to do this we define for any dimension $(d, \preceq)$ the notation $\preceq_s = \preceq$ if $s = 1$ and $\preceq_s = \succeq$ if $s = 0$.

Next we recall Definitions 3.3 and 3.5. Given fact situations Y and X we say X is *at least as strong* as Y for an outcome s, denoted $Y \preceq_s X$, if it is at least as strong for s on every dimension d:

$$Y \preceq_s X \quad \text{if and only if} \quad Y(d) \preceq_s X(d) \text{ for all } d \in D.$$

The decision of a fact situation X is *forced* for an outcome s by a case base $\mathscr{C}$, denoted $\mathscr{C}, X \vDash s$, if there is a case $(Y, s) \in \mathscr{C}$ with $Y \preceq_s X$.

Lastly, we will add some notation needed for the rest of the definitions of the AF-CBA method, which is related to partial fact situations. Recall from Section 3.4 that a partial fact situation is a partial choice function on D. Given a function $f : X \to Y$ and a subset $Z \subseteq X$ we write $f \restriction Z$ for the function $f \restriction Z : Z \to Y$ satisfying $(f \restriction Z)(z) = f(z)$ for all $z \in Z$.

Let X and Y be fact situations, and s an outcome. We define

$$W_s(Y, X) = X \restriction \{d \in D \mid Y(d) \npreceq_s X(d)\},$$
$$B_s(Y, X) = X \restriction \{d \in D \mid Y(d) \preceq_s X(d)\}.$$

In other words, $W_s(Y, X)$ is X restricted to the set of dimensions on which it is worse than Y with respect to outcome s. Similarly, $B_s(Y, X)$ corresponds to the part of X on which X is at least as good as Y with respect to outcome s. Note that $W_s(Y, X) \cup B_s(Y, X) = X$.

Example 6.1. Continuing Example 3.4, we consider the recidivism related dimensions:

$$\begin{aligned}\textbf{Age} &= (\{18,19,20,\dots\},\geq),\\ \textbf{Priors} &= (\{0,1,2,\dots\},\leq),\\ \textbf{Male} &= (\{0,1\},\leq).\end{aligned}$$

Suppose we have two cases $(Y,1)$ and $(X,1)$ with fact situations defined by

$$\begin{aligned} Y(\textbf{Age}) &= 45, & Y(\textbf{Priors}) &= 4, & Y(\textbf{Male}) &= 1,\\ X(\textbf{Age}) &= 50, & X(\textbf{Priors}) &= 5, & X(\textbf{Male}) &= 1.\end{aligned}$$

Now we can compute that $W_1(Y,X) = \{(\textbf{Age},50)\}$ and $B_1(Y,X) = \{(\textbf{Priors},5),(\textbf{Male},1)\}$.

6.1.2 Case citability

An important aspect of AF-CBA is the selection of the precedent case (Y,s) with which it initiates its explanation of the outcome of the focus case (X,s). We will now describe how this selection procedure works; later in Section 6.2 we return to this topic to suggest improvements. We begin with the notion of case citability.

Definition 6.2. A case (Y,s) is *citable* for a case (X,s) if $B_s(Y,X)$ is nonempty.

Since this is a quite weak requirement there may in general be very many citable cases (Y,s) for any given focus case (X,s). For this reason the notion is strengthened by requiring that (Y,s) should have a minimal number of relevant differences with (X,s), according to some suitable notion of minimality. To formalize this we first recall the definition of relevant differences used by Prakken and Ratsma (2022, Definition 11).

Definition 6.3. Given fact situations X,Y and an outcome s, the set $D_s(Y,X)$ of *relevant differences* between Y and X, with respect to outcome s, is given by

$$D_s(Y,X) = Y\restriction\{d\in D \mid Y(d)\not\preceq_s X(d)\}.$$

In other words, the relative differences between Y and X, with respect to s, correspond to that part of Y on which X is not at least as good as (or better than) Y. The set $D_s(Y,X)$ is comparable to the set $W_s(Y,X)$, except that the former is a restriction to Y and the latter is a restriction to X.

Now a best citable precedent is defined as a citable precedent which additionally minimizes this set of differences, in the following sense.

Definition 6.4. A case (Y,s) is a *best citable* case for a case (X,s) if

(a) (Y,s) is citable for (X,s), and

(b) any other case (Z,s) which is citable for (X,s) satisfies $D_s(Z,X)\not\subset D_s(Y,X)$.

6.1.3 Compensation of relevant differences

An idea central to the explanation dialogues is that when a precedent (Y,s) does not force a focus case (X,s), i.e. it does not hold that $Y\preceq_s X$, the values $W_s(Y,X)$ on which X is worse

than Y for s can be *compensated* for by the values $B_s(Y, X)$ on which X is better than (or equal to) Y. This idea is often encountered in the literature on case-based reasoning; in the words of Aleven (2003), these compensations "[show] that at a more abstract level, a parallel exists between the cases, arguing in effect that the apparent distinction is merely a mismatch of details."

In our context we assume the existence of a relation SC on partial fact situations V, U, where $SC(V, U)$ says that V compensates for U. This is used in practice as follows. Consider a precedent (Y, s) and a focus case (X, s). If (Y, s) forces the decision of (X, s) then $Y \preceq_s X$, so $\emptyset = W_s(Y, X)$, or equivalently $B_s(Y, X) = X$. If (Y, s) does not force the decision of (X, s), then $\emptyset \subset W_s(Y, X)$, or equivalently $B_s(Y, X) \subset X$. In order for the precedent case (Y, s) to be said to justify the outcome of (X, s) we should have that $B_s(Y, X)$ compensates for $W_s(Y, X)$, as determined by whether the relation $SC(B_s(Y, X), W_s(Y, X))$ holds.

6.1.4 Opposing citations and case transformations

The last component of the explanation dialogue generated by the AF-CBA method is that of opposing citations. The idea is that a proponent of the decision of (X, s), who cites a precedent case (Y, s) to argue for its outcome s, can have their citation countered by the citation of a case $(Z, \bar{s})$ by a proponent, where $\bar{s}$ denotes the opposite outcome of s. This counter-citation corresponds to the claim that the precedent case $(Z, \bar{s})$ is a more appropriate precedent to draw on with respect to the focus case (X, s). This is analogous to the argument between lawyers in a legal case.

Definition 6.5. A case (Y, s) can be *transformed* into a case (Z, s), relative to a focus case (X, s) and a compensation relation SC relative to a compensation relation SC, iff $Y = Z$ or $U = W_s(Y, X) \neq \emptyset$ and there exists $V \subseteq B_s(Y, X)$ such that $SC(V, U)$, and $Z = (Y \setminus Y \restriction \mathrm{dom}(U)) \cup U$.

The goal of these transformations is to change (Y, s) into a case (Z, s) that forces the outcome of (X, s). It does so by replacing the values of the precedent case with those of the focus case, on those dimensions on which the focus case is not at least as strong as the precedent.

6.1.5 An abstract argumentation framework for explanation

We are now ready to give a formal account of AF-CBA through the use of an *abstract argumentation framework*—a concept introduced by Dung (1995). An abstract argumentation framework $AF = (\mathrm{Arg}, \mathrm{Attack})$ is a directed graph, in which the nodes are interpreted as arguments and the edges as an attack relation between them.

A particular argumentation framework $(\mathrm{Arg}, \mathrm{Attack})$ is used by Prakken and Ratsma (2022) that combines the types of arguments defined in the preceding Sections 6.1.2, 6.1.3, and 6.1.4, relative to a *focus case* (X, s). To do so we first define, for a particular precedent (Y, s) that may be cited in defense of the decision of X for s, a subset $\mathbb{A}_{(Y,s)} \subseteq \mathrm{Arg}$ as follows:

$$\mathbb{A}_{(Y,s)} = \bigcup \Big\{ \{\mathrm{worse}_{(Y,s)}(U) \mid U = W_s(Y, X) \neq \emptyset\}, \quad (6.1.1)$$
$$\{\mathrm{compensates}_{(Y,s)}(V, U) \mid \mathrm{worse}_{(Y,s)}(U) \in \mathbb{A}_{(Y,s)},\ V \subseteq B_s(Y, X),\ SC(V, U)\},$$
$$\{\mathrm{transformed}_{(Y,s)}(Z) \mid (Y, s) \text{ can be transformed into } (Z, s) \text{ with } Z \preceq_s X\} \Big\}.$$

Definition 6.6. Given a finite case base $\mathscr{C}$, a focus case (X, s), and a compensation relation SC, an *abstract argumentation framework for explanation with dimensions* is a pair $AF = (\text{Arg}, \text{Attack})$ where the arguments Arg are given by

$$\text{Arg} = \mathscr{C} \cup \bigcup \{\mathbb{A}_{(Y,s)} \mid (Y, s) \in \mathscr{C}\},$$

and for arguments $A, B \in \text{Arg}$ we have $\text{Attack}(A, B)$ if and only if either:

- $(Y, s), (Z, \bar{s}) \in \mathscr{C}$ and $\{d \in D \mid Y(d) \not\preceq_s X(d)\} \not\subset \{d \in D \mid Z(d) \not\preceq_s X(d)\}$;
- $(Y, s) \in \mathscr{C}$ and A is of the form $\text{worse}_{(Y,s)}(U)$;
- B is of the form $\text{worse}_{(Y,s)}(U)$ and A is of the form $\text{compensates}_{(Y,s)}(V, U)$; or
- $(Z, \bar{s}) \in \mathscr{C}$ and A is of the form $\text{transformed}_{(Y,s)}(W)$.

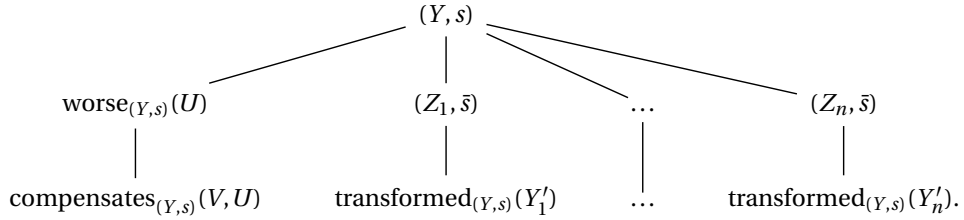
A dialogue now takes the form of a *grounded argument game* played on $(\text{Arg}, \text{Attack})$. For the sake of brevity we only give an intuitive explanation of how this works, the reader is referred to Prakken and Ratsma (2022) for a detailed treatment of the subject.

An argument game on an AF (A, R) is a two-player game, in which the players take turns playing arguments from A which must attack the previously played argument according to the attack relation R. A player can win the game by moving an argument to which the other player cannot reply, and a *winning strategy* for a player is a method of playing that ensures a win regardless of how the opponent plays.

We can now formally define the explanations generated by the AF-CBA method.

Definition 6.7. An *explanation* of a focus case (X, s) is a winning strategy in the grounded argument game starting with the citation of a best citable precedent $(Y, s) \in \mathscr{C}$, played on the abstract argumentation framework for explanation with dimensions $(\text{Arg}, \text{Attack})$.

The winning strategies may be viewed as trees and have the following general shape:



6.2 Best citable cases

We now turn to some suggestions for modifications of Definition 6.4 that might be closer to the intuitive notion of a most closely related case (Y, s) of our focus case (X, s).

Firstly, since Definition 6.3 does not gather just the dimensions on which (X, s) is worse than (Y, s) but also the value of (Y, s) at that dimension, a situation can arise where there is some case (Z, s) with $\{d \in D \mid Z(d) \not\preceq_s X(d)\} \subset \{d \in D \mid Y(d) \not\preceq_s X(d)\}$ but $Z \restriction \{d \in D \mid Z(d) \not\preceq_s X(d)\} \not\subset Y \restriction \{d \in D \mid Y(d) \not\preceq_s X(d)\}$, just because there is some dimension $d \in \{d \in D \mid Z(d) \not\preceq_s X(d)\}$ with $Z(d) \neq Y(d)$. It does not seem correct to dismiss (Z, s) as a good citation simply because it disagrees with (Y, s) on a single dimension,

especially when $\{d \in D \mid Z(d) \not\preceq_s X(d)\}$ is only a very small subset of $\{d \in D \mid Y(d) \not\preceq_s X(d)\}$. Let us look at an example to illustrate this point.

Example 6.8. We consider three cases $(Y, s), (Z, s), (X, s)$ (so they were all judged a high risk of recidivism) in the recidivism scenario of Example 6.1:

$$\begin{aligned} Y(\textbf{Age}) &= 20, & Y(\textbf{Male}) &= \text{M}, & Y(\textbf{Priors}) &= 3, \\ Z(\textbf{Age}) &= 50, & Z(\textbf{Male}) &= \text{M}, & Z(\textbf{Priors}) &= 1, \\ X(\textbf{Age}) &= 40, & X(\textbf{Male}) &= \text{M}, & X(\textbf{Priors}) &= 2. \end{aligned}$$

We have that $D_s(Y, X) = \{(\textbf{Age}, 20), (\textbf{Priors}, 3)\}$ and $D_s(Z, X) = \{(\textbf{Priors}, 1)\}$. Therefore, even though there are fewer dimensions on which (Z, s) has relevant differences with (X, s)—as $\{\textbf{Priors}\} \subset \{\textbf{Age}, \textbf{Priors}\}$—this does not prevent (Y, s) from being considered a best citable precedent for (X, s)—as $\{(\textbf{Priors}, 1)\} \not\subset \{(\textbf{Age}, 20), (\textbf{Priors}, 3)\}$.

This consideration suggests the definition should require minimality of $\{d \in D \mid Y(d) \not\preceq_s X(d)\}$ instead of $Y \restriction \{d \in D \mid Y(d) \not\preceq_s X(d)\}$. However, this modification leaves room for a second type of scenario where there is some precedent (Z, s) which is intuitively much closer to the focus case relatively to some other (Y, s), without hindering (Y, s) from being considered best citable. To see why we consider a set of $n+1$ dimensions $\{d_0, \ldots, d_n\}$. Now we may have that $\{d \in D \mid Z(d) \not\preceq_s X(d)\} = \{d_0\}$ and $\{d \in D \mid Y(d) \not\preceq_s X(d)\} = \{d_1, \ldots, d_n\}$. This means that the presence of (Z, s) does not hinder (Y, s)'s being considered a best citable precedent for (X, s), even though (X, s) is worse than (Y, s) on n times as many dimensions as it is worse on than (Z, s). To remedy this, we could require minimality of the number of dimensions rather than the set of dimensions itself, i.e. of $|\{d \in D \mid Y(d) \not\preceq_s X(d)\}|$.

In addition to looking just at differences between the precedent and focus case it may be beneficial to also consider their similarities, since after all, the *stare decisis* doctrine states that similar cases must be decided similarly. To achieve this we can require the best citable precedent to subsequently maximize $|\{d \in D \mid Y(d) = X(d)\}|$, so that it both minimizes differences and maximizes similarities. In all, this leads us to the following definition.

Definition 6.9. A case (Y, s) is a *best citable* case for a case (X, s) if it satisfies the conditions

(a) (Y, s) is citable for (X, s);

(b) there is no other (Z, s) satisfying (a) with $|\{d \in D \mid Z(d) \not\preceq_s X(d)\}| < |\{d \in D \mid Y(d) \not\preceq_s X(d)\}|$;

(c) there is no other (Z, s) satisfying (a) and (b) with $|\{d \in D \mid Z(d) = X(d)\}| > |\{d \in D \mid Y(d) = X(d)\}|$.

The experimental results of Prakken and Ratsma (2022) showed that there are in general many cases satisfying Definition 6.4 for any (X, s). Measured on three datasets, the mean and standard deviation of the number of best citable cases were respectively 82 ± 123.6, 76 ± 134, and 106 ± 116.5 (Prakken & Ratsma, 2022, Table 5). Recalculating these statistics for the same datasets with Definition 6.9 instead results in respectively 5.6 ± 2.0, 2.1 ± 2.6, and 2.6 ± 2.5 average number of best citable cases; a substantial decrease. Still, the definition remains somewhat ad-hoc, and more research is needed to assess its adequacy in actual applications.

6.3 Specifying the compensation relation

Prakken and Ratsma (2022) do not make further explicit assumptions of the compensation relation *SC*—which is why their method is considered top-level. However, in order for this relation to function according to our intuitions it may be necessary to do so, and we now consider a few such requirements. Let us first illustrate *SC* through a continuation of Example 6.8.

Example 6.10. We saw two example cases $(Y, s), (Z, s)$ where (Z, s) was worse than (Y, s) on the dimensions **Age** and **Male**, but better on **Priors**. Suppose that for a number of priors higher than 4, we no longer care about values besides the number of priors. Then we may define

$$SC(V, U) \quad \text{if and only if} \quad V(\textbf{Priors}) \geq 4.$$

In this case the worse values $W_s(Y, Z)$ would indeed be compensated for by the better values $B_s(Y, Z)$, since $Z(\textbf{Priors}) = 5$.

Another point to consider is whether the compensation relation should itself adhere to an a fortiori principle. That is to say, if a set V is capable of compensating for a set U, should a superset $W \supseteq V$ be capable of compensating for U as well? This property is captured by the following definition.

Definition 6.11. A compensation relation *SC* is *monotone* if for any partial fact situations U, V, W it holds that $SC(V, U)$ implies $SC(V \cup W, U)$.

The same goes for values that are being compensated for; if a set V can compensate for a set U then we might require of it to compensate any subset $W \subseteq U$ as well.

Definition 6.12. A compensation relation *SC* is *antitone* if for any partial fact situations U, V, W it holds that $SC(V, U)$ implies $SC(V, U \cap W)$.

In the factor-based version of AF-CBA, i.e. in the special case where the dimensions are all two-element sets with a linear order, it is possible to compensate for a set of worse values in parts through the use of a pSubstitutes(V, U, X, s)&cCancels(V', U', X, s) move (Prakken & Ratsma, 2022, Definition 5). We can translate this to the dimensional setting as follows.

Definition 6.13. A compensation relation *SC* is *linear* if for any partial fact situations T, U, V, W it holds that $SC(T, U)$ and $SC(V, W)$ imply $SC(T \cup V, U \cup W)$.

A more fundamental question regarding the compensation relation is that of *context dependence*; should the compensation of two sets be allowed to depend on the context in which it takes place? This question and its consequences are the subject of Section 6.4.

6.4 Justification as an extension of forcing

An interesting way to think of the compensation relation is as an extension of the notion of forcing between cases (Definition 3.5). In essence a compensation says that while a precedent (Y, s) might not force the decision of some other case (Z, s), the obstructing relevant differences can be compensated, and so the precedent (Y, s) may still be said to *justify* the outcome of (Z, s).

6.4.1 Context-dependent compensations

A downside of the formal specification of this compensation relation is that it is defined on partial fact situations, rather than just fact situations. This makes it impossible for compensations to take the values of the precedent into account when allowing compensations to be made.

Example 6.14. In Example 6.1 the difference in age between (Y, s) and (Z, s) is only 5, and we may want to say that $B_s(Y, Z)$ compensates for $W_s(Y, Z)$ in this case if we find this difference small enough to be insignificant. To make this compensation possible formally we would need to postulate $SC(\{(\textbf{Age}, 50)\}, \{(\textbf{Priors}, 5), (\textbf{Male}, \text{M})\}$ but this would inadvertently sanction compensations where the age of the precedent case is, say, 20, in which case we may find the difference in age large enough to be significant.

Modifying SC so that it takes the precedents' values into account yields a relation on full fact situations. A natural requirement of any such relation is that it *extends* the strength order $\preceq$ of Definition 3.5. This is akin to saying that any set can compensate for the empty set. This leads us to the following definition.

Definition 6.15. A relation $\sqsubseteq$ on cases is called a *justification* relation if it extends the forcing relation $\preceq$, i.e. if $\preceq \subseteq \sqsubseteq$.

Note that any compensation relation SC gives rise to a justification relation $\sqsubseteq_{SC}$:

$$(Y, s) \sqsubseteq_{SC} (Z, s) \quad \text{if and only if} \quad Y \preceq_s Z \text{ or } SC(B_s(Y, Z), W_s(Y, Z)). \tag{6.4.1}$$

The converse does not hold, precisely because a justification relation takes into account the *context* of the compensation. To see this, consider the naive approach of obtaining a compensation relation $SC_{\sqsubseteq}$ from a justification relation $\sqsubseteq$:

$$SC_{\sqsubseteq}(V, U) \quad \text{if and only if} \quad (Y, s) \sqsubseteq (Z, s) \text{ for } Y, Z \text{ with } U = W_s(Y, Z), V = B_s(Y, Z). \tag{6.4.2}$$

The problem is that this definition is not necessarily *well-defined*, meaning that the truth value of $SC_{\sqsubseteq}(V, U)$ may depend on the particular representatives Y and Z that are used for its evaluation. This leads us to define the notion of a context-independent $\sqsubseteq$, requiring exactly that the relation $SC_{\sqsubseteq}$ above is well defined.

Definition 6.16. A justification relation $\sqsubseteq$ is *context-independent* with respect to an outcome s, if for any four fact situations Y, Z, W, X with $W_s(Y, Z) = W_s(W, X)$ and $B_s(Y, Z) = B(W, X)$ it holds that $(Y, s) \sqsubseteq (Z, s)$ iff $(W, s) \sqsubseteq (X, s)$.

6.4.2 Winning strategies and justification

The terminology of Definition 6.15 is inspired by Prakken and Ratsma (2022), where an argument is said to be justified if and only if the proponent has a winning strategy in the grounded argument game about the argument. We will now formally justify this comparison by showing that for any compensation relation SC the proponent of an initial citation (Y, s) has a winning strategy in the game on the argumentation framework if and only if $(Y, s) \sqsubseteq_{SC} (X, s)$ (of Eq. (6.4.1)).

We fix a precedent case (Y, s) and a focus case (X, s), and introduce some terminology for convenience. We will say a case (Y, s) has a winning strategy if the proponent has a winning strategy in the grounded argument game on the explanation AF (Arg, Attack) of Definition 6.6, starting with a citation of (Y, s). Following Prakken and Ratsma (2022) we distinguish between *nontrivial* winning strategies for (Y, s), in which (Y, s) can be attacked by a $\text{worse}_{(Y,s)}(U)$ move, and *trivial* winning strategies for (Y, s), in which there is no $\text{worse}_{(Y,s)}(U)$ attack possible. This means a winning strategy for (Y, s) is nontrivial if $\text{worse}_{(Y,s)}(U) \in \mathbb{A}_{(Y,s)}$ and trivial if $\text{worse}_{(Y,s)}(U) \notin \mathbb{A}_{(Y,s)}$, with $\mathbb{A}_{(Y,s)}$ as in Eq. (6.1.1).

Proposition 6.17. *There is a trivial winning strategy for* (Y, s) *if and only if* $Y \preceq_s X$.

Proof. Note that $\text{worse}_{(Y,s)}(U) \notin \mathbb{A}_{(Y,s)}$ iff $W_s(Y, X) = \emptyset$ iff $Y \preceq_s X$. Hence left to right is immediate. For right to left we note in addition that any citation made by the opponent can be attacked with a $\text{transformed}_{(Y,s)}(Y, s)$ move, and so since there is no reply possible to a Transformed move the proponent has a (trivial) winning strategy for (Y, s). □

Proposition 6.18. *There is a nontrivial winning strategy for* (Y, s) *if and only if* $W_s(Y, X) \neq \emptyset$ *and* $SC(B_s(Y, X), W_s(Y, X))$.

Proof. Suppose the proponent has a winning strategy. Since $\text{worse}_{(Y,s)}(U) \notin \mathbb{A}_{(Y,s)}$ attacks the initial citation of (Y, s) there should be a $\text{compensates}_{(Y,s)}(V, U)$ response to the $\text{worse}_{(Y,s)}(U)$ move available to the proponent, with $V = B_s(Y, Z)$. This implies that $SC(B_s(Y, X), W_s(Y, X))$.

For the other direction we begin by noting that because $W_s(Y, Z) \neq \emptyset$ there is $\text{worse}_{(Y,s)}(U) \in \mathbb{A}_{(Y,s)}$, and so the assumption $SC(B_s(Y, X), W_s(Y, X))$ guarantees that there is $C = \text{compensates}_{(Y,s)}(V, U) \in \mathbb{A}_{(Y,s)}$. Now, there are two types of moves available to the opponent to which we need a reply.

1. The first is $\text{worse}_{(Y,s)}(U) \in \mathbb{A}_{(Y,s)}$. As mentioned we have a reply C available, and since a compensation move cannot be replied to the game is won by the proponent.
2. The second is the citation of a case $(Z, \bar{s}) \in \mathscr{C}$ for which it holds that $\{d \in D \mid Z(d) \not\preceq_s X(d)\} \not\subset \{d \in D \mid Y(d) \not\preceq_s X(d)\}$. By Definition 6.5 we have that (Y, s) can be transformed into (Y', s), and so we can reply to the citation with $\text{transformed}_{(Y,s)}(q) \in \mathbb{A}_{(Y,s)}$. There are no more moves available to the opponent and so the proponent wins the game. □

Corollary 6.19. *There is a winning strategy for* (Y, s) *if and only if* $(Y, s) \sqsubseteq_{SC} (X, s)$.

Proof. Applying Eq. (6.4.1) and then Propositions 6.17 and 6.18 we get

$$\begin{aligned}(Y, s) \sqsubseteq_{SC} (X, s) \text{ iff } & Y \preceq_s X \text{ or } SC(B_s(Y, X), W_s(Y, X)) \\ & \text{iff } (Y, s) \text{ has a (non)trivial winning strategy} \\ & \text{iff } (Y, s) \text{ has a winning strategy.}\end{aligned}$$

□

Under this view of the winning strategies, and employing a fully general definition of compensation through a justification relation $\sqsubseteq$, we can now rephrase Definition 6.7 of explanations in the following way.

Definition 6.20. An *explanation* of a case (X, s) is a best citable precedent $(Y, s) \in \mathscr{C}$ with $(Y, s) \sqsubseteq (X, s)$.

The theory of precedential constraint describes how the outcome of a fact situation can be forced by precedent. However the collection of precedents may not be sufficient to force the outcome of all possible new fact situations. If such an undecided fact situation presents itself there may still be a precedent which, on the basis of additional reasoning, can be argued to *justify* an outcome for the fact situation. This is the view suggested by Corollary 6.19; a justification relation goes beyond the forcing relation by sanctioning citations of precedents that do not strictly force the outcome of the focus case.

6.4.3 A relational description of the explanation model

Corollary 6.19 shows that a justification relation corresponds to winning strategies underlying the explanations of AF-CBA, and this allows us to give a succinct description of the explanation method just through the use of relations on cases. Let us think of citability as a relation $\trianglelefteq$, then those $(Y, s) \in \mathscr{C}$ related to the focus case through the intersection $\sqsubseteq \cap \trianglelefteq$ with (X, s) are said to explain the focus case (X, s), i.e. those (Y, s) with $(Y, s) \sqsubseteq (X, s)$ and $(Y, s) \trianglelefteq (X, s)$.

The AF-CBA model is top-level as it does not give explicit definitions of these notions, apart from suggesting a definition for the citability relation $\trianglelefteq$ as in Definition 6.4, and a method for determining $\preceq$ on the basis of Pearson correlation coefficients. In its running example and the experiments of Prakken and Ratsma (2022, Section 6) all compensations are allowed, so that $\sqsubseteq \cap \trianglelefteq = \trianglelefteq$. These inputs are summarized through the relations as:

1. The forcing relation $\preceq$, determined by specifying the dimensions and their orders.
2. The justification relation $\sqsubseteq$, determined by specifying the compensations.
3. The citability relation $\trianglelefteq$, determined by the definition of a best citable precedent.

This view considerably simplifies the presentation of the model as it does not rely on the concepts of argumentation frameworks and winning strategies.

6.5 Discussion and conclusion

We have described the AF-CBA explanation model of Prakken and Ratsma (2022) in Section 6.1, which provides explanations as winning strategies on the grounded argument game of an abstract argumentation theory. In Section 6.4 we showed that this model admits an equivalent rephrasing in terms of the *justification*, *citability*, and *forcing* relations. In this view, explanations are provided as cases that are related to the focus case through these relations. Most notably this shows that the explanation model can in some sense be seen as adding a notion of justification to the theory of precedential constraint as a relation $\sqsubseteq$ extending the forcing relation $\preceq$.

As mentioned, the AF-CBA method discussed in this chapter is motivated by an analogy between the way in which a machine learning system draws on training data to assign a label to a new data point and the way in which a court of law draws on previously decided cases to make a decision about a new fact situation. In the next chapter, we will take a closer look at this analogy, by fitting the various versions of the result model which we studied in Part I of this thesis to datasets. We then consider how well the analogy matches up to the data by examining various statistics such as consistency.

Chapter 7

Case Base Consistency

USING THE DIMENSIONAL RESULT MODEL (DRM, cf. Chapter 3), Prakken and Ratsma (2022) developed a case-based reasoning method for explaining data-driven decisions, which was the topic of the previous chapter. In addition, they studied whether the DRM is useful as a tool for modeling AI datasets. The present chapter will be spent performing similar analyses, by concretely applying the DRM to various datasets and evaluating the degree to which the model fits the data, as measured in terms of *case-base consistency*—the relative frequency of consistent cases.

To do this we first use the reformulation of the model in terms of many-sorted logic, described in Section 3.10.2, to write a Python implementation on the basis of the Satisfiability Modulo Theories (SMT) solver Z3 (de Moura & Bjørner, 2008). In Section 7.1, we describe how this implementation works, and then use it to fit the model to various datasets. We measure this fit in terms of consistency. In addition we evaluate various questions such as: Is the dataset consistent and/or complete? If not, what is causing the inconsistency or incompleteness? If the dataset is inconsistent, how many of its cases are inconsistent? How many landmarks does the data contain, and what do they look like? We also compare different ways of automatically determining the dimension orders and the effect that they have on the aforementioned statistics. Some of these datasets have known ground truth labels, which allows us to analyze exactly how well the model fits the data.

After having presented our SMT-based implementation, we put it to work in Section 7.2 and 7.3 by fitting the DRM to the well-known COMPAS recidivism dataset published by Angwin et al. (2016), as well as on several variations of this dataset. This dataset consists of real-world data which is representative of the domain on which we would like to apply XAI methods based on the a fortiori model. As such, the results of this experiment are indicative of the feasibility of such XAI methods.

Subsequently, in Sections 7.4, we consider datasets used by Steging et al. (2021, 2023). We use these because they have known ground truth labels, which allows us to precisely evaluate the model's fit to the data. An overview of our findings is given in Table 7.1. We end with some concluding remarks in Section 7.5.

Dataset	Size	Pearson Corr.		Logistic Regr.	
		$\|\mathscr{L}\|$	Cons.	$\|\mathscr{L}\|$	Cons.
Churn	7,010	1,259	59.2%	6,009	95.6%
Admission	500	41	80.2%	90	91.2%
Mushrooms	8,124	23	98.8%	23	100%
COMPAS (full)	5,873	88	8.1%	–	–
COMPAS (simp.)	1,342	12	4.2%	–	–
CORELS	907	6	100%	–	–
Tort	1,024	18	98.6%	–	–
Welfare (full)	99,988	634	71.1%	462	48.5%
Welfare (simp.)	32,876	10	67.3%	5	66%

Table 7.1: An overview of the various datasets used in our experiments. For each dataset we list its size, number of landmarks, and its consistency percentage. We do this for both the Pearson correlation and logistic regression methods for determining the dimension orders. In some cases both methods produce the same dimension orders, which means that all statistics will also be the same; such duplicate statistics are replaced by dashes.

7.1 Implementing the dimensional result model using Z3

In this section we describe how the DRM can be implemented in Python using the SMT solver Z3 (de Moura & Bjørner, 2008). In order to be able to compute with the a fortiori model we require two main components. First—to construct the model—we need a method for determining the dimension orders. Secondly, we need a way to operationalize it, so that we can compute, for example, whether some new fact situation has its outcome forced by a case base. Other necessary ingredients like data representations can be handled with built-in Python functionality. We describe the method for determining the dimension orders in Section 7.1.1, and then how we use Z3 to operationalize the model in Section 7.1.2. In Section 7.1.3 we discuss a preliminary test of the resulting implementation, by recomputing the statistics reported by Prakken and Ratsma (2022).

7.1.1 Determining dimension orders

Determining appropriate orders for the dimensions is not a straightforward task. They constitute an assumption that the values along the dimension tend to prefer either of the binary outcomes. For instance, in our example with recidivism data we have an **Age** dimension, and to determine its order is to say whether the elderly are more likely to recidivate than the young, or vice versa. Knowledge engineering techniques and statistical methods can be used for this purpose. For instance, for the **Age** dimension, much has been written on the interplay between age and recidivism, the conclusion of which is summarized by the adage that “people age out of crime," meaning that as people age they become decreasingly likely to recidivate. Another option is to look at statistical trends in available data, for instance, by considering the sign of the Pearson correlation between age and recidivism. If it is positive, we can say that likelihood of recidivism increases with age,

and if it is negative, we can say it decreases.

For our implementation, we employ the statistical method. We will use the same underlying idea as used by Prakken and Ratsma (2022), which is to use a function c that associates each numerical feature x with a *coefficient* $c(x)$ indicating the degree to which the values of x favor outcome 1. If $c(x)$ is positive we order the values of x with the usual less-than order $\leq$ on the number line, and if it is negative we order it using the greater-than order $\geq$; so more precisely $\preceq := \leq$ if $c(x) \geq 0$ and $\preceq := \geq$ if $c(x) < 0$.

If x is categorical we cannot apply c directly so we use *dummy variables*. More specifically, if x is a categorical feature that can take the possible (unordered) values $v_1, \ldots, v_n$, then we introduce for each value v_i a dummy variable d_{v_i} which is a binary feature indicating whether $x = v_i$. Then we define $v_i \preceq v_j$ if and only if $c(d_{v_i}) \leq c(d_{v_j})$.

Prakken and Ratsma (2022) define c on the basis of Pearson correlation, but for the present work we define c using logistic regression. Supposing we have features $x_1, \ldots, x_n$ the logistic model has parameters $\beta_0, \ldots, \beta_n$, and models the probability that a given sample belongs to class 1 by the formula

$$p(x_1, \ldots, x_n) = \frac{1}{1 + e^{-(\beta_0 + \sum_i \beta_i x_i)}}. \tag{7.1.1}$$

We find appropriate values for the β parameters using the scikit-learn implementation of a maximum likelihood estimation with default parameters (Pedregosa et al., 2011), and after this is done we can simply put $c(x_i) := \beta_i$.

As mentioned we opt to use logistic regression rather than Pearson correlation. There are several reasons for this. Firstly, logistic regression seems to be a better choice conceptually, since it optimizes the coefficients collectively rather than compute them independently of one another. Secondly, logistic regression seems to perform better in practice, as we will demonstrate in the coming sections. Lastly, the method using Pearson correlation seems to work poorly with categorical features, as we will now illustrate.

Given n samples $(x_1, y_1), \ldots, (x_n, y_n)$ of binary variables x and y, the estimate of the Pearson correlation $r(x, y)$ between x and y is given by

$$r(x, y) = \frac{n\mathbb{1}_{xy} - \mathbb{1}_x \mathbb{1}_y}{\sqrt{n\mathbb{1}_x - \mathbb{1}_x^2}\sqrt{n\mathbb{1}_y - \mathbb{1}_y^2}}, \tag{7.1.2}$$

where $\mathbb{1}_x$ is the number of times x takes value 1 in the samples, $\mathbb{1}_y$ the number of times y takes value 1, and $\mathbb{1}_{xy}$ the number of times x and y both take value 1.

In order to get a sense of how this function behaves we plot its values for a fixed n and with $\mathbb{1}_y := n/2$, see Figure 7.1. This plot shows that when $\mathbb{1}_x$ is relatively low, or relatively high, the range of $r(x, y)$ (as a function of $\mathbb{1}_{xy}$) is not $[-1, 1]$ but some restricted interval near 0. More precisely, writing $s(x) := \sqrt{n\mathbb{1}_x - \mathbb{1}_x^2}$, we can calculate that for $0 \leq \mathbb{1}_x \leq n/2$ the range of $r(x, y)$ is $[-\mathbb{1}_x/s(x), \mathbb{1}_x/s(x)] \approx [-\mathbb{1}_x/(n/2), \mathbb{1}_x/(n/2)]$, i.e. is roughly proportional to $\mathbb{1}_x$. This is undesirable when x is a dummy variable, as then $\mathbb{1}_x$ simply indicates the number of times the original categorical feature took the value which the dummy variable represents, i.e. the number of samples we have of that class.

Example 7.1. Let us consider an example to illustrate this point. The original COMPAS data includes a **Race** variable, with possible values including "Asian" and "Caucasian." The value Asian occurs much less often than Caucasian (0.4% against 34%), meaning

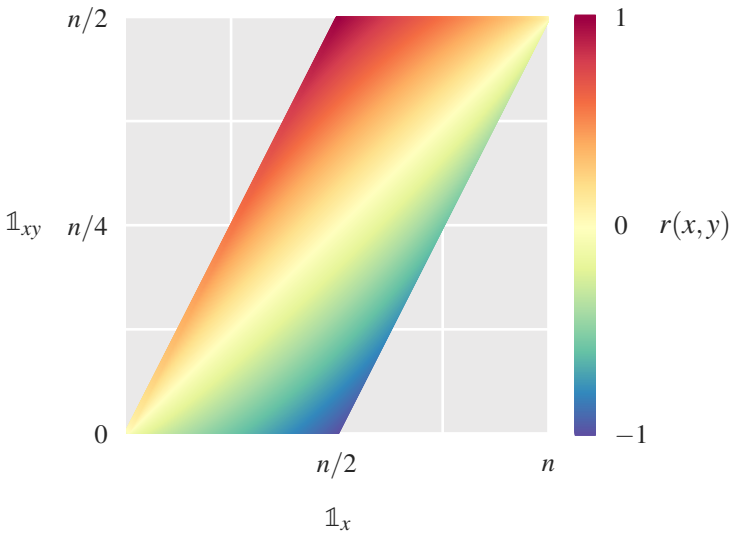


Figure 7.1: A plot of Eq. (7.1.2), the Pearson correlation coefficient for binary vectors x and y for a fixed value of $n := 400$ and with $\mathbb{1}_y := n/2$. The gray area marks points that violate one of the inequalities $\mathbb{1}_x + \mathbb{1}_y - n \leq \mathbb{1}_{xy} \leq \mathbb{1}_x$, and as such could not result from a sample.

that the value of $\mathbb{1}_x$ for the dummy variable for **Race** = Asian is much lower than that for the **Race** = Caucasian variable. As a result, its Pearson correlation must land in a very small interval around 0, while the one for Caucasian has almost the full range available. Indeed, the order for the **Race** dimension on the basis of the Pearson correlation method puts Caucasian in the last position (i.e. comparatively least prone to recidivate), and Asian a little over halfway in the order. To compare this with a measure that does not place such great importance on the number of samples that we have of each race, we consider the relative frequency $\mathbb{1}_{xy}/\mathbb{1}_x$ of recidivism within that class. The picture is now the opposite of what we see with Pearson correlation, with Asian ending lowest in the ranking (at 28% prevalence) and Caucasian a little over halfway (at 40% prevalence).

7.1.2 Using Z3 to operationalize the result model

Satisfiability Modulo Theories (SMT) is about procedurally checking the satisfiability of formulas over a theory, in the sense described in Section 3.10.1 of Chapter 3. In particular: given a many sorted signature Σ, a formula $\phi \in L^\Sigma$, a structure $\mathbb{A}$ and a theory $T \subseteq L^\Sigma$, an SMT solver is concerned with deciding whether there is a satisfying assignment $\alpha \in [\![\phi]\!]$ or not. Using the equivalence between validity of ϕ and unsatisfiability of $\neg\phi$ this means an SMT solver can also be used to check validity. The decidability and complexity of answering this satisfiability problem greatly depend on the theories in question. Bradley and Manna (2007, Section 3) have given an overview of some supported theories and their complexities.

In this section, we describe how we use Z3—a state-of-the-art SMT solver—to answer

questions for a given set D of dimensions and a case base $\mathscr{C}$, such as:

(1) Given two fact situations $F, G \in \mathscr{X}$ does $F \preceq_s G$ holds?

(2) For a fact situation F and an outcome s, does $\mathscr{C}, F \models s$ hold?

(3) Is $\mathscr{C}$ consistent and/or complete?

(4) Given a case $(F, s) \in \mathscr{C}$, is (F, s) a landmark of $\mathscr{C}$?

Answering these questions with Z3 is a relatively straightforward application of our work in Section 3.10.2. For instance, to answer question (1) we use the formula $G \stackrel{\circ}{=} x \wedge \phi_s(F)$ of Eqs. (3.10.1) and (3.10.3) because Z3 can determine whether there is a satisfying assignment in $[\![G \stackrel{\circ}{=} x \wedge \phi_s(F)]\!] = \{G\} \cap \uparrow_s F$, which is inhabited if and only if $F \preceq_s G$. Similarly, to see if F is forced by $\mathscr{C}$ for s we use the formula $F \stackrel{\circ}{=} x \wedge \Phi_s$, since $[\![F \stackrel{\circ}{=} x \wedge \Phi_s]\!] = \{F\} \cap \uparrow_s \mathscr{C}$ is inhabited if and only if $\mathscr{C}, F \models s$. To answer question (3) we can use Proposition 3.35; the case base is inconsistent if and only if $\Phi_0 \wedge \Phi_1$ is satisfiable, and incomplete if and only if $\neg(\Phi_0 \vee \Phi_1)$ is satisfiable. Finally, to check whether a case (F, s) is a landmark of $\mathscr{C}$ we can use the formula $\lambda_s(F)$ of Eq. (3.10.4), since $[\![\lambda_s(F)]\!]$ is inhabited if and only if $(F, s) \in \mathscr{L}$. The full implementation, and the rest of our code, is available online.[1]

7.1.3 A preliminary test of the implementation

As a first test of this implementation we repeated the experiments reported on by Prakken and Ratsma (2022, Section 6) on the Churn,[2] Mushroom (Schlimmer, 1981), and Admission datasets (Acharya et al., 2019). We do this so we can compare the output of our implementation to known results, and so that we can test the logistic-regression method for automatically learning appropriate dimension orders from the data. Note that all three of these datasets are, or at least appear to be, largely synthetic. The Churn dataset contains "information about a fictional telco company that provided home phone and Internet services to 7,043 customers in California in Q3." The Mushroom dataset contains "descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family." Lastly, the Admission dataset contains information about the chance of university admission on the basis of data like undergraduate grade point average. Again, this dataset seems to contain at least some synthetic elements, as its author writes that it had values "entered manually with no specific pattern. It was random assignment." A more extensive description of these datasets and their features can be found in the work by Prakken and Ratsma (2022).

We report the findings of our implementation in Table 7.1, which can be compared to the results found by Prakken and Ratsma (2022, Table 3). We list the number of landmarks $|\mathscr{L}|$ as well as the consistency percentage, which is computed as the relative frequency of consistent cases in the dataset:

$$\mathrm{Cons}(\mathscr{C}) := 100 \cdot \left(1 - \frac{|\mathscr{C}_0 \cap \uparrow \mathscr{C}_1| + |\mathscr{C}_1 \cap \downarrow \mathscr{C}_0|}{|\mathscr{C}|}\right).$$

We find an identical consistency percentage for the Mushrooms dataset, but only approximately equal percentages for the Churn and Admission datasets. The difference

[1] https://git.science.uu.nl/ics/responsible-ai/van-woekom/afcbr.

[2] https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113.

in the percentage for Churn is because Prakken and Ratsma did not delete duplicate occurrences of cases. We did delete duplicate cases for the sake of our landmark analysis; if two cases have identical fact situations and outcomes, but are not considered equal, then they will 'force' each other's outcome and so are not considered landmarks when they otherwise might have been. The consistency percentage on the Admission dataset also differs, even though the number of cases there is equal. It is not entirely clear why this is. Since the difference in percentages is small—only 0.4%—and the results are otherwise in agreement, we do not further investigate the source of this difference.

As we can see, the approach using logistic regression tends to increase both the number of landmarks as well as the consistency percentage—in the case of the Churn dataset by as much as 36.4%. This suggests to us that logistic regression is indeed a better method for the purpose of automatically assigning dimension orders to the features.

In the next few sections we will perform similar analyses of consistency percentages and landmark cases of various datasets, starting with the COMPAS dataset published by Angwin et al. (2016).

7.2 The COMPAS dataset

We turn our attention to the COMPAS recidivism dataset, published by Angwin et al. (2016), which contains information on convicts and whether they recidivated within two years after being arrested for an initial charge. We chose this dataset because it consists of real-world data that is closely related to the type of situations for which we want to develop XAI methods: data-driven methods with legal, ethical, or social impact to end users.

For this evaluation we proceed just as we did for the preliminary test in Section 7.1.3 on the Churn, Mushroom, and Admission datasets—we fit a logistic regression model to the data to determine the dimension orders and subsequently evaluate various statistics to measure the degree to which the a fortiori model fits the data. However, unlike for the aforementioned datasets, we need to do more extensive preprocessing in order to get the data in an appropriate format. This results in a dataset that we will refer to as the "COMPAS dataset." In order to get a better understanding of our experimental results, we also make two variations on this dataset. The first, which we call the "simplified COMPAS dataset," contains only a subset of the features of the COMPAS set. Then, we relabel the simplified version according to a rule found by Angelino et al. (2018) using their Certifiably Optimal Rule Lists (CORELS) algorithm. We name this last dataset the "CORELS dataset."

The preprocessing steps we took are described in Section 7.2.1. We then describe in Section 7.2.2 our results for the COMPAS dataset, in Section 7.2.3 our results for the simplified COMPAS dataset, and in Section 7.2.4 our results for the CORELS dataset.

7.2.1 Data preprocessing

Before analyzing the COMPAS data we preprocess it. In particular, we discard features that are not of interest, delete rows that do not have values for the remaining features, create new features on the basis of old ones, and finally delete duplicate rows. Below follows a more detailed description of the steps taken.

First, we discard features that are not of interest. For instance, many of the features in

Table 7.2: An overview of the COMPAS features of interest. Angwin et al. (2016) did not give a comprehensive overview of the meaning of all the features used in their analysis, so we should note that this is only our best attempt at an interpretation.

Feature	Description	Order
`Age`	Age of the convict at the time of the COMPAS assessment.	Descending
`Sex`	Gender as specified when the convict was arrested, can take on the values 'Male' or 'Female'.	Female $\prec$ Male
`ChargeDegree`	Indicates whether the charge that led to the assessment was a felony (F) or a misdemeanor (M).	M $\prec$ F
`DaysInJail`	Number of days the convict spends in jail for the crime, computed by comparing (and rounding down) the number of days between the `c_jail_in` and `c_jail_out` fields.	Ascending
`DaysInCustody`	Number of days the convict spends in custody, computed in the same way as `DaysInJail` but with the `c_custody_in` and `c_custody_out` fields.	Ascending
`Priors`	Number of offenses committed prior to the one that led to the COMPAS assessment. The value of this field is computed as the sum of the values of `juv_fel_count`, `juv_misd_count`, `juv_other_count`, and `priors_count` fields in the original dataset.	Ascending
`Label`	The label, indicating whether there was "a criminal offense that resulted in a jail booking and took place after the crime for which the person was COMPAS scored [...] within two years after the first." (Larson et al., 2016)	N/A

the original dataset pertain to the COMPAS system, but presently we are only interested in the data describing the convicts and whether they recidivated or not, not in the COMPAS system itself. For example, one of the features describes the recidivism risk score (on a 1–10 scale) which COMPAS assigned to the individual.

Some features are of interest to us but are not in the right format. For instance, the two columns `c_jail_in` and `c_jail_out` together tell us how many days the convict spend in prison, but are represented in a date format, so we replace them with a new `DaysInJail` feature holding the number of days spent in prison. A complete overview of the resulting features and their meaning can be found in Table 7.2.

Lastly, we remove any rows that do not have values for any of the relevant features, or which occur more than once in the data. This last step is necessary for our landmark analysis; a case c may be a landmark, but if there is a second case d with exactly the same fact situation and outcome as c but not *equal* to c, then neither c nor d are landmarks.

We are then left with a total of 5,873 rows and we will henceforth refer to that set when we say 'COMPAS dataset'. In addition, we will look at two variations on that set. The first we will call the 'simplified COMPAS dataset', which is obtained from the COMPAS dataset by omitting all features except `Age` and `Priors`, and then deleting all duplicates. The second we call the 'CORELS dataset', and is obtained by changing the labels in the simplified COMPAS dataset according to the recidivism prediction rule found by Angelino et al. (2018, Figure 1) using their CORELS algorithm, see Figure 7.2.

if (*age* = 18 − 20) **then predict** *yes*
else if (*age* = 21 − 23) **and** (*priors* = 2 − 3) **then predict** *yes*
else if (*priors* > 3) **then predict** *yes*
else predict *no*

Figure 7.2: A rule list for the COMPAS dataset found by Angelino et al. (2018) using their CORELS algorithm. The clause related to sex has been excluded since this feature is omitted from the simplified COMPAS dataset for the sake of visualizability.

Table 7.3: On the left is a summary of the strength order on the COMPAS dataset, and the impact of the landmarks l_0 and l_1 defined in Definition 7.2. On the right is a concrete description of l_0 and l_1. Notice that they are archetypal examples of the *opposite* class that they belong to; l_0 is a young male with many priors, who did not recidivate; while l_1 is an older female with no priors, who did recidivate.

Property	**Label 0**	**Label 1**	**Total**
Consistent	76	397	473
Inconsistent	2,783	2,617	5,400
Forced by l_0	2,271	1,765	4,036
Forced by l_1	2,296	2,700	4,969
Landmark	70	18	88

d	$l_0(d)$	$l_1(d)$
`Age`	23	49
`Sex`	Male	Female
`ChargeDegree`	F	M
`DaysInJail`	70	0
`DaysInCustody`	70	0
`Priors`	11	0

7.2.2 Results on the COMPAS dataset

Having selected the dimensions, assigned their orders, and constructed the case base, we can now evaluate various statistics. We start by looking at the consistency percentage, i.e. the relative frequency of cases that do not have their outcome disputed by the strength order on the case base. We find the COMPAS dataset is only 8% consistent, see Table 7.1. This low percentage is caused by a small number of landmarks—outliers in the data that one would expect to have the opposite label of the one they received. We identify two landmarks l_0 and l_1 as being most impactful, which are defined as follows.

Definition 7.2. Given a finite case base $\mathscr{C}$ and an outcome s we define the set L_s of cases with outcome s that force the outcome of the greatest number of other cases in $\mathscr{C}$:

$$L_s := \operatorname{argmax}_{F \in \mathscr{C}_s} |\uparrow_s F \cap (\mathscr{C}_0 \cup \mathscr{C}_1)|.$$

When L_s is a singleton we write l_s for its sole element.

By transitivity of the strength order the cases in L_s are also landmarks, i.e. we have $L_s \subseteq \mathscr{L}_s$. In the datasets we consider in this work the L_s sets are singletons, so we will just refer to their sole elements l_0 and l_1. The l_s cases in the COMPAS dataset are shown in Table 7.3. In Figure 7.3 an overview of the collective impact of the landmarks is shown.

Remark 7.3. The notions of landmark and outlier, while similar, are not quite the same: a landmark need not be an outlier (cf. Figure 7.5) and an outlier need not be a landmark (for instance, when there is an outlier even further across the best-fit decision boundary).

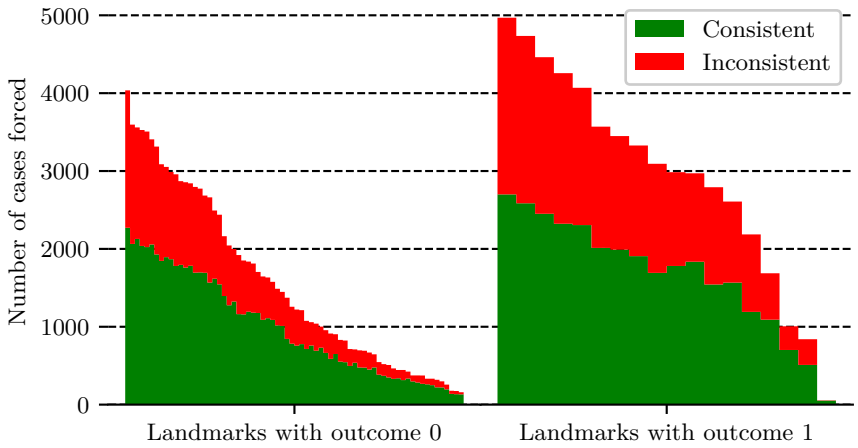


Figure 7.3: A visualization of the impact of the landmarks in the COMPAS data. Each vertical bar represents one landmark and shows the number of cases for which it forces the decision. The green area indicates the cases with an outcome equal to that of the landmark, and the red area the cases with an outcome different from the landmark (and which are therefore made inconsistent by the landmark). More precisely, for each landmark $(F, s) \in \mathscr{L}$ the green area represents $|\mathscr{C}_s \cap \uparrow_s F|$ and the red area represents $|\mathscr{C}_{\bar{s}} \cap \uparrow_s F|$.

7.2.3 Results on the simplified COMPAS dataset

High dimensional data is difficult to visualize, so in order to get a better view of these results we repeat our analysis on a subset of the data with only the two most predictive variables—**Age** and **Priors**. We call this the simplified COMPAS dataset. The resulting order on the variables remains the same as in the larger version. This lets us visualize the data, the decision surface of our logistic model, and the landmarks; see Figure 7.4 for the resulting plot. The landmarks highlight the cause for the inconsistency: there are many cases that lie on the opposite side of the decision boundary for their class, causing large overlap in their forcing cones.

7.2.4 Results on the CORELS dataset

The preceding results have shown that the model of precedential constraint is a poor fit on the COMPAS data. This makes sense intuitively, because when someone of a certain age and with some number of priors recidivates, we cannot expect this to set a precedent that future convicts will abide by. For example, when an elderly lady with no prior offenses recidivates, this will have very little influence on the behavior of convicts thereafter. In other words, the process underlying recidivism does not respect precedence.

This type of reasoning should be more suited to our running example from Part I in which we judge *risk* of recidivism (see e.g. Example 3.4). When a person is assigned low or high risk of recidivism, we would expect this assignment to obey the a fortiori principle.

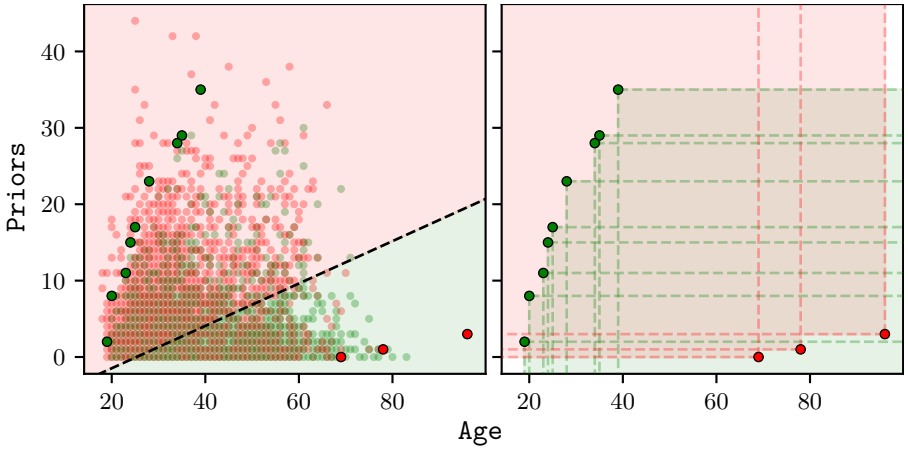


Figure 7.4: Two illustrations of the simplified COMPAS dataset. The green dots correspond to cases with outcome 0, and the red dots to those with outcome 1. The enlarged circles indicate the landmarks. On the left, all cases in the case base are shown, together with a dotted line indicating the decision boundary associated with the logistic regression coefficients. On the right, only the landmarks are shown, together with their forcing cones.

To test this hypothesis we change the labels of the simplified COMPAS data according to a sensible risk assessment rule, mined from the original COMPAS data by Angelino et al. (2018, Figure 1), as a demonstration of their CORELS algorithm. This rule is listed in Figure 7.2, with the only modification being that we omit the clause related to sex from the first case distinction since we have omitted this feature for the sake of visualizability. Changing all labels according to this rule, and then removing duplicates, results in a new dataset that we refer to as the CORELS dataset.

Now, we again fit our model to this data and visualize the decision boundary of the logistic regression model, along with the forcing cones of the landmarks; see Figure 7.5 for the resulting plot. As expected, the decision rule of Figure 7.2 does satisfy the a fortiori principle, and as a result the consistency is very high (in fact the dataset is fully consistent). The forcing cones of the landmarks are in agreement with the decision boundary determined by the logistic regression analysis.

In all, our results on the COMPAS datasets suggest that we can think of the phenomenon of inconsistency in two ways. The first is the mathematical view that the theory of precedential constraint contains a linearity assumption and that the consistency percentage is a measure of the degree to which the data is linearly separable. Of each class, the landmarks are then those cases that lie furthest in the direction of the best fit linear decision boundary, and the farther they cross it the more inconsistency they cause. The second is the semantic view that tells us to what degree the labelling process relies on a fortiori reasoning, or the degree to which we can expect precedent to be obeyed. If this is the case, then the landmarks are those cases that most reveal the nature of the underlying labelling process.

Our results also suggest that the presence of a small number of landmarks that force the decision of the rest is what we can expect of an average dataset, because in general a partial order will have far fewer minimal elements than that it will have elements in total. Two

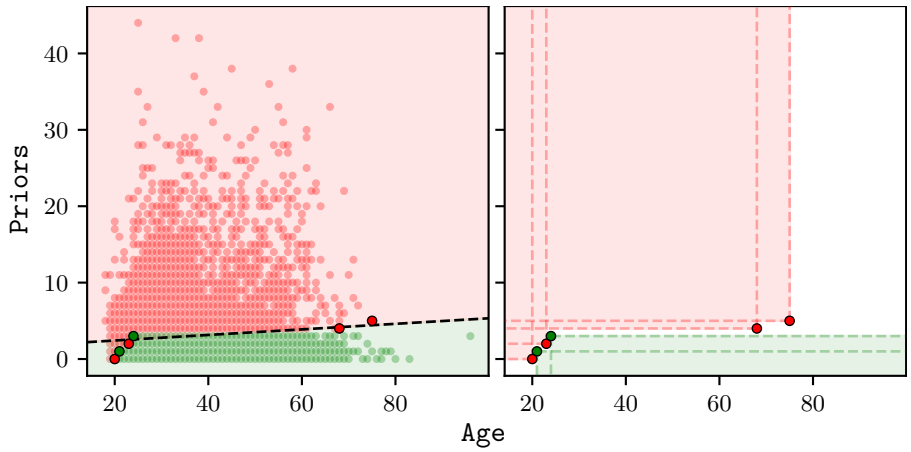


Figure 7.5: Two illustrations of the simplified CORELS dataset. The green dots correspond to cases with outcome 0, and the red dots to those with outcome 1. The enlarged circles indicate the landmarks. On the left, all cases in the case base are shown, together with a dotted line indicating the decision boundary associated with the logistic regression coefficients. On the right, only the landmarks are shown, together with their forcing cones.

factors that can influence this is the number of dimensions and the way in which we order them. For instance, if we have a dimension with more than two values and we order them so that they are all incomparable, it will immediately become impossible for any case to force the outcome of another, and so every case becomes a landmark.

7.3 A logical analysis of the CORELS dataset

An interesting fact of the CORELS dataset is that its labels are determined by a logical rule, which can be expressed in the same many-sorted language we used in Section 3.10.2 to formulate the a fortiori model. More specifically, the rule in Figure 7.2 corresponds to a formula $\Psi \in L^{\Sigma(D)}$ defined by:

$$\begin{aligned} \Psi &:= C_1 \vee C_2 \vee C_3, \\ C_1 &:= 18 \leq x_{\textbf{Age}} \leq 20, \\ C_2 &:= (21 \leq x_{\textbf{Age}} \leq 23) \wedge (2 \leq x_{\textbf{Priors}} \leq 3), \\ C_3 &:= 3 < x_{\textbf{Priors}}. \end{aligned} \tag{7.3.1}$$

A fact situation F is assigned label 1 if $\mathscr{D}, F \vDash \Psi$, and 0 otherwise—i.e. when $\mathscr{D}, F \vDash \neg\Psi$. Letting $\mathscr{X}_1 = [\![\Psi]\!]$ and $\mathscr{X}_0 = [\![\neg\Psi]\!] = \mathscr{X} \setminus [\![\Psi]\!]$. This means we have a situation as described in Section 3.10.3, in which the set of fact situations $\mathscr{X}$ is equal to a disjoint union $\mathscr{X}_0 \cup \mathscr{X}_1 = \mathscr{X}$ indicating binary ground truth labels. Moreover, since these $\mathscr{X}_0$ and $\mathscr{X}_1$ sets are defined in the logical language of the a fortiori model, we can use Z3 to reason about the relation between the ground truth labels and the forcing relation induced by the CORELS case base.

Let us illustrate how this works by looking at the Φ_s formulas. The CORELS dataset contains very few landmarks—only 6 in total—which allows us to write them down; see

Table 7.4: On the left are the landmarks in the consistent and incomplete CORELS dataset, and on the right are the landmarks of the modified, consistent and complete CORELS dataset.

Age	Priors	Label
21	1	0
24	3	0
20	0	1
23	2	1
68	4	1
75	5	1

$\text{Age}_{\leq 100}$	Priors	Label
21	1	0
24	3	0
20	0	1
23	2	1
100+	4	1

Table 7.4 for an overview. This also means we can write out the corresponding Φ_s formulas:

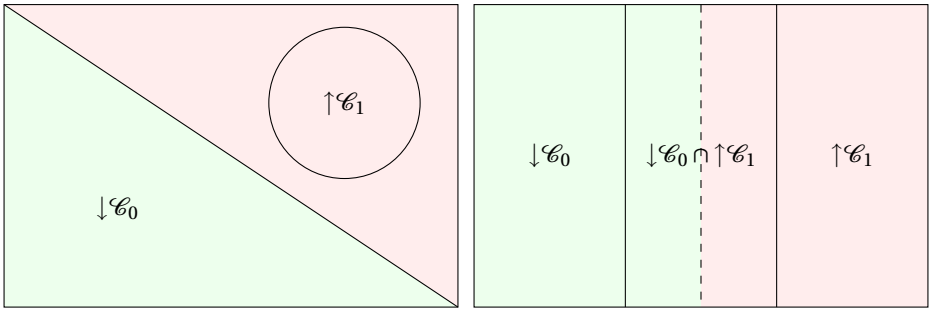
$$\Phi_0 = (24 \leq x_{\textbf{Age}} \wedge x_{\textbf{Priors}} \leq 3) \vee (21 \leq x_{\textbf{Age}} \wedge x_{\textbf{Priors}} \leq 1), \tag{7.3.2}$$

$$\Phi_1 = (75 \geq x_{\textbf{Age}} \wedge x_{\textbf{Priors}} \geq 5) \vee (68 \geq x_{\textbf{Age}} \wedge x_{\textbf{Priors}} \geq 4) \vee (20 \geq x_{\textbf{Age}} \wedge x_{\textbf{Priors}} \geq 0) \vee (23 \geq x_{\textbf{Age}} \wedge x_{\textbf{Priors}} \geq 2). \tag{7.3.3}$$

Using these we can precisely analyze the degree to which the forcing relation on cases is in accordance to the ground truth labels. For instance, is it always the case that when the case base forces a fact situation F for outcome 0, that F has ground truth label 0? In other words, does the inclusion $\downarrow\mathscr{C}_0 \subseteq \mathscr{X}_0$ hold? And what about the converse, $\downarrow\mathscr{C}_0 \supseteq \mathscr{X}_0$? Recall from Section 3.10.2 that these questions have logical counterparts. To check $\downarrow\mathscr{C}_0 = \mathscr{X}_0$ is the same as to check that $\Phi_0 \leftrightarrow \neg\Psi$ is valid, and the CORELS dataset is simple enough that this can be done by hand, using the basic rules for manipulating logical formulas:

$$\begin{aligned}
\neg\Psi &\leftrightarrow \neg\bigvee\{18 \leq x_{\textbf{Age}} \leq 20,\\
&\qquad (21 \leq x_{\textbf{Age}} \leq 23) \wedge (2 \leq x_{\textbf{Priors}} \leq 3),\\
&\qquad 3 < x_{\textbf{Priors}}\}\\
&\leftrightarrow \bigwedge\{21 \leq x_{\textbf{Age}},\\
&\qquad (x_{\textbf{Age}} \leq 20 \vee 24 \leq x_{\textbf{Age}}) \vee (x_{\textbf{Priors}} \leq 1 \vee 4 \leq x_{\textbf{Priors}}),\\
&\qquad x_{\textbf{Priors}} \leq 3\}\\
&\leftrightarrow \bigwedge\{21 \leq x_{\textbf{Age}},\\
&\qquad 24 \leq x_{\textbf{Age}} \vee x_{\textbf{Priors}} \leq 1,\\
&\qquad x_{\textbf{Priors}} \leq 3\}\\
&\leftrightarrow (24 \leq x_{\textbf{Age}} \wedge x_{\textbf{Priors}} \leq 3) \vee (21 \leq x_{\textbf{Age}} \wedge x_{\textbf{Priors}} \leq 1)\\
&\leftrightarrow \Phi_0.
\end{aligned}$$

In other words, we apply De Morgan's law to $\neg\Psi$, simplify the resulting expressions, distribute the conjunction over the disjunction, and then finally simplify the expression again to obtain Φ_0. Thankfully, we do not have to do this by hand, as Z3 can quickly perform such verifications. Any subsequent claims that we make about the validity of formulas was checked using Z3.



(a) CORELS **(b)** Tort law

Figure 7.6: Euler diagram representations of the relation between the CORELS case base and the ground truth labels determined by the decision rule of Figure 7.2 (in 7.6a), and of the relation between the tort law case base and the labels determined by Eq. (7.4.1) (in 7.6b).

Similarly to the derivation above, we can show that $\Phi_1 \rightarrow \Psi$ is valid, which tells us that $\uparrow\mathscr{C}_1 \subseteq \mathscr{X}_1$. However, as we saw in Proposition 3.15, this inclusion is necessarily strict. An Euler diagram representation for the CORELS case base can be found in Figure 7.6a.

The proof of Proposition 3.15 shows that the problem with making the CORELS case base complete is that the values for **Priors** and **Age** can become infinitely large. Since case bases are finite by definition, we can always find fact situations on the northeast part of the (**Priors**, **Age**) plane that do not have their outcome forced. If we put a cap on either of these values it would be possible to make the case base complete. Let $\mathbf{Age}_{\leq 100}$ denote the dimension equal to the **Age** dimension with the exception that it has a highest value '100+', i.e. fact situations have their age represented along this dimension, and any value that would normally be above 100 gets assigned the 100+ value. More specifically, let $\mathbf{Age}_{\leq 100}$ be a dimension consisting of the set $\{18, 19, 20, \ldots, 99, 100+\}$ ordered by $\geq$. Now, the CORELS case base can be made into a complete (and consistent) case base for the $\mathbf{Age}_{\leq 100}$ and **Priors** dimensions by the addition of a case with $\mathbf{Age}_{\leq 100}$ value 100+ and **Priors** value 4. See Table 7.4 for the landmarks of the resulting case base.

7.4 The tort and welfare datasets

In Section 7.3 we saw an example of a dataset that has its labels determined on the basis of a logical formula Ψ. The set of fact situations $\mathscr{X}$ was partitioned in two parts $\mathscr{X}_1 = [\![\Psi]\!]$ and $\mathscr{X}_0 = [\![\neg\Psi]\!] = \mathscr{X} \setminus [\![\Psi]\!]$, indicating the ground truth labels of the fact situations. This allowed us to precisely measure the fit of the a fortiori model by looking at the relationships between the sets $\uparrow_s \mathscr{C}_s$ and $\mathscr{X}_s$.

We now turn our attention to two more datasets that come with such a labelling formula Ψ. Firstly, we look at data on a real legal setting, namely the domain of Dutch tort law (Verheij, 2017). Secondly, we consider the fictional welfare benefit domain introduced by Bench-Capon (1993), and several variations on this dataset. Both of these datasets were recently used by Steging et al. (2021) to see if modern machine learning systems can learn the rules used to label the examples in these datasets. In this section we will essentially do

Table 7.5: A description of the features appearing in the welfare set together with a description of their meaning (Bench-Capon, 1993).

Feature	Values	Description
Age	0 – 100	The person's age; should be of pensionable age to be eligible (60 for a woman, 65 for a man).
Sex	Male or female	The person's sex, used to determine pension age.
$\mathbf{Con}_1, \ldots, \mathbf{Con}_5$	0 or 1	The person should have paid contributions in four out of the last five relevant contribution years.
Spouse	True or false	The person should be a spouse of the patient.
Absent	True or false	The person should not be absent from the UK.
Resources	0 – 10,000	The person should have capital resources not amounting to more than 3,000£.
Type	In or out	If the relative is an in-patient the hospital should be within a certain distance: if an out-patient, beyond that distance.
Distance	0 – 100	Distance to the hospital.

the same but for the a fortiori model, through an analysis similar to the one we performed for the CORELS dataset in Section 7.3. We perform this analysis first for the tort dataset, and then for the welfare dataset, in Section 7.4.

Tort law describes when a wrongful act is committed, and when the resulting damages must be repaired. The label of the training examples in the dataset we consider states whether such a *duty to repair* holds according to the law in that particular fact situation. Fact situations are described along 12 binary features. Examples of these features are `vun`, which states that the act was a violation of unwritten law against proper social conduct, or `imp`, which states the act can be imputed to the person that committed the act. For a complete overview of the features and their meaning the reader is referred to the work by Verheij (2017, Table 1) and Steging et al. (2023).

This duty to repair can be formalized according to the following rule:

$$
\begin{aligned}
&\Psi := \textstyle\bigwedge_{1 \leq i \leq 5} C_i, \\
&C_1 := x_{\mathbf{cau}}, \\
&C_2 := x_{\mathbf{ico}} \vee x_{\mathbf{ila}} \vee x_{\mathbf{ift}}, \\
&C_3 := x_{\mathbf{vun}} \vee (x_{\mathbf{vst}} \wedge \neg x_{\mathbf{jus}}) \vee (x_{\mathbf{vrt}} \wedge \neg x_{\mathbf{jus}}), \\
&C_4 := x_{\mathbf{dmg}}, \\
&C_5 := \neg(x_{\mathbf{vst}} \wedge \neg x_{\mathbf{prp}}).
\end{aligned}
\tag{7.4.1}
$$

The consistency percentage and number of landmarks for this dataset can be found in Table 7.1. Since there are only 10 binary features there are only $2^{10} = 1{,}024$ possible fact situations for this domain. The dataset we use contains all 1,024 of them, and so this case base is necessarily complete. Using Z3, we can furthermore prove that $\neg\Psi \rightarrow \Phi_0$ and $\Psi \rightarrow \Phi_1$ are valid, which means that $\mathscr{X}_0 \subseteq {\downarrow}\mathscr{C}_0$ and $\mathscr{X}_1 \subseteq {\uparrow}\mathscr{C}_1$. The corresponding Euler diagram representation can be found in Figure 7.6b.

Next, we turn to the welfare datasets, first used by Bench-Capon (1993) to investigate whether neural networks can handle open texture in law. They contain data about a fictional

welfare benefit paid to pensioners to defray expenses for visiting a spouse in a hospital. An overview of the features appearing in this dataset can be found in Table 7.5. The labels are determined as a logical function Ψ of these features, defined by:

$$\begin{aligned}
\Psi &:= \textstyle\bigwedge_{1\le i\le 6} C_i, \\
C_1 &:= (x_{\mathbf{Sex}} = \mathrm{F} \wedge x_{\mathbf{Age}} \ge 60) \vee (x_{\mathbf{Sex}} = \mathrm{M} \wedge x_{\mathbf{Age}} \ge 65), \\
C_2 &:= 4 \le \textstyle\sum_{1\le i\le 5} x_{\mathbf{Con}_i}, \\
C_3 &:= x_{\mathbf{Spouse}}, \\
C_4 &:= \neg x_{\mathbf{Absent}}, \\
C_5 &:= x_{\mathbf{Resources}} \le 3{,}000, \\
C_6 &:= (x_{\mathbf{Type}} = \mathrm{in} \wedge x_{\mathbf{Distance}} < 50) \vee (x_{\mathbf{Type}} = \mathrm{out} \wedge x_{\mathbf{Distance}} \ge 50).
\end{aligned} \tag{7.4.2}$$

Steging et al. (2021) used several different versions of the original welfare dataset for their experiments. Amongst these are two datasets each containing 50,000 examples, randomly sampled in the ranges described in Table 7.5, and each labelled according to the formula Ψ in Eq. (7.4.2). These were designed to either fail on a random number of the conditions $C_1, \ldots, C_6$, or to fail on just one specific condition. For our purposes, this distinction is not important, so we merge the datasets into one set that we will henceforth refer to as the welfare dataset. After merging and removing duplicates it contains 99,988 cases; its number of landmarks and consistency percentage can be found in Table 7.1.

Interesting to note is that the Pearson correlation method yields a substantially higher consistency percentage on this set: 71.1% as opposed to 48.5%. An inspection of the dimension orders shows that this is arguably the result of chance. The Pearson correlation and logistic regression methods agree on the signs of the coefficients of all dimensions except that of the **Distance** dimension. The Pearson correlation coefficient of this dimension is 0.001, while its coefficient from the logistic regression analysis is -0.01. We see that both methods assign a negligibly small value, which is because the **Distance** dimension violates the assumption that its values tend to favor either of the outcomes; if $x_{\mathbf{Type}} = \mathrm{in}$ then lower values of the **Distance** dimension are better for outcome 1, and if $x_{\mathbf{Type}} = \mathrm{out}$ then higher values are better for outcome 1. The Pearson correlation method happened to assign a small positive value to the coefficient, but it could have just as well produced a small negative coefficient for a slightly different sample; and the same holds for the logistic regression method.

What about the relation between the Φ_s formulas and Ψ? In fact, none of the possible inclusions hold, so its Euler diagram is the most general one, depicted in Figure 3.3a.

Remark 7.4. Note that the formulas involved in these situations can become very big: the Welfare dataset contains 12 features and 99,988 cases, so the forcing formula Φ_s will contain approximately $12 \cdot 99{,}988 \approx 1.2$ million atomic subformulas. Nevertheless, Z3 is capable of handling big formulas such as these.

Part of the analysis performed by Steging et al. (2021) used a simplified version of the welfare set containing only a subset of the features of the original set: namely **Sex**, **Age**, **Type**, and **Distance**. The labels of this set are determined only by conditions C_1 and C_6,

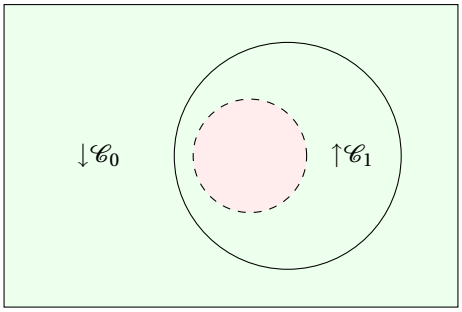


Figure 7.7: An Euler diagram representation of the relation between the simplified welfare case base and the ground truth labels determined by the formula in Eq. (7.4.3).

i.e. its labelling formula Ψ is defined as:

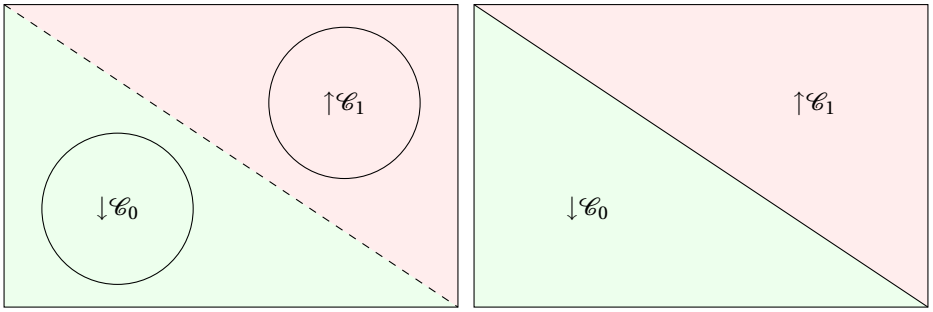
$$\begin{aligned} \Psi &:= \textstyle\bigwedge_{1\leq i\leq 6} C_i, \\ C_1 &:= (x_{\textbf{Sex}} = \mathrm{F} \wedge x_{\textbf{Age}} \geq 60) \vee (x_{\textbf{Sex}} = \mathrm{M} \wedge x_{\textbf{Age}} \geq 65), \\ C_6 &:= (x_{\textbf{Type}} = \mathrm{in} \wedge x_{\textbf{Distance}} < 50) \vee (x_{\textbf{Type}} = \mathrm{out} \wedge x_{\textbf{Distance}} \geq 50). \end{aligned} \tag{7.4.3}$$

We refer to this set as the simplified welfare dataset and performed a similar analysis on it as with the other sets. The results can be found in Table 7.1.

The consistency percentage on this dataset is not great—only about 67.3% for the Pearson correlation method, and only 66% for the logistic regression method. However, when we break down this percentage for both classes we see that the situation is more dire than it at first appears. The consistency percentage for class 0 is 86.4%, but that of class 1 is 0%. This is caused by a single landmark with label 0, which forces all cases with outcome 0 for outcome 1, which means that $\uparrow\mathscr{C}_1 \subseteq \downarrow\mathscr{C}_0$: any case forced for outcome 1 by the case base is also forced for outcome 0. Z3 can prove that the case base is complete, and so since $\uparrow\mathscr{C}_1 \subseteq \downarrow\mathscr{C}_0$ this means $\mathscr{X} \subseteq \downarrow\mathscr{C}_0$: all fact situations are forced for outcome 0 by the case base. Lastly, it can be shown that $\mathscr{X}_1 \subseteq \uparrow\mathscr{C}_1$. The Euler diagram corresponding to this situation is shown in Figure 7.7.

Why is the a fortiori model such a poor fit for this dataset? The reason, as mentioned previously, is that the features violate the assumption that their possible values have a preference for either of the binary outcomes. The way that **Distance** values prefer outcome 0 or 1 depends on the value of the **Type** dimension. Similarly, the **Type** and **Sex** dimensions do not themselves prefer outcome 0 or 1; they are just information to be conditioned on in the labelling formula. The only exception is the **Age** dimension, for which higher values clearly prefer outcome 1.

The simplified welfare set isolates almost exactly the variables that violate the dimension order assumption, which is also indicated by the fact that both the Pearson correlation and logistic regression methods assign coefficients to these dimensions which are very close to 0. What if we do the opposite: isolate from the original welfare dataset exactly the variables that satisfy the dimension order assumption? This means removing the **Distance**,



(a) Before completion **(b)** After completion

Figure 7.8: Euler diagram representations of the relation between the second simplified welfare case base and the ground truth labels determined by the formula in Eq. (7.4.4), before (7.8a) and after (7.8b) running Algorithm 1.

Type and **Sex** dimensions, and relabelling the data according to the following formula:

$$\begin{aligned}
&\Psi := \textstyle\bigwedge_{1\leq i\leq 5} C_i, \\
&C_1 := x_{\textbf{Age}} \geq 60, \\
&C_2 := 4 \leq \textstyle\sum_{1\leq i\leq 5} x_{\textbf{Con}_i}, \\
&C_3 := x_{\textbf{Spouse}}, \\
&C_4 := \neg x_{\textbf{Absent}}, \\
&C_5 := x_{\textbf{Resources}} \leq 3{,}000.
\end{aligned} \tag{7.4.4}$$

Performing this modification and subsequently removing duplicates yields a new dataset with 96,348 cases, which we will refer to as the second simplified welfare dataset. Fitting the a fortiori model on this set we get a consistency percentage of 100%. Moreover, Z3 can prove that $\mathscr{X}_0 \subseteq \downarrow\mathscr{C}_0$ and $\mathscr{X}_1 \subseteq \uparrow\mathscr{C}_1$. The Euler diagram corresponding to this situation is depicted in Figure 7.8a.

We see that the only property missing now is completeness. This means that it might be possible to add certain cases, so that the result is a consistent and complete case base. To finish this section on data analysis, we show that Z3 can potentially be used to complete a case base in such a scenario. This works because in order to prove completeness Z3 tries to find a counterexample, i.e. a fact situation $F \in \mathscr{X} \setminus (\downarrow\mathscr{C}_0 \cup \uparrow\mathscr{C}_1)$. If it succeeds at finding such a fact situation, we can determine its label using Eq. (7.4.4) and add it to the case base, after which we ask Z3 to prove completeness again. This yields an algorithm for completing a case base, described in Algorithm 1. This algorithm does not necessarily terminate; e.g. Proposition 3.15 tells us it would loop endlessly on the CORELS dataset. If it does terminate, this is either because the case base was made inconsistent, or it was made complete by the addition of the last added case, while retaining the consistency property.

Running Algorithm 1 on the second simplified welfare case base yields a consistent and complete case base with 19 landmarks; see Figure 7.8b for the corresponding Euler Diagram visualization.

Algorithm 1: Completing a case base using Z3.

Data: A consistent, incomplete case base $\mathscr{C}$
while $\mathscr{C}$ *is consistent and incomplete* **do**
  $F \leftarrow$ the counterexample to completeness generated by Z3;
  $s \leftarrow$ the ground truth label of F according to the labelling formula Ψ;
  $\mathscr{C} \leftarrow \mathscr{C} \cup \{(F, s)\}$;
end

7.5 Conclusion

In this chapter, we have explored the application of the dimensional result model (DRM) to various datasets to evaluate its fit in terms of case base consistency. We implemented the DRM using the Z3 SMT solver and tested it on several datasets, including the COMPAS recidivism dataset, the tort law dataset, and the welfare benefit datasets. Our findings indicate that the DRM can effectively model datasets where the features have a clear preference for one of the binary outcomes. However, in cases where this assumption is violated, the model's fit is poor. We also demonstrated the use of Z3 to analyze the logical structure of datasets and to complete case bases while maintaining consistency.

The a fortiori case-based reasoning method of explanation developed by Prakken and Ratsma (2022), which we discussed in Chapter 7, operates on the basis of the DRM. In Chapter 5 we presented an extension of this model—the dimensional hierarchical result model (DHRM)—which is capable of modeling a fortiori constraint of decisions that can take more than two values. If the explanation method of Prakken and Ratsma were modified to operate on the basis of the DHRM it would be more broadly applicable, because most AI systems output real numbers, and not binary values.

This raises the question how the analyses considered in this chapter would look for the DHRM, as opposed to the DRM, and this will be the topic of the next chapter.

Chapter 8

COMPAS Risk Scores Case Study

AST CHAPTER, we fit the dimensional result model (DRM, cf. Chapter 3) to machine learning datasets.[1] In particular, we computed the consistency of the COMPAS dataset with respect to the binary labels indicating whether a person did, or did not, recidivate within a two-year timeframe of being COMPAS scored. However, we did not measure the consistency of the COMPAS scores themselves, as these are not on binary scales. In this chapter, we extend our Z3-based implementation to compute constraint of the dimensional-hierarchical result model (DHRM, cf. Chapter 5), which does allow us to measure the consistency of COMPAS. The contributions of this chapter are twofold. We prove two formal results on case base consistency and relate those to COMPAS. In both cases, the consistency of COMPAS contradicts the theoretical predictions, and we reflect on the causes of these discrepancies.

The first theoretical result states that a large class of popular statistical models, called generalized linear models, always produces fully consistent case bases. It has been claimed by the developers of COMPAS that the program computes its scores based on a regression model, and so this should mean that the COMPAS risk scores display a high level of consistency. In contrast, we find that the consistency of the risk scores is actually very low. We conclude that this can be explained by the fact that there are features missing from the COMPAS dataset which are used by COMPAS to compute its scores.

The second theoretical result relates case base consistency to *binning*—the practice of subdividing a range of values into smaller, consecutive, non-overlapping intervals, which are called bins. We show that consistency scores should decrease when input features are binned. Again, an analysis of COMPAS outputs produces results that contradict this theoretical result. This brings to light an underdiscussed aspect of COMPAS: the use of norm groups for the interpretation of COMPAS outputs. We show that information on these norm groups can be reverse-engineered from the COMPAS data, by use of a graph coloring algorithm.

The contents of this chapter are structured as follows. We start by giving some background information in Section 8.1 on COMPAS, its risk scores, and on the dataset

[1] The material in this chapter stems from van Woerkom (2025) and van Woerkom et al. (2024b), with the exception of Section 8.2.3 which is new.

containing examples of these scores. In Section 8.2 we recall the DHRM model of constraint which we use for the analysis, together with a description of its Z3-based implementation. In Section 8.3 we give the first theoretical result on generalized linear models, and relate it to the COMPAS risk scores. Then, in Section 8.4, we give the second theoretical result, and relate it to the COMPAS recommended supervision level scores. Finally, we end in Section 8.5 with some concluding remarks.

8.1 The COMPAS risk assessment dataset

The primary outputs of the COMPAS program are its need and risk scale assessments, computed based on answers to questionnaires; see the Berkman Klein Center for Internet and Society (2025) for examples. The COMPAS need scales measure constructs like financial problems, substance abuse, and depression, while the risk scales predict factors like recidivism, violence, and failure to appear (Equivant, 2019). Additionally, COMPAS provides a "level of supervision" recommendation, ranging from 1 (lowest) to 4 (highest).

The inputs to these scales are answers to questionnaires about prior offenses, education, work experience, et cetera. Some data are self-reported. These answers inform the need scales, which in turn inform the risk scales. The COMPAS risk scales, particularly the "General Recidivism Risk Score" (GRRS) and the "Violent Recidivism Risk Score" (VRRS), have been debated. Angwin et al. (2016) published a dataset containing COMPAS risk scores assigned between 2013 and 2014 in Broward County, Florida, which we discussed in Section 7.2 of the previous chapter. This dataset includes scores, information on which these scores were presumably computed, and information about whether a person recidivated or committed a violent act after being scored (Larson et al., 2016).

The dataset has been criticized for missing features needed to compute the COMPAS risk scores (Michelle Bao et al., 2021). Rudin et al. (2020a) supplemented ProPublica's dataset with probation data from the Broward Clerk's office to fill in some missing features. However, some information is still missing (Rudin et al., 2020a, Table 1). The dataset published by Rudin et al. (2020a) contains data on 9 features used to compute the GRRS and 13 for the VRRS (Rudin et al., 2020a, Table 1 & Tables A4–A8; Engel et al., 2024, Tables 1–5).

According to Northpointe (2009) the failure to appear risk score "is based largely on prior history of a failure to appear, current charges for failure to appear, prior recidivism on community placement, general criminal involvement, and unstable residential ties and transience." This score has received less attention in the literature, presumably because the dataset published by Angwin et al. (2016) does not contain "true label" information on this scale. Again, we do not have access to most of the features on the basis of which the failure to appear risk scores are computed. As an approximation we will use the history of criminal involvement subscale and marital status information, totaling 8 input features.

Lastly, the COMPAS risk scores can be presented in various ways. Initially, COMPAS produces *raw* scores, which can take any value, such as -1.54, for example (Equivant, 2019). For interpretability, these raw scores are converted to *decile* scores by comparing them to those of a *norm group*—a representative sample of the target population of the agency using COMPAS. For instance, a raw score of -1.54 might be converted to a decile score of 6, indicating it is higher than the lowest 50% but lower than the highest 40% of

scores in the norm group. These decile scores may then be further represented as "Low" (scores 1–4), "Medium" (5–7), and "High" (8–10). (According to Nisbet et al., 2018, "a surprising number of agencies prefer the traditional labels of low, medium, and high risk.")

We will use the extended dataset published by Rudin et al. (2020a) to relate formal results about case-based reasoning consistency, to our analysis of the COMPAS risk scores. Before turning to these results, we first recall the model of case-based reasoning and its associated notion of consistency.

8.2 The model of a fortiori constraint

As the basis of our analysis we use the dimensional-hierarchical result model (DHRM) which we developed in Chapter 5. We will now recall the definitions, and illustrate their applicability to the COMPAS risk scores through some examples. However, we note in advance that we will make some minor modifications to simplify the rest of the analyses. The essence of the DHRM will remain unchanged, and we will note any deviations from the definitions of Chapter 5.

8.2.1 Knowledge representation

A dimension d will be a partially ordered set, but in this chapter, we will additionally require it to be *total*. This means that for any $v, w \in d$, either $v \preceq w$ or $w \preceq v$. In other words, we assume that a dimension is *linearly* ordered. This assumption is made because non-total dimension orders are uncommon, particularly in the context of AI. For instance, the statistical methods we use to determine dimension orders inherently produce total orders (see Section 7.1.1 of the previous chapter). Moreover, some of the results we present and prove rely on the totality property. While we believe that these results could be adapted to apply to non-total dimensions with appropriate modifications, this remains a subject for future research.

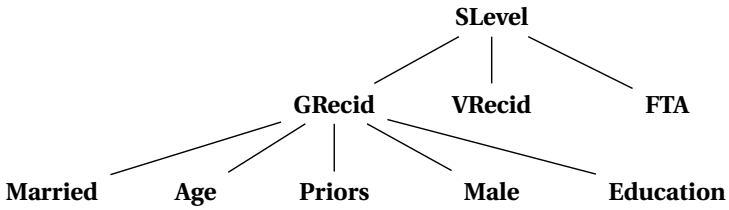
As before, we may refer to just the set d as the dimension, leaving the reference to its linear order implicit. Furthermore, in the context of a set of dimensions D, we will, for the sake of brevity, refer to all the orders of the dimensions by just $\preceq$, because confusion as to which dimension order is being referred to is unlikely to arise. For example, given two dimensions d and e, we will refer to both of the dimensions orders of d and e by $\preceq$, rather than introducing separate notations such as $\preceq_d$ and $\preceq_e$.

Definition 8.1. A *dimension hierarchy* (D, H) is a finite set of dimensions D with a relation H on D such that the transitive closure of H is irreflexive. A dimension is *base-level* if it is H-minimal, and *abstract* otherwise. We denote the pre-image of a dimension d under a hierarchical structure H by $\mathsf{H}(d) = \{e \in D \mid \mathsf{H}(e, d)\}$.

Remark 8.2. Note that Definition 8.1—unlike Definition 5.1 of a dimension hierarchy—does not assume that the links in the hierarchy have a polarity. Link polarity is not strictly necessary because the constraint induced by a hierarchy with link polarity can always be reproduced by a hierarchy with only positive links. This is done by adding a copy of each dimension, corresponding to its inverse order: for any $d \in D$ with order $\preceq$, add a dimension d' with order $\succeq$, and replace negative links from d by positive links from d', so that the resulting hierarchy contains only positive links.

A link $\mathsf{H}(d, e)$ between dimensions d, e in a dimension hierarchy (D, H) indicates that there is a positive correlation between the values of d and e, relative to their orders.

Example 8.3. Below is an example of a dimension hierarchy in the context of the COMPAS risk scores:



(8.2.1)

The hierarchical structure H is indicated by the lines, where the higher dimensions indicate an increasing level of abstraction. The associated sets, orders, and meanings of the base-level dimensions, displayed in the bottom row of (8.2.1), are as follows:

$$\begin{aligned}
\textbf{Married} &= (\{0,1\}, \geq), \text{ is married or not,}\\
\textbf{Age} &= (\{18,19,20,\ldots\}, \geq), \text{ the age of the defendant,}\\
\textbf{Priors} &= (\{0,1,2,\ldots\}, \leq), \text{ the number of prior offenses,}\\
\textbf{Male} &= (\{0,1\}, \leq), \text{ is male or not,}\\
\textbf{Education} &= (\{0,1\}, \geq), \text{ completed high school or not.}
\end{aligned}$$

The order $\geq$ of the **MS**, **Age**, and **Education** dimensions indicates that higher values for these dimensions should generally lead to lower values for the overlying **GRecid** dimensions. Conversely, the $\leq$ order of the other dimensions, such as the number of priors **Priors**, indicates that higher values generally lead to higher values for the **GRecid** dimension.

Above the base-level dimensions is a row of more abstract dimensions. Note that **VRecid** and **FTA** are technically base-level dimensions in this example—in a more realistic version of this hierarchy these would also be dependent on less abstract dimensions. We use the decile scores, ranging from 1 to 10, for this example:

$$\begin{aligned}
\textbf{GRecid} &= (\{1,2,\ldots,10\}, \leq), \text{ the general recidivism risk score,}\\
\textbf{VRecid} &= (\{1,2,\ldots,10\}, \leq), \text{ the violent recidivism risk score,}\\
\textbf{FTA} &= (\{1,2,\ldots,10\}, \leq), \text{ the failure to appear risk score.}
\end{aligned}$$

At the top of the hierarchy is the **SLevel** dimension:

$$\textbf{SLevel} = (\{1,2,3,4\}, \leq), \text{ the recommended supervision level.}$$

Definition 8.4. A *fact situation* X for a set of dimensions D is a partial choice function on D. We denote the domain of a fact situation X by $\mathrm{dom}(X)$. The set of all fact situations for D is denoted by $\mathscr{X}(D)$, and a *case base* is a finite subset $\mathscr{C} \subseteq \mathscr{X}(D)$.

8.2.2 Constraint

Definition 8.5. Given a case base $\mathscr{C}$ and a value $v \in d$, a fact situation X is *lower bounded* in d by $\mathscr{C}$ to v, denoted by $\mathscr{C} \models v \preceq X(d)$, if and only if either

Table 8.1: Three fact situations X, Y, Z for the example dimension hierarchy for the recidivism risk domain depicted in (8.2.1). A dash indicates that the fact situation is undefined on that particular dimension.

	MS	**Age**	**Priors**	**Male**	**Educ**	**GRecid**	**VRecid**	**FTA**	**SLevel**
X	1	25	3	0	1	7	4	7	2
Y	0	30	2	0	0	5	8	—	3
Z	1	20	3	1	—	—	9	5	—

(1) v is the least element of d, or

(2) $v \preceq X(d)$, or

(3) d is abstract, and there is $Y \in \mathscr{C}$ satisfying $v \preceq Y(d)$ such that $\mathscr{C} \models Y(e) \preceq X(e)$ holds for all $e \in \mathsf{H}(d) \cap \mathrm{dom}(Y)$.

The *upper bound* $\mathscr{C} \models X(d) \preceq v$ is defined similarly.

Remark 8.6. Note that the role of the supp and opp functions, which we used in Chapters 3 and 5 for the DRM and the DHRM, is now played by disjunct (1) of Definition 8.5. The difference is minimal: we now have $\mathscr{C} \models v \preceq X(d)$ when X is undefined on d and v is the least element of d, whereas before this was not necessarily the case.

Example 8.7. An example case base for the dimension hierarchy of Example 8.3 is listed in Table 8.1.

$$\{X, Y\} \models 5 \preceq Z(\textbf{GRecid}) \tag{8.2.2}$$
$$\text{if } \{X, Y\} \models X(e) \preceq Z(e) \text{ for all } e \in \mathsf{H}(\textbf{GRecid}) \cap \mathrm{dom}(X) \tag{8.2.3}$$
$$\text{if } 25 \preceq Z(\textbf{Age}) \text{ and } 3 \preceq Z(\textbf{Priors}) \tag{8.2.4}$$
$$\text{if } 25 \geq 20 \text{ and } 3 \leq 3 \tag{8.2.5}$$

Step (8.2.3) corresponds to disjunct (3) of Definition 8.5, and may be applied because **GRecid** is abstract and $5 \preceq X(\textbf{GRecid}) = 7$. Step (8.2.4) follows from disjuncts (8.2.3) and (8.2.4), as X selects the least elements of **MS**, **Male**, and **Education**. Step (8.2.5) simply fills in the definition of Z, and is a true statement, so that we indeed have a lower-bound constraint $\{X, Y\} \models 5 \preceq Z(\textbf{GRecid})$. Using this, we can in turn derive:

$$\{X, Y\} \models 3 \preceq Z(\textbf{SLevel})$$
$$\text{if } \{X, Y\} \models Y(e) \preceq Z(e) \text{ for all } e \in \mathsf{H}(\textbf{SLevel}) \cap \mathrm{dom}(Y)$$
$$\text{if } \{X, Y\} \models 5 \preceq Z(\textbf{GRecid}) \text{ and } \{X, Y\} \models 8 \preceq Z(\textbf{VRecid})$$

We have already verified that $\{X, Y\} \models 5 \preceq Z(\textbf{GRecid})$ holds, and $\{X, Y\} \models 8 \preceq Z(\textbf{VRecid})$ holds because $8 \preceq Z(\textbf{VRecid}) = 9$, and so we indeed have a lower-bound constraint $\{X, Y\} \models 3 \preceq Z(\textbf{SLevel})$.

Next, we recall Definition 5.7 of case base consistency.

Definition 8.8. Given a dimension d, a fact situation X, and a case base $\mathscr{C}$, we say X is *d-inconsistent with respect to $\mathscr{C}$* if there are values $v, w \in d$ with $v \prec w$, such that both $\mathscr{C} \models X(d) \preceq v$ and $\mathscr{C} \models w \preceq X(d)$; otherwise X is *d-consistent*. The *d-consistency percentage* of $\mathscr{C}$, denoted $\mathrm{Cons}_d(\mathscr{C})$, is the relative frequency of d-consistent cases in $\mathscr{C}$:

$$\mathrm{Cons}_d(\mathscr{C}) = \frac{|\{X \in \mathscr{C} \mid X \text{ is } d\text{-consistent}\}|}{|\mathscr{C}|} \tag{8.2.6}$$

Remark 8.9. Note that the usage of sets in this model is somewhat informal. For instance, in Example 8.3 we treat **GRecid** and **VRecid** as different dimensions, even though they should strictly speaking be considered identical as sets. We follow the tradition in the literature and refer to dimensions as sets in spite of these concerns (as e.g. Horty did in Horty, 2019), but strictly speaking it may be more accurate to speak of *multisets*, which are sets that can can contain multiple copies of an element. This will be particularly relevant for case bases, as the cardinality of a case base appears in the denominator of (8.2.6) in Definition 8.8, so we reiterate that case bases can contain multiple instances of a case.

Example 8.10. Reconsidering Table 8.1, it can be checked that X and Y are **GRecid**- and **SLevel**-consistent with respect to $\{X, Y\}$, so

$$\mathrm{Cons}_{\mathbf{GRecid}}(\{X, Y\}) = \mathrm{Cons}_{\mathbf{SLevel}}(\{X, Y\}) = 1.$$

Now, suppose we assigned a **GRecid** score of 3 to the fact situation Z, so $Z(\mathbf{GRecid}) = 3$. We saw in Example 8.7 that $\{X, Y\} \models 5 \preceq Z(\mathbf{GRecid})$, but now we also have $\{X, Y\} \models Z(\mathbf{GRecid}) \preceq 3$ by Definition 8.5 as $Z(\mathbf{GRecid}) = 3$. Therefore, since $3 \prec 5 \in \mathbf{GRecid}$, we see that Z would become **GRecid**-inconsistent with respect to the case base $\{X, Y\}$ as a result of the assignment $Z(\mathbf{GRecid}) = 3$.

The rest of this chapter revolves around Definition 8.8 of case base consistency. To start, we derive a formal result about what can be expected of the consistency of decisions made by a program such as COMPAS, and then compare it to its actual consistency in the data that is available. For this, we need to expand the Z3-based implementation which we developed in Section 7.1.2 so that it can compute constraint for the DHRM.

8.2.3 An implementation in Z3

In Chapter 3 we showed that Horty's dimensional result model (DRM) can be rephrased in terms of many-sorted logic, and can thus be implemented in the Z3 solver (de Moura & Bjørner, 2008). With some small changes the same can be done for the DHRM. Due to the similarity of these approaches we will only give a brief description.

The basic idea is to phrase the constraint induced by a case base in terms of formulas of many-sorted logic. More specifically, given fact situations $X, Y \in \mathscr{X}(D)$ for some set of dimensions D, the statements that "X is upper-bounded by Y on d" and "X is lower-bounded by Y on d" can, respectively, be expressed logically by two formulas:

$$\lambda(X, Y, d) = \left(\bigwedge\nolimits_{e \in \mathsf{H}(d)} Y(e) \preceq X(e)\right) \rightarrow Y(d) \preceq X(d), \tag{8.2.7}$$

$$\upsilon(X, Y, d) = \left(\bigwedge\nolimits_{e \in \mathsf{H}(d)} X(e) \preceq Y(e)\right) \rightarrow X(d) \preceq Y(d). \tag{8.2.8}$$

Similarly, given a fact situation $X \in \mathscr{X}(D)$ a case base $\mathscr{C} \subseteq \mathscr{X}(D)$, and a dimension $d \in D$, the λ and υ formulas can be used to express all the d-constraint induced by $\mathscr{C}$ on X:

$$\Phi(X, \mathscr{C}, d) = \bigwedge_{Y \in \mathscr{C}} \lambda(X, Y, d) \wedge \upsilon(X, Y, d). \tag{8.2.9}$$

Note that these formulas do not mirror the recursive aspect of Definition 8.5. This could easily be implemented by modifying $\Phi(X, \mathscr{C}, d)$ to also include the lower and upper bounds induced by cases in $\mathscr{C}$ for dimensions in $e \in \mathsf{H}^+(d)$, where H^+ denotes the transitive closure of H. We do not need this in our implementation because the hierarchies we use will be flat.

Given a flat hierarchy culminating in d, the d-consistency of the fact situation X with respect to $\mathscr{C}$ is equivalent to the satisfiability of $\Phi(X, \mathscr{C}, d)$ (which, since $\Phi(X, \mathscr{C}, d)$ does not contain free variables, is just its truth value). An SMT solver is a program designed to check satisfiability of formulas, so the Eqs. (8.2.7)–(8.2.9) are what allows us to use Z3 to compute with the model. The full implementation, and the rest of our code, is available online.[2]

8.3 Consistency of generalized linear models

It is unknown exactly how the COMPAS program works due to its proprietary nature, but its developers have previously indicated that the scores it produces are (at least partially) based on regression models (Brennan & Dieterich, 2018; Brennan et al., 2009; Equivant, 2019; Jackson & Mendoza, 2020). For example, Brennan et al. (2009) stated that the "Recidivism Risk Scale is a regression model that has been used in COMPAS since 2000," and that "the COMPAS risk and classification models use logistic regression [...] in [...] prediction and classification procedures." Furthermore, the COMPAS Practitioner's Guide (Equivant, 2019) states that "linear equations are used to calculate the [general recidivism and violent recidivism] risk scales," and that the violent recidivism risk score, the VRRS, is computed as the following weighted sum:[3]

$$\begin{aligned} \text{VRRS} = {} & (-w_1 \cdot \text{age}) + (-w_2 \cdot \text{age at first arrest}) \\ & + (w_3 \cdot \text{history of violence}) + (w_4 \cdot \text{vocation education}) \\ & + (w_5 \cdot \text{history of noncompliance}) \end{aligned}$$

Despite these claims by the developers of COMPAS, Rudin et al. (2020a) argue, on the basis of a data analysis of the COMPAS dataset, that the scores assigned by COMPAS depend *nonlinearly* on age. Equivant disputed these claims (Jackson & Mendoza, 2020), and reiterated that the COMPAS risk scales make use of logistic regression. Rudin et al. (2020b) responded that if COMPAS does operate on the basis of a logistic regression model, the age variable might first undergo a nonlinear transformation before being fed as input to the model.

As we can see, the question of whether the risk scores produced by COMPAS are the output of a relatively simple linear model has been the topic of debate. Indeed, Rudin (2019)

[2] https://git.science.uu.nl/ics/responsible-ai/van-woekom/afcbr.

[3] Though Equivant has later stated that this description should not be taken as a complete technical description of the violent recidivism risk scale (Jackson & Mendoza, 2020).

has argued that it is always better to use interpretable models, such as linear regression or logistic regression, for high stakes decision-making, rather than complex black-box machine learning algorithms. In the case of COMPAS, it might be that it is only a black box because of its proprietary nature, and not because it makes use of uninterpretable machine learning algorithms such as neural networks. Ideally, we would be able to verify whether COMPAS is a linear model without compromising Equivant's intellectual property protections.

In this section we will show that the model of a fortiori case-based reasoning, which we reviewed in Section 8.2, can theoretically be used to falsify the claim that a given set of outputs was produced by a linear model. More specifically, we show that for a large class of linear models called *generalized linear models* (GLMs), introduced by Nelder and Wedderburn (1972), a case base of model decisions is necessarily consistent in the sense of Definition 8.8.

We start by recalling the basic definition of a GLM, and give a concrete example in the form of a logistic regression model trained on the COMPAS dataset. This is a representative example, since it has been claimed that COMPAS is itself a form of logistic regression. We then prove a theorem stating that a case base of GLM decisions is necessarily fully consistent. Given this result, we would expect that the consistency of the COMPAS risk score assignments in the dataset made available by Angwin et al. (2016) and Rudin et al. (2020a) is high. We will show that quite the opposite is the case, and we discuss some possible causes of these low consistency percentages.

8.3.1 Regression analysis

Regression analysis is a statistical modeling technique used to predict the expected value of a random variable $\boldsymbol{y}$ as a function of a set of observed values $\boldsymbol{x} = (x_1, \ldots, x_n)$. The simplest form of regression is *linear* regression, in which the expected value $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{x})$ of $\boldsymbol{y}$ given $\boldsymbol{x}$ corresponds to the linear combination of $\boldsymbol{x}$ with a vector of coefficients $\beta_0, \beta_1, \ldots, \beta_n$, so $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{x}) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$. This linear combination of the β_i and x_i is called the *linear predictor*. There are many variations on this idea, such as logistic regression, Poisson regression, gamma regression, and so forth.

Nelder and Wedderburn (1972) showed that many forms of regression fit in a common class which they called *generalized linear models* (GLMs). A GLM assumes that the random variable $\boldsymbol{y}$ is distributed according to a member of the exponential family of probability distributions, and that the expected value of $\boldsymbol{y}$, conditioned on an observed set of values $\boldsymbol{x}$, is related to the linear predictor by a monotone *link function* g, i.e. that $g(\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{x})) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$ (Dobson, 2001). Note that linear regression is obtained as a GLM by using an identity link function $g(x) = x$. The inverse of the link function is often assumed to exist and is called the *mean function*, denoted by m, as it maps the linear predictor to the mean of the random variable.

The β_i coefficients of the GLM are often estimated from data using techniques such as maximum likelihood estimation (Dobson, 2001, Chapter 4). Once this is done, the GLM can be used for prediction. More specifically, given a new observation $(x_1, \ldots, x_n)$, the GLM can estimate a value $\hat{\boldsymbol{y}}$ of the target random variable $\boldsymbol{y}$ by simply applying the mean function to the linear predictor: $\hat{\boldsymbol{y}} = m(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)$. Note that for the purpose of prediction the choice of distribution for $\boldsymbol{y}$ is no longer relevant. Since we are primarily interested in the

application of GLMs to predictive modeling, we will not further consider this aspect of GLMs in this work, and instead consider an n-ary GLM $(m, \boldsymbol{\beta})$ to be parameterized by two components: a vector of $n+1$ real coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_n)$, and an n-ary monotone mean function m.

Example 8.11. Logistic regression is obtained as a GLM by choosing the sigmoid σ as the mean function:

$$m(x) = \sigma(x) = (1 + \exp(-x))^{-1} \tag{8.3.1}$$

We will now demonstrate how logistic regression can be used to produce a risk score such as the COMPAS general recidivism risk scale. We do so by fitting the $\boldsymbol{\beta}$ coefficients to a selection of variables from the COMPAS dataset published by Rudin et al. (2020a), where the target variable is a binary indicator whether the person in question recidivated or not (see Section 8.1 and Rudin et al., 2020a, for a more detailed description of the data). This is, presumably, representative of how the actual COMPAS risk scales were developed. For example, Brennan et al. (2009) stated that "The Recidivism Risk Scale is a regression model [...] that was trained to predict new offenses in a probation sample." This is very similar to the data contained in the COMPAS dataset published by Rudin et al. (2020a), as it does not only contain COMPAS risk scores, but also contains labels stating whether the person in question recidivated or not.

We select five features from the dataset, corresponding to some of the dimensions in the hierarchy we discussed in Example 8.3: **MS**, **Age**, **Priors**, and **Male**. Here, the number of priors is given in the data as the number of offenses committed in the 30 days leading up to the COMPAS assessment. The target variable $\boldsymbol{y}$ is the binary label indicating whether the person committed a new crime within two years after the assessment. We used the default Python *Sci-kit learn* implementation to estimate the $\boldsymbol{\beta}$ coefficients (Pedregosa et al., 2011).

The resulting logistic regression model can be specified as the GLM

$$(\sigma, -0.04, -0.2, -0.03, 0.11, 0.55),$$

which means that its prediction for values of the features **MS**, **Age**, **Priors**, **Male** is given by

$$(1 + \exp(0.04 + 0.2\mathbf{MS} + 0.03\mathbf{Age} - 0.11\mathbf{Priors} - 0.55\mathbf{Male}))^{-1} \tag{8.3.2}$$

Some example rows of the COMPAS dataset, together with their predicted (raw) general recidivism risk score (according to our example GLM), are displayed in Table 8.2. The $\mathbf{GRecid}_{\text{rs}}$ values are calculated according to Eq. (8.3.2).

8.3.2 The consistency of GLM decisions

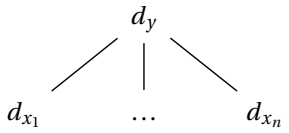
In this section we will prove our first main result regarding case base consistency, which states that the predictions made by a GLM are always fully consistent in the sense of Definition 8.8. To do this, we first show that any GLM can naturally be associated with a dimension hierarchy. Using this associated hierarchy, a dataset of GLM outputs can be translated to a case base $\mathscr{C}$ of which the consistency can be calculated with respect to its target variable. Theorem 8.16 below states that the consistency of such a dataset is necessarily equal to 1.

Consider an n-ary GLM $(m, \boldsymbol{\beta})$. We may, without loss of generality, assume that the coefficients β_i are all nonzero. This is because if any coefficient β_i were zero, the

Table 8.2: Example general recidivism risk assessments based on rows of the COMPAS dataset, according to the GLM $(\sigma, -0.04, -0.2, -0.03, 0.11, 0.55)$, where σ is defined in Eq. (8.3.1), and the parameters are estimated based on the data published by Rudin et al. (2020a). The example recidivism risk scores, i.e. the values of $\mathbf{GRecid}_{\text{rs}}$, are computed according to Eq. (8.3.2).

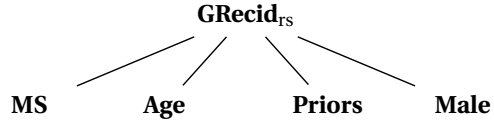
MS	Age	Priors	Male	$\mathbf{GRecid}_{\text{rs}}$
1	25	1	1	0.41
0	21	3	1	0.54
1	67	1	0	0.10
0	18	10	1	0.73
0	46	1	1	0.30

corresponding term $\beta_i x_i$ would not contribute to the linear predictor, and so the model's behavior would be the same as if that term were omitted. We define a set of dimensions $D = \{d_{x_i} \mid 1 \leq i \leq n\} \cup \{d_y\}$. Each d_{x_i} is the set of real numbers $\mathbb{R}$, and is ordered by $\leq$ if $\text{sign}(\beta_i) = 1$, and by $\geq$ if $\text{sign}(\beta_i) = -1$. Similarly, d_y is the set of real numbers $\mathbb{R}$, and is ordered by $\leq$ if m is order-preserving, and by $\geq$ if m is order-reversing. We order D by the structure $\mathsf{H} = \{(d_{x_i}, d_y) \mid 1 \leq i \leq n\}$:



Definition 8.12. Let $(m, \boldsymbol{\beta})$ be an n-ary GLM, and (D, H) its associated dimension hierarchy. An $(m, \boldsymbol{\beta})$-*decision* is a fact situation $X \in \mathscr{X}(D)$ with $\text{dom}(X) = D$, and $X(d_y) = m(\beta_0 + \sum_{i=1}^{n} \beta_i X(d_{x_i}))$. An $(m, \boldsymbol{\beta})$-*case base* is a case base $\mathscr{C} \subseteq \mathscr{X}(D)$ of $(m, \boldsymbol{\beta})$-decisions.

Example 8.13. Consider the GLM $(\sigma, -0.04, -0.2, -0.03, 0.11, 0.55)$ of Example 8.11, and note that its associated hierarchy is given by:



(8.3.3)

The dimension orders are $\geq$ for **MS** and **Age**, because their coefficients in the GLM are negative, and $\leq$ for **Priors** and **Male**, because their coefficients are positive. The order for $\mathbf{GRecid}_{\text{rs}}$ is $\leq$ because σ is (strictly) order-preserving. Relative to this hierarchy, the rows of Table 8.2 constitute a $(\sigma, -0.04, -0.2, -0.03, 0.11, 0.55)$-case base.

To prove our theorem, we will need two lemmas—the first of which states that GLM decisions naturally satisfy the a fortiori principle underlying the model of case-based reasoning.

Lemma 8.14. *For an n-ary GLM $(m, \boldsymbol{\beta})$ and $(m, \boldsymbol{\beta})$-decisions X and Y: If $Y(d_{x_i}) \preceq X(d_{x_i})$ for all $1 \leq i \leq n$, then $Y(d_y) \preceq X(d_y)$. Similarly, if $X(d_{x_i}) \preceq Y(d_{x_i})$ for all $1 \leq i \leq n$, then $X(d_y) \preceq Y(d_y)$.*

Proof. The first implication can be derived as follows:

$$Y(d_{x_i}) \preceq X(d_{x_i}) \text{ for } 1 \leq i \leq n \tag{8.3.4}$$
$$\text{implies} \quad \beta_i Y(d_{x_i}) \preceq \beta_i X(d_{x_i}) \text{ for } 1 \leq i \leq n \tag{8.3.5}$$
$$\text{implies} \quad \beta_0 + \textstyle\sum_{i=1}^{n} \beta_i Y(d_{x_i}) \preceq \beta_0 + \sum_{i=1}^{n} \beta_i X(d_{x_i}) \tag{8.3.6}$$
$$\text{implies} \quad m\big(\beta_0 + \textstyle\sum_{i=1}^{n} \beta_i Y(d_{x_i})\big) \preceq m\big(\beta_0 + \sum_{i=1}^{n} \beta_i X(d_{x_i})\big) \tag{8.3.7}$$
$$\text{implies} \quad Y(d_y) \preceq X(d_y). \tag{8.3.8}$$

Step (8.3.5) follows by definition of the dimension order of d_{x_i}: If $\beta_i > 0$ then $\preceq = \leq$ so $Y(d_{x_i}) \leq X(d_{x_i})$ and $\beta_i Y(d_{x_i}) \leq \beta_i X(d_{x_i})$; while if $\beta_i < 0$ then $\preceq = \geq$, so $Y(d_{x_i}) \geq X(d_{x_i})$ and $\beta_i Y(d_{x_i}) \leq \beta_i X(d_{x_i})$. Step (8.3.7) follows by monotonicity of m and the definition of the dimension order of d_y, and step (8.3.8) follows by Definition 8.12. The proof of the second implication is analogous so we omit it. □

The second lemma uses the first to show that when a GLM case base induces constraint on one of its decisions, then this is necessarily the results of disjunct (2) of Definition 8.5.

Lemma 8.15. *For an n-ary GLM $(m, \boldsymbol{\beta})$, an $(m, \boldsymbol{\beta})$-case base $\mathscr{C}$, some fact situation $X \in \mathscr{C}$, and a value $v \in d_y$: If $\mathscr{C} \models v \preceq X(d_y)$ then $v \preceq X(d_y)$. Similarly, if $\mathscr{C} \models X(d_y) \preceq v$ then $X(d_y) \preceq v$.*

Proof. Assume $\mathscr{C} \models v \preceq X(d_y)$, we proceed by a case distinction on the disjuncts (1)–(3) in Definition 8.5. We can rule out condition (1) because d_y does not have a minimal element. If condition (2) holds we are done immediately. Lastly, if condition (3) holds, then for some $Y \in \mathscr{C}$ with $v \preceq Y(d_y)$ we have that $Y(d_{x_i}) \preceq X(d_{x_i})$ for all $1 \leq i \leq n$. Hence, we get $Y(d_y) \preceq X(d_y)$ from Lemma 8.14, and so $v \preceq X(d_y)$ by transitivity of $\preceq$. The proof of the second implication follows the same pattern, so we omit it. □

It is now straightforward to derive our first main result from Lemma 8.15: a case base of GLM decisions is always fully consistent.

Theorem 8.16. *If $(m, \boldsymbol{\beta})$ is a GLM, (D, H) its associated hierarchy, and $\mathscr{C} \subseteq \mathscr{X}(D)$ an $(m, \boldsymbol{\beta})$-case base, then* $\mathrm{Cons}_{d_y}(\mathscr{C}) = 1$.

Proof. Assume, for sake of contradiction, that $\mathrm{Cons}_{d_y}(\mathscr{C}) < 1$; then there is a case $X \in \mathscr{C}$ which is d_y-inconsistent. This means that both $\mathscr{C} \models X(d_y) \preceq v$ and $\mathscr{C} \models w \preceq X(d_y)$ for some values $v, w \in d_y$ with $v \prec w$. By Lemma 8.15 this implies $X(d_y) \preceq v$ and $w \preceq X(d_y)$, so $X(d_y) \prec X(d_y)$, meaning $X(d_y) \neq X(d_y)$—a contradiction. □

It is important to note that Theorem 8.16 relies on an accurate construction of the dimension hierarchy associated with the GLM. In practice, when the GLM is a black box, we cannot inspect the signs of its coefficients, or whether its mean function is order-preserving or -reversing. However, estimating the signs of the coefficients is a much easier task than estimating the precise values of the coefficients. Likewise, it is easier to determine whether the function connecting the linear predictor to the output is order-preserving or -reversing, than it is to precisely estimate the function itself.

8.3.3 The consistency of the COMPAS risk scores

We can now use our implementation from Section 8.2.3 to compute the consistency scores of the COMPAS risk scores, using the scale inputs as described in Section 8.1. The dimension orders we use for this analysis are in line with their description in Example 8.3; a complete overiew can be found in our source code. The results of this analysis are quite surprising: the consistency of the raw **GRecid**, **VRecid**, and **FTA** risk scores are all 0%—the opposite of what we would expect given Theorem 8.16.

There are multiple possible explanations for this. One possibility is that our estimations of the dimension orders are incorrect. However, we consider this unlikely, as the effects of the features involved are quite self-evident. For example, the majority of the features in the dataset correspond to criminal history, and it is clear that higher values of these features should generally lead to higher risk scores. Another possibility is that COMPAS depends on the features in a nonlinear manner, as hypothesized by Rudin et al. (2020a).

However, it seems to us that the most likely cause is the absence of dimensions from the dataset which were used by COMPAS to arrive at the scores. Reconsider, for example, Example 8.10 of inconsistency. Assigning $Z(\textbf{GRecid}) = 3$ makes Z **GRecid**-inconsistent with respect to the case base $\{X, Y\}$, because X lower-bounds Z on **GRecid** to 5. However, the recidivism risk scores assignment might in reality be based on an additional drug-problem dimension **Drugs**—a 1–10 score ordered by $\leq$—which is missing from our hierarchy. Suppose, furthermore, that $X(\textbf{Drugs}) = 8$ and $Z(\textbf{Drugs}) = 4$; then Z appears inconsistent in the version of the hierarchy that does not include the drug-problem dimension, but is in fact consistent with respect to the hierarchy that does include it. Indeed, the COMPAS general recidivism risk score does rely on such a dimension according to Dieterich et al. (2016), and this dimension is not present in the data that we have used, so it is highly likely that our analysis is suffering from the effect illustrated by this example.

8.4 The effects of binning on consistency

In the second half of this chapter we consider the effects of binning input dimensions on consistency. Data binning is a preprocessing technique used in machine learning and data science to convert continuous numeric variables into categorical ones by subdividing a range of values into smaller, consecutive, non-overlapping intervals called bins (Nisbet et al., 2018, Chapter 4). A common application of data binning is the histogram, which visualizes the distribution of a dataset by replacing individual data points with their corresponding bins, thus smoothing the data and making general trends easier to see.

We begin by proving a theorem stating that binning input dimensions decreases case base consistency. We then show that this theorem can be used to find a problem with the COMPAS dataset that has not received much attention in the literature, namely that multiple norm groups were used in the raw-to-decile score conversions. This means that the decile scores, which are often the focus of data analyses based on the COMPAS dataset, can only be compared with each other after checking that they were normed using the same group. This is difficult because information on the norm groups is not made available by Equivant. We conclude by demonstrating that this information can, to some extent, be reverse engineered from the COMPAS dataset, by using a graph coloring algorithm.

8.4.1 Dimension binning

To begin, we need a formal definition of data binning within our framework, for which we propose the following.

Definition 8.17. Let d be a dimension; a *binning* (bin, e) of d is a dimension e together with a surjective order-preserving function $\mathrm{bin} : d \to e$. The elements of e may be referred to as *bins*.

For the sake of convenience we will refer to just the function $\mathrm{bin} : d \to e$ as the "binning" of d. The requirement that bin is order-preserving, which means that $v \preceq w$ implies $\mathrm{bin}(v) \preceq \mathrm{bin}(w)$, ensures that the order of the bins reflects the order of the original dimension d. The surjectivity requirement, which states that any bin $v \in e$ has a nonempty pre-image, ensures that all bins correspond to some region of the original dimension.

Example 8.18. An example of dimension binning in the context of COMPAS is given by the various presentations of the risk scores which we discussed in Section 8.1. Let $\mathbf{GRecid}_{\mathrm{rs}}$ denote the dimension corresponding to the raw version of the recidivism risk score, $\mathbf{GRecid}_{\mathrm{ds}}$ the decile version, and $\mathbf{GRecid}_{\mathrm{txt}}$ the textual version (i.e., with the possible values low, medium, and high). The conversion of the raw scores to decile scores corresponds to a binning $\mathrm{bin}_{\mathrm{ds}} : \mathbf{GRecid}_{\mathrm{rs}} \to \mathbf{GRecid}_{\mathrm{ds}}$. The way this mapping works depends on the specific norm group that is used for the conversion. Assuming the conversion specified by Equivant (2019, Table 2.3) can compute the pre-images of the bins; for example, $\mathrm{bin}_{\mathrm{ds}}^{-1}(4) = (-0.7, -0.4]$, and $\mathrm{bin}_{\mathrm{ds}}^{-1}(5) = (-0.4, -0.2]$, etc. In turn, we have a binning converting deciles scores to text $\mathrm{bin}_{\mathrm{txt}} : \mathbf{GRecid}_{\mathrm{ds}} \to \mathbf{GRecid}_{\mathrm{txt}}$:

$$\mathrm{bin}_{\mathrm{txt}}^{-1}(\mathrm{low}) = [1, 4],\ \mathrm{bin}_{\mathrm{txt}}^{-1}(\mathrm{med}) = [5, 7],\ \mathrm{bin}_{\mathrm{txt}}^{-1}(\mathrm{high}) = [8, 10].$$

Example 8.19. In general, for any dimension d there is a trivial *identity* binning $\mathrm{id}_d : d \to d$ defined by $\mathrm{id}_d(v) = v$ for all $v \in d$. This corresponds to putting every value of d in its own unique bin.

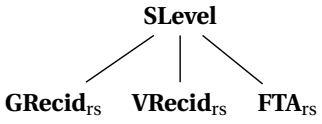
We want to investigate the effect that binning one or more dimensions has on case base consistency. To this end, we introduce the following definition.

Definition 8.20. Given a set of dimensions D, a *D-binning* is an assignment of a binning $\mathrm{bin}_d : d \to \underline{d}$ to every dimension $d \in D$. Furthermore, given $e \in D$, a D-binning is an *input binning of e* if $\mathrm{bin}_e = \mathrm{id}_e$.

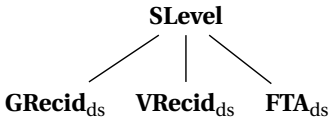
Given a hierarchy (D, H) with a D-binning, the binned version of a dimension $d \in D$ is denoted by $\underline{d}$. Likewise, we will write $\underline{D} = \{\underline{d} \mid d \in D\}$ for the set of all binned dimensions. This set can be given the same hierarchical structure as the original hierarchy by defining $\underline{\mathsf{H}} = \{(\underline{e}, \underline{d}) \mid (e, d) \in \mathsf{H}\}$. There is a canonical way of transforming a fact situation $X \in \mathscr{X}(D)$ into a fact situation $\underline{X} \in \mathscr{X}(\underline{D})$ for the binned version of the hierarchy, by defining $\underline{X}(\underline{d}) = \mathrm{bin}_d(X(d))$. This operation extends in the obvious way to a case base $\mathscr{C} \subseteq \mathscr{X}(D)$: We define $\underline{\mathscr{C}} = \{\underline{X} \mid X \in \mathscr{C}\} \subseteq \mathscr{X}(\underline{D})$.

Example 8.21. An example of an input binning is given by the various presentations of the COMPAS risk scores, in their connection to the recommended supervision level score.

Together the binnings $\mathbf{GRecid}_{\mathrm{rs}} \to \mathbf{GRecid}_{\mathrm{ds}}$, $\mathbf{VRecid}_{\mathrm{rs}} \to \mathbf{VRecid}_{\mathrm{ds}}$ and $\mathbf{FTA}_{\mathrm{rs}} \to \mathbf{FTA}_{\mathrm{ds}}$ give an input **SLevel** binning, with associated hierarchies:



(8.4.1)



(8.4.2)

8.4.2 Input binning decreases consistency

We now prove the second theorem of this chapter, which states that input binning decreases case base consistency. In fact this result follows readily from the following lemma, which states that binning preserves constraint.

Lemma 8.22. *Consider a hierarchy (D, H) with a D-binning. For a dimension $d \in D$, a value $v \in d$, a fact situation $X \in \mathscr{X}(D)$, and a case base $\mathscr{C} \subseteq \mathscr{X}(D)$: If $\mathscr{C} \models v \preceq X(d)$ then $\underline{\mathscr{C}} \models \mathrm{bin}(v) \preceq \underline{X}(\underline{d})$; and similarly, if $\mathscr{C} \models X(d) \preceq v$ then $\underline{\mathscr{C}} \models \underline{X}(\underline{d}) \preceq \mathrm{bin}(v)$.*

Proof. We proceed by structural induction on the position of d in H, and apply a case distinction on $\mathscr{C} \models v \preceq X(d)$.

(1) If v is the least element of d, then $\mathrm{bin}(v)$ is the least element of $\underline{d}$ because bin is order-preserving and surjective, so that indeed $\underline{\mathscr{C}} \models \mathrm{bin}(v) \preceq \underline{X}(\underline{d})$.

(2) Likewise, if $v \preceq X(d)$ then because bin is order-preserving we have $\mathrm{bin}(v) \preceq \mathrm{bin}(X(d)) = \underline{X}(\underline{d})$, so $\underline{\mathscr{C}} \models \mathrm{bin}(v) \preceq \underline{X}(\underline{d})$.

(3) If $\mathscr{C}, Y \models v \preceq X(d)$, then $d \in A$, $v \preceq Y(d)$, and for all $e \in \mathsf{H}(d) \cap \mathrm{dom}(Y)$: $\mathscr{C} \models Y(e) \preceq X(e)$. The induction hypothesis states that for all $w \in e \in \mathsf{H}(d)$: $\mathscr{C} \models w \preceq X(e)$ implies $\underline{\mathscr{C}} \models \mathrm{bin}(w) \preceq \underline{X}(\underline{e})$. We are done if $\underline{\mathscr{C}}, \underline{Y} \models \mathrm{bin}(v) \preceq \underline{X}(\underline{d})$. Note that $\underline{d}$ is abstract in $(\underline{D}, \underline{\mathsf{H}})$, and that $\mathrm{bin}(v) \preceq \mathrm{bin}(Y(d)) = \underline{Y}(\underline{d})$ as bin is order-preserving. Therefore, it only remains to show that for all $\underline{e} \in \underline{\mathsf{H}}(\underline{d}) \cap \mathrm{dom}(\underline{Y})$: $\underline{\mathscr{C}} \models \underline{Y}(\underline{e}) \preceq \underline{X}(\underline{e})$, so consider such $\underline{e}$. By definition of $\underline{\mathsf{H}}$ and $\underline{Y}$, this means that $e \in \mathsf{H}(d) \cap \mathrm{dom}(Y)$ and so $\mathscr{C} \models Y(e) \preceq X(e)$, which means $\underline{\mathscr{C}} \models \underline{Y}(\underline{e}) \preceq \underline{X}(\underline{e})$ follows from the induction hypothesis.

The proof of the other implication is analogous, so we omit it. □

Theorem 8.23. *Given a dimension hierarchy (D, H), a case base $\mathscr{C} \subseteq \mathscr{X}(D)$, and an input D-binning of $d \in D$, we have* $\mathrm{Cons}_d(\underline{\mathscr{C}}) \leq \mathrm{Cons}_d(\mathscr{C})$.

Proof. It suffices to show that if $X \in \mathscr{C}$ is d-inconsistent with respect to $\mathscr{C}$, then $\underline{X}$ is d-inconsistent with respect to $\underline{\mathscr{C}}$, so consider such $X \in \mathscr{C}$. This means there are $v \prec w \in d$ with $\mathscr{C} \models X(d) \preceq v$ and $\mathscr{C} \models w \preceq X(d)$. Lemma 8.22 gives us $\underline{\mathscr{C}} \models \underline{X}(d) \preceq v$ and $\underline{\mathscr{C}} \models w \preceq \underline{X}(d)$, so $\underline{X}$ is indeed d-inconsistent with respect to $\underline{\mathscr{C}}$. □

Intuitively, the reason for this result is that as we bin input dimensions, it becomes easier to satisfy disjunct (3) of Definition 8.5 of constraint. As such, more constraint is induced, and so there are more opportunities for inconsistencies to arise.

Table 8.3: Two examples of fact situations X, Y in the COMPAS dataset that are **SLevel**-inconsistent with respect to their raw scores, but **SLevel**-consistent with respect to their decile scores. The input dimensions are the three risk scores produced by COMPAS.

	FTA		**GRecid**		**VRecid**		**SLevel**
	Raw	**Decile**	**Raw**	**Decile**	**Raw**	**Decile**	
Y	21	3	0.14	7	−0.95	9	3
X	19	3	0.11	8	−1.21	8	4

8.4.3 Consistency of the recommended supervision levels

For our second analysis, we look at the consistency percentage of the recommended supervision level (**SLevel**) scores in the dataset. This score is an overall recommendation that the COMPAS program outputs, based on its various needs and risk scales (Equivant, 2019). This recommendation is primarily based on the general recidivism and violent recidivism risk scores, and so—unlike in the first analysis—we can now be (more or less) certain that we are using the same dimensions as the ones used by the COMPAS program.

Each of the datasets we examined contained approximately 12,000 rows. The dimension orders were assigned based on the assumption that there is a positive correlation between the risk scores and the recommended supervision level, as described in Section 8.3. We used the three risk scores as the inputs, so that the resulting hierarchies are the same as the ones depicted in Example 8.21. Our implementation reports that the consistency of the data with respect to (8.4.1) is 81%, and with respect to (8.4.2) it is 100%.

In other words, the consistency of the data increases after applying a binning to the input dimensions, which directly contradicts Theorem 8.23. An examination of this contradiction reveals that the cause lies in the use of multiple norm groups for the raw-to-decile score conversion. To see this, consider the two fact situations X and Y listed in Table 8.3, taken from the dataset. If we look at the raw scores, we have the constraint $\{Y\} \vDash X(\textbf{SLevel}) \preceq 3$ because X assigns lower values to all three of the raw COMPAS risk scores. This means that X is **SLevel**-inconsistent because $X(\textbf{SLevel}) = 4 \succ 3$. However, when we look at the decile scores rather than the raw scores we get a different picture, because according to those we have $X(\textbf{GRecid}) = 8 \not\preceq 7 = Y(\textbf{GRecid})$; so Y no longer constrains X, and X has become **SLevel**-consistent. Of course, this should not happen: when a general recidivism risk score of 0.11 is higher than 70–80% of the normative group, then 0.14 should also be higher than 70–80% of the normative group.

Remark 8.24. A manual of the version of COMPAS used in the state of New York explicitly states how its recommended supervision levels are computed (New York State Division of Criminal Justice Services, Office of Probation and Correctional Alternatives, 2015, Appendix C). This description matches the scores found in the Broward County data, which suggests that the version of COMPAS used there computes the recommendation in the same way. This would also explain why the decile scores give a higher consistency percentage than the raw scores.

Table 8.4: The reconstructed cut-points of norm groups used for the raw-to-decile score conversions in the COMPAS dataset. We indicate the highest score for each given decile, in accordance with Equivant (2019, Table 2.3). The dashes indicate that there were no defendants grouped in that particular decile of the hypothesized norm group, and so no information is available on its cut-point.

Score	Order	Male	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
	5,230	×	−1.66	−1.31	−1.03	−0.82	−0.59	−0.37	−0.14	0.12	0.43	2.36
GRecid	4,440	✓	−1.39	−0.92	−0.60	−0.39	−0.19	0.01	0.19	0.39	0.67	—
	2,840	✓	—	—	−0.92	−0.60	−0.39	−0.19	0.01	0.19	—	—
VRecid	12,510	×	−2.95	−2.56	−2.24	−1.98	−1.74	−1.50	−1.26	−1.00	−0.63	0.93
	6,631	×	16	19	21	23	25	27	29	31	35	50
FTA	4,950	×	—	16	19	21	22	24	26	28	31	35
	929	✓	—	—	—	—	23	25	27	29	—	—

8.4.4 Reconstructing the norm groups

Decile scores computed with different norm groups should not be compared, as they are on different scales. However, many studies, including the original publication by Angwin et al. (2016), do exactly that. A more accurate approach uses raw scores, as that of Rudin et al. (2020a), but users prefer decile or textual scores (Brennan & Dieterich, 2018). The study by Engel et al. (2024) renorms decile scores based on all raw scores, which lowers the average decile risk score and bases the study on hypothetical rather than actual COMPAS output.

To compare decile risk scores accurately, they should be split according to the norm groups used to convert the scores. This is difficult because Equivant does not make their norm groups public. We estimate the number of norm groups by constructing a graph of COMPAS risk assessments, drawing edges between nodes with mutually inconsistent raw-to-decile score conversions. Applying a graph coloring algorithm (Hagberg et al., 2008), we identified three norm groups; see Table 8.4 for their cut-points. The **GRecid** and **FTA** graphs have fully connected subgraphs of size 3, indicating at least three norm groups for these scores. The **VRecid** scores were converted using a single norm group.

We note that, because there is some overlap in the cut-points of the groups, defendants can be "shuffled around." For example, the size of the second largest group of the **GRecid** score could be increased to at least 6,698. Therefore, those wishing to analyze the recidivism scores and the outcome labels could analyze this group in isolation, without having to re-norm the deciles.

Some of the reconstructed groups consist of over over 99.5% males. Since the graph labeling algorithm we used was not in any way instructed to group defendants by sex (in fact, it did not even have access to this information), we consider it likely that gendered norm groups were used for the raw-to-decile conversion in the dataset. We have marked these groups in Table 8.4. Moreover, one of these groups (corresponding to the second row of Table 8.4) closely aligns with the cut-points given by Equivant (2019, Table 2.3) for the GRRS. Due to this, we consider it likely that our reconstructed cut-points are good approximations of the true cut-points.

8.5 Conclusion

In this chapter, we analyzed the consistency of the COMPAS risk scores using the dimensional-hierarchical result model that we introduced in Chapter 5. We proved two formal results on case base consistency, and related them to this analysis. Firstly, we showed that generalized linear models always produce fully consistent case bases. We showed that the COMPAS dataset is not fully consistent (in fact, it is fully *in*consistent), and conclude that this is due to input features missing from the data. Secondly, we showed that binning input dimensions generally decreases consistency. However, our analysis of the COMPAS recommended supervision level scores revealed an increase in consistency after binning, which we attribute to the use of multiple norm groups for the raw-to-decile score conversions. We demonstrated that these norm groups can be reverse-engineered from the dataset using a graph coloring algorithm, providing a more accurate basis for analyzing the COMPAS risk scores.

We remark, as a point for future work, that Rudin et al. (2020a) found that many defendants with numerous prior offenses received low risk scores in the COMPAS dataset. Van Woerkom et al. (2024a) showed that such outlier cases, there termed landmarks, can significantly increase inconsistency in a dataset. We suspect that these cases might overlap with those identified by Rudin et al. (2020a, Table 6), potentially explaining the low consistency percentages observed for the COMPAS scores in this chapter. If this is indeed the case, this would suggest that inconsistency measures could be useful for detecting outliers in similar datasets.

Chapter 9

Conclusion

FROM THE OUTSET, this thesis pursued a dual aim: to develop a general theory of a fortiori reasoning, and to study its application to the analysis and justification of decisions made by data-driven artificial intelligence systems. After introducing these subjects in Chapter 1, we developed variations of the result model (RM) of precedential constraint, introduced by Horty (2011). In Chapter 2 we reviewed the RM, and considered how to incorporate incomplete information in its knowledge representation. In Chapter 3 we conducted a similar analysis for the dimension-based version of this model, the DRM, and related it to order theory and logic. In Chapter 4, we extended the RM to operate on factor hierarchies, resulting in the HRM. Then, as the final chapter of Part I, we introduced the DHRM, which incorporates both dimensional and hierarchical information. In Part II we turned to applications of these models to artificial intelligence. Beginning in Chapter 6, we reviewed the a fortiori case-based argumentation method of explaining data-driven decisions, and expanded the method with formal notions of justification, compensation, and citability. In Chapter 7, we used the theory developed in Chapter 3 to build an implementation of the DRM based on the Z3 SMT solver. This was used to analyze the consistency of several machine learning datasets. Lastly, in Chapter 8, we extended this implementation to compute constraint produced by the DHRM, and used this to analyze the consistency of the COMPAS risk scores.

9.1 Answers to the research questions

We will now recall and answer the research questions set out at the beginning of this thesis in Chapter 1. The first of these, addressed in Part I, concerns expanding the result model of precedential constraint to a general theory of a fortiori reasoning:

> **Research question 1:** Can we extend the result model of precedential constraint to a general theory of a fortiori case-based reasoning?

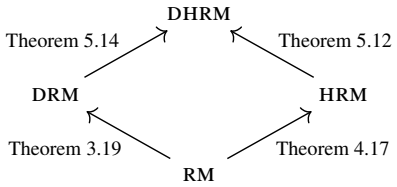


Figure 9.1: A graphical representation of the conservative extension relations between the models we introduced; e.g., the arrow from RM to DRM indicates that all reasoning done in an instance of the RM can be faithfully reproduced in an instance of the DRM.

In particular, the knowledge representation used by the RM makes simplifying assumptions with regards to the inputs of the model that do not always hold in practice.

> **Research question 1A:** How can incomplete, dimensional, and hierarchical information be incorporated in the knowledge representation, and what should the corresponding notion of constraint be?

We answered this question by first modifying the RM and the DRM to operate on the basis of incomplete information. We proposed a modified principle of constraint, according to an assumption that the available information was sufficient to reach a decision. As such, we adapted the notion of constraint accordingly, to operate only on the factors or dimensions that have known truth values. Subsequently, we introduced the HRM, which operates on the basis of a factor hierarchy. We developed a notion of constraint for the HRM that utilizes this hierarchical structure by means of recursion. This enables multi-case precedential reasoning. Lastly, we introduced the DHRM, which operates on the basis of a dimension hierarchy. In this model, outcomes of the decision-maker are nonbinary, and accordingly we proposed a notion of constraint in the form of lower- and upper-bounds.

> **Research question 1B:** How can the models developed in response to 1A be formally compared, and what are their differences and similarities?

We showed that the models we developed in response to Question 1A are conservative extensions of each other. This relies on two observations: firstly, a factor can be interpreted as a binary dimension; and secondly, the RM can be interpreted as a "flat" factor hierarchy, which moves in one step from the base-level factors to a decision. We gave formal proofs of each of these relations, which are visualized in Figure 9.1. This showed that the DHRM subsumes the other models and, as such, is the most general of the models of a fortiori reasoning that we presented. Furthermore, we saw that the DHRM is also more expressive than the other models which we considered, because it produces constraint in the form of lower- and upper-bounds, rather than directly for either of two sides.

Research question 1C: What is the relation between this theory and other reasoning formalisms, such as logic?

We have positioned the DRM in the context of order theory and logic. Key to this association is the observation that the *strength order* of the DRM corresponds to the *product order* of order theory. Through this connection, constraint in the DRM can be understood in terms of up- and down-sets. This provides a clarifying perspective on the DRM. For example, the notion of what we call a *landmark case*—which is a case that does not have its outcome forced by the rest of a case base—can be understood as a minimal element in the order-theoretic sense. Building on this, we showed that the DRM can be phrased in terms of many-sorted logic. This led to new insights. In particular, both theories define a notion of inconsistency, and we showed that these are related in a precise sense. Furthermore, an important topic in the literature on logic is that of *nonmonotonicity*. We connected our formalism to this topic by phrasing a monotonicity principle in the setting of a fortiori reasoning, and gave formal proofs that our models satisfy this principle. In a similar vein, we briefly compared our notion of constraint for incomplete information to the notion of *default* truth values in nonmonotonic logic.

In Part II, we addressed the second of our research questions, pertaining to applications of the theory of a fortiori reasoning to artificial intelligence:

Research question 2: How can the models of a fortiori precedential constraint be applied to artificial intelligence?

In particular, we reviewed the AF-CBA method of explaining data-driven decisions proposed by Prakken and Ratsma (2022), and asked the following question:

Research question 2A: How can the theory of a fortiori precedential constraint be used to formalize compensation and citability, to aid in justifying data-driven decisions?

The AF-CBA model provides explanations as winning strategies in the grounded argument game of an abstract argumentation theory. We showed that this model admits an equivalent rephrasing in terms of relations, in which explanations are provided as cases related to the focus case through justification and citation relations. Most notably, this shows that the explanation model can, in some sense, be seen as adding a notion of justification to the theory of precedential constraint, in the form of a relation that extends the forcing relation. We further introduced criteria for selecting precedent cases that minimize differences while maximizing similarities to the focus case in order to provide an appropriate starting point for the explanation.

Research question 2B: How can we write capable and efficient computer implementations of these models?

To answer this question, we used the connection that we established between the models of a fortiori reasoning and logic in response to Question 1C, in order to write an implementation in the Z3 SMT solver. We showed how this can be used to compute constraint, to check whether a case is a landmark, and to give automatic proofs of case base properties such as consistency and completeness. We demonstrated the efficacy of this implementation on a number of datasets. As part of this implementation, we proposed a way to use logistic regression in order to determine the dimension orders used by the DRM and the DHRM, and showed that it performs well for the datasets that we considered.

Research question 2C: Is precedential constraint useful as a measure of data-driven decision consistency?

For this question, we used our implementation to analyze the internal consistency of the recidivism risk scores produced by the COMPAS program. By comparing the results of this analysis to formal results providing predictions about how these scores should behave, we were able to discover inconsistencies in the data that many previous studies on this data did not take into account. This affirms that the notion of constraint can be useful for measuring artificial intelligence.

9.2 Relevance for artificial intelligence and law

In the past, AI researchers have devoted much of their attention to the development of systems of ever-increasing complexity. The results speak for themselves: modern systems are capable of outperforming humans in many domains, and have become so complex that it is no longer possible to have a complete understanding of how they function. This problem is compounded by the proprietary nature of many of these systems—even if they could be understood by a human, their inner workings are hidden behind intellectual property rights. Naturally, this has raised concerns regarding their use in sensitive contexts, in which decisions made by AI systems have social, ethical, or legal consequences. Increasingly, interest in AI research is rediverted to the *responsible* use of AI, which is focused on developing AI techniques that emphasize the role of humans in their interaction with AI systems. Examples of topics receiving attention are the development of methods to explain AI decisions, mitigating bias, aligning decisions with international law and intellectual property rights, ensuring privacy, and combating misinformation.

In recent years, models of CBR from the AI & law literature have been applied to this purpose: to explain data-driven decisions (Čyras et al., 2016; Prakken & Ratsma, 2022), to develop interpretable decision systems (Odekerken & Bex, 2020), and to formulate computational normative reasoning principles (Canavotto & Horty, 2022). This dissertation fits in this line of research: We developed a general model of a fortiori reasoning, together with an efficient software implementation, and applied these to case studies that demonstrated their efficacy with respect to analyzing data-driven decisions. We now discuss the relation between the findings of this thesis and the goals of the responsible AI research line, and AI & law in general, before closing with thoughts on future research directions.

Firstly, we consider the goal of explaining, or justifying, data-driven decisions. Research in this direction has been one of our primary motivators to further develop the formal theory of case-based reasoning. Miller (2019) argued that explanations should be selective, contrastive, and are fundamentally social. As such, explanations should not, for example, overemphasize probabilities associated to the black box system, as this type of explanations tend to be unsatisfying to people. It is generally better to provide a causal explanation, focused on generalisations. Accordingly, Miller (2019) argues that research on explainable AI should draw on the social sciences literature in order to meet these standards. This is what Prakken and Ratsma (2022) have aimed to do through the development of their AF-CBA explanation method: to draw on the way in which courts explain and justify decisions, and apply these techniques to explain data-driven decisions. While the generation of argumentative dialogue was not the focus of this thesis, we did aim to contribute to the development of such explanation methods by expanding the theoretical basis on which they are built. For example, courts often make use of hierarchically structured concepts, or factors, to explain their reasoning, and we have incorporated these in the model of case-based reasoning (Bench-Capon, 2024; Canavotto & Horty, 2023a).

It has often been argued that explanations of AI decisions should have the form of an argumentative dialogue (Miller, 2019; Prakken & Ratsma, 2022; Verheij, 2020). In an ideal scenario, an end-user could engage in a back-and-forth conservation with the system, in natural language, in which arguments for a decision are sequentially provided and questioned. Work such as that by Prakken and Ratsma (2022) draws on argumentation theory and models of reasoning which are akin to logic to try to reach this goal. However, recently developed AI systems based on large language models (LLMs) seem to have off-the-shelf capabilities for providing this gold standard of explanations (OpenAI, 2024). For example, by using *chain-of-thought* prompting, LLMs can give answers to questions by providing a sequence of reasons which can then individually be challenged or questioned (Wei et al., 2022). These developments would seem to suggest that symbolic approaches to explanation, such as the AF-CBA method, have been rendered obsolete.

However, in practice the self-explanation capabilities of LLMs lack *fidelity* (Parcalabescu & Frank, 2024), meaning their explanations do not necessarily reflect the true reasons behind their decisions (Molnar, 2024). Turpin et al. (2023) showed that chain-of-thought explanations can be biased by input features, leading models to rationalize incorrect answers without mentioning the biasing factors in their self-explanation. Related to this is the problem of LLM *hallucinations*, which refers to the model presenting incorrect but plausible sounding information. Dahl et al. (2024) have shown that the mean hallucination rate of popular LLMs on common legal tasks lies between 69% and 88%. Ironically, the problem of determining whether a self-explanation accurately reflects the internal reasoning, or whether it is just an incorrect but plausible sounding rationalization, is exactly the type of problem for which we need explanation methods to begin with. Work is being done on improving the self-explanatory capabilities of LLMs (Chuang et al., 2024), but we believe that it is clear that there is still merit to symbolic, argumentation-based approaches, which are inherently reliable and transparent.

A risk associated with black-box machine learning systems, and LLMs in particular, is their tendency to perpetuate bias that is present in the data used to train them. For example, bias has been present in the earliest forms of word embeddings (on the basis of which LLMs operate) and have turned out to be very difficult to remove (Kurita et al., 2019). Another

prominent example, which we discussed in Chapters 7 and 8, is the COMPAS recidivism risk prediction system, and it has often been claimed that it is biased with regards to race and age—in fact, Engel et al. (2024) have recently argued that COMPAS is biased against *all* defendants. Bias is a complicated topic. The debate surrounding COMPAS is partially due to its proprietary nature, but it is also due to the difficulty with giving a universal and precise definition of fairness. Consider, for example, the issue of recidivism *base rates*. It is well-known that younger individuals tend to recidivate more often, which is to say that the group of young individuals has a higher recidivism base rate than the group of old individuals. A data-driven risk prediction system will pick up on this trend and assign higher risk scores to young individuals, and this leads to scenarios where a person's age determines if they remain detained or not. Whether this is fair depends on one's definition of fairness; on the one hand, age is not something a person has any control over, but on the other hand it is a highly predictive feature. As a matter of fact, it is prohibited in New York to base a pretrial risk assessment on age (Equivant, 2019).

We believe that the models presented in Part I of this thesis encode a formal stare decisis principle for data-driven decisions, which can be used as a concrete definition of fairness. More research is needed to ascertain the usefulness of this definition, but we think that our findings in Part II of this thesis are promising. Furthermore, we think that stare decisis is a natural principle which people tend to agree on. Consider, for instance, the example presented by Tripathi et al. (2024, Figure 1), which demonstrates that an LLM can change opinion on the applicability of a law to a certain case only when seemingly irrelevant attributes such as the subject's name and religion are changed. It seems clear that a stare decisis principle underlies the surprise expressed at such an example: Similar cases should be treated similarly.

An alternative to the attempts of mitigating the risks associated with black-boxes is simply to use systems that are inherently interpretable. Rudin (2019) warns that attempts to explain black-box are potentially dangerous as explanations can be misleading, and suggests to exclusively use interpretable models for high-stakes decisions. In fact, Rudin (2019) argues that the commonly held conception that there is a trade-off between accuracy and interpretability is a myth. For example, it has been shown repeatedly that simple predictive models can achieve the same accuracy as complicated black-box models on the task of predicting recidivism; see Rudin and Radin (2019) for references. It should be noted, however, that these claims predate the sudden rise of LLMs, of which the performance is currently unmatched by interpretable models.

Models of a fortiori case-based reasoning have been used by Odekerken (2025) as part of interpretable decision systems for fraud intake and the classification of web shops. These are both implemented and used by the Dutch Police Force. This indicates that our models from Part I can also be used for this purpose. In this vein, we investigated in Chapter 3 under which conditions the DRM constitutes a *classifier* in the machine-learning sense of the word. It should be noted, however, that the DHRM does not lend itself as well to regression tasks, because its notion of constraint produces lower- and upper-bounds, rather than precise outcomes. Furthermore, the dimensions constitute quite a strong assumption that the features involved in the classification or prediction task are linearly correlated with the output, and this is an assumption that is not always satisfied, particularly in non-legal contexts.

9.3 Future work

We conclude by outlining some directions for future research.

Order theory and logic Our comparison between the a fortiori models of reasoning and order theory and logic has focused on the DRM, and to a lesser extent on the DHRM. An evident continuation of this comparison would focus more on the DHRM, and in particular, work out the details of the way in which the recursive aspect of the DHRM is handled in the logical framework.

Incomplete information We have proposed a notion of constraint in the context of incomplete information. Concurrently, Odekerken et al. (2023a) has investigated very similar questions. A preliminary comparison between our proposed definitions and that of Odekerken et al. (2023a) suggests that there is agreement between our work, but we have not worked this out thoroughly.

Computing stability In the same vein, Odekerken et al. (2023b) have proposed algorithms for computing possible assignments of values to undefined dimensions, so that the resulting fact situations are forced for a side. This is exactly the type of question that our Z3-based implementation can answer, because Z3 tries to find satisfying assignments. For example, a question such as, "Is there a value that can be assigned to dimension d so that outcome s is forced?" is native to the language of an SMT solver such as Z3. It would be interesting to explore this application in more detail and to see whether Z3 can offer a performance improvement on this task.

The hierarchical reason model Throughout this dissertation, we focused on the result model of precedential constraint. When it comes to modeling common law jurisdictions, Horty (2011) prefers the *reason* model of constraint, which has similarly seen applications in Artificial Intelligence and Law (Canavotto & Horty, 2022). The reason model has also been recently extended by Canavotto and Horty (2023b) to operate on the basis of factor hierarchies. A superficial comparison of this extension and our HRM suggests the models are very comparable. It would be interesting to give a complete formal comparison. Specifically, the comparison between the RM and the reason model, which we gave in Section 2.7, could be done for the HRM and the model hierarchical reason model developed by Canavotto and Horty (2023b).

Expanding the AF-CBA method The AF-CBA explanation method, developed by Prakken and Ratsma (2022), operates on the basis of the DRM. As such, it can only be applied to binary classifiers—models that learn to sort an input into either of two classes. In practice, there are many AI systems that produce real-valued, nonbinary output, such as regression models, and to these the AF-CBA method cannot be applied. Modifying the AF-CBA method to operate on the basis of the DHRM, as opposed to the DRM, could rectify this. These modifications would likely be nontrivial, as they would require changes to the underlying argumentation framework of the method.

Further data analyses The result model, and variations thereof, have initially been proposed as a model of human decision-making. It would therefore be interesting to see whether decisions of courts and judges are consistent in the sense proposed by these models. An example of where such an analysis could be performed is in the domain of bail decisions, which has featured as a running example in this dissertation. A dataset was made available by Williams and Kolter (2021) of cash bail decisions made by magistrates in the Allegheny and Philadelphia counties in Pennsylvania. Our DHRM is capable of quantifying the decision consistency of these magistrates, so our Z3 implementation could be used to compute them based on the available dataset.

Determining dimension orders As part of our software implementation, we have proposed a logistic regression-based method for estimating appropriate choices for the dimension orders of the a fortiori models. This method is a modified version of the one proposed by Prakken and Ratsma (2022), who use Pearson correlation coefficients, and we compared the performance of these methods. Since there exist many other measures of correlation, it is easy to come up with alternatives to these approaches, and we have not given a comprehensive comparison of these options. In a similar vein, our formal results regarding the DHRM consistency of generalized linear models require an accurate reconstruction of the signs of the coefficients used by the linear model, and of the direction of its mean function. The feasibility of estimating this information from a dataset could be further explored; for example, how many datasets would be needed to obtain a 95% confidence that the signs are correctly estimated?

Bibliography

Acharya, M. S., Armaan, A., & Antony, A. S. (2019). A comparison of regression models for prediction of graduate admissions. *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, 1–5. https://doi.org/10.1109/ICCIDS.2019.8862140

Aleven, V. (1997). *Teaching case-based argumentation through a model and examples* [Doctoral dissertation]. https://doi.org/10.5555/926270

Aleven, V. (2003). Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment. *Artificial Intelligence*, *150*(1), 183–237. https://doi.org/10.1016/S0004-3702(03)00105-X

Aleven, V., & Ashley, K. D. (1997). Evaluating a learning environment for case-based argumentation skills. *Proceedings of the Sixth International Conference on Artificial Intelligence and Law*, 170–179. https://doi.org/10.1145/261618.261650

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2018). Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*, *18*(234), 1–78. http://jmlr.org/papers/v18/17-716.html

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias [newspaper]. *ProPublica*. Retrieved February 26, 2024, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Ashley, K. D. (1991). Reasoning with cases and hypotheticals in HYPO. *International Journal of Man-Machine Studies*, *34*(6), 753–796. https://doi.org/10.1016/0020-7373(91)90011-U

Ashley, K. D. (1992). Case-based reasoning and its implications for legal expert systems. *Artificial Intelligence and Law*, *1*(2), 113–208. https://doi.org/10.1007/BF00114920

Atkinson, K., Bench-Capon, T., & Bollegala, D. (2020). Explanation in AI and law: Past, present and future. *Artificial Intelligence*, *289*, 103387. https://doi.org/10.1016/j.artint.2020.103387

Barenstein, M. (2019). *ProPublica's COMPAS data revisited*. arXiv: 1906.04711 [cs, econ, q-fin, stat]. https://doi.org/10.48550/arXiv.1906.04711

Bench-Capon, T. (1993). Neural networks and open texture. *Proceedings of the Fourth International Conference on Artificial Intelligence and Law*, 292–297. https://doi.org/10.1145/158976.159012

Bench-Capon, T. (2023). The role of intermediate factors in explaining precedential constraint. *Proceedings of the 23nd Workshop on Computational Models of Natural Argument*, *3614*, 21–32. https://ceur-ws.org/Vol-3614

Bench-Capon, T. (2024). Intermediate factors and precedential constraint. *Artificial Intelligence and Law*. https://doi.org/10.1007/s10506-024-09405-x

Bench-Capon, T., & Atkinson, K. (2021). Precedential constraint: The role of issues. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 12–21. https://doi.org/10.1145/3462757.3466062

Berkman Klein Center for Internet and Society. (2025). *The risk assessment tools database*. https://criminaljustice.tooltrack.org/tool/16627

Bjørner, N., & Nachmanson, L. (2020). Navigating the universe of Z3 theory solvers. *Formal Methods: Foundations and Applications*, 8–24. https://doi.org/10.1007/978-3-030-63882-5_2

Bradley, A. R., & Manna, Z. (2007). *The calculus of computation*. Springer. https://doi.org/10.1007/978-3-540-74113-8

Brennan, T., & Dieterich, W. (2018). Correctional Offender Management Profiles for Alternative Sanctions (COMPAS). In *Handbook of Recidivism Risk/Needs Assessment Tools* (pp. 49–75). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119184256.ch3
Chapter: 3.

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, *36*(1), 21–40. https://doi.org/10.1177/0093854808326545

Bruninghaus, S., & Ashley, K. D. (2003). Predicting outcomes of case based legal arguments. *Proceedings of the Ninth International Conference on Artificial Intelligence and Law*, 233–242. https://doi.org/10.1145/1047788.1047838

Canavotto, I., & Horty, J. (2022). Piecemeal knowledge acquisition for computational normative reasoning. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 171–180. https://doi.org/10.1145/3514094.3534182

Canavotto, I., & Horty, J. (2023a). The importance of intermediate factors. In G. Sileno, J. Spanakis, & G. van Dijck (Eds.), *Legal Knowledge and Information Systems. JURIX 2023: The Thirty-sixth Annual Conference* (pp. 13–22). IOS Press. https://doi.org/10.3233/FAIA230941

Canavotto, I., & Horty, J. (2023b). Reasoning with hierarchies of open-textured predicates. *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, 52–61. https://doi.org/10.1145/3594536.3595148

Chuang, Y.-N., Wang, G., Chang, C.-Y., Tang, R., Zhong, S., Yang, F., Du, M., Cai, X., & Hu, X. (2024, June 26). *FaithLM: Towards Faithful Explanations for Large Language Models* (3). arXiv: 2402.04678 [cs]. https://doi.org/10.48550/arXiv.2402.04678

Čyras, K., Birch, D., Guo, Y., Toni, F., Dulay, R., Turvey, S., Greenberg, D., & Hapuarachchi, T. (2019). Explanations by arbitrated argumentative dispute. *Expert Systems with Applications*, *127*, 141–156. https://doi.org/10.1016/j.eswa.2019.03.012

Čyras, K., Satoh, K., & Toni, F. (2016). Explanation for case-based reasoning via abstract argumentation. In P. Baroni, T. F. Gordon, T. Scheffler, & M. Stede (Eds.), *Computational Models of Argument. Proceedings of COMMA 2016* (pp. 243–254). IOS Press. https://doi.org/10.3233/978-1-61499-686-6-243

Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, *16*(1), 64–93. https://doi.org/10.1093/jla/laae003

Davey, B. A., & Priestley, H. A. (2002). *Introduction to lattices and order* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511809088

de Moura, L., & Bjørner, N. (2008). Z3: An efficient SMT solver. *Tools and Algorithms for the Construction and Analysis of Systems*, 337–340. https://doi.org/10.1007/978-3-540-78800-3_24

de Moura, L., & Bjørner, N. (2009). Satisfiability modulo theories: An appetizer. *Formal Methods: Foundations and Applications*, 23–36. https://doi.org/10.1007/978-3-642-10452-7_3

Dieterich, W., Mendoza, C., & Brennan, T. (2016). *COMPAS risk scales: Demonstrating accuracy equity and predictive parity* (Research report). Northpointe Inc. Research Department.

Dobson, A. J. (2001, November 28). *An Introduction to Generalized Linear Models* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781420057683

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and *n*-person games. *Artificial Intelligence*, *77*(2), 321–357. https://doi.org/10.1016/0004-3702(94)00041-X

Engel, C., Linhardt, L., & Schubert, M. (2024). Code is law: How COMPAS affects the way the judiciary handles the risk of recidivism. *Artificial Intelligence and Law*. https://doi.org/10.1007/s10506-024-09389-8

Equivant. (2019, April 4). Practitioner's guide to COMPAS core. https://equivant-supervision.com/resources/white-papers-research-studies/

Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks." *The Administrative Office of the U.S. Courts*, *80*(2), 38–46.

Grabmair, M. (2017). Predicting trade secret case outcomes using argument schemes and learned quantitative value effect tradeoffs. *Proceedings of the Sixteenth International Conference on Articial Intelligence and Law*, 89–98. https://doi.org/10.1145/3086512.3086521

Hagberg, A., Swart, P. J., & Schult, D. A. (2008, January 1). *Exploring network structure, dynamics, and function using NetworkX* (LA-UR-08-05495; LA-UR-08-5495). Los Alamos National Laboratory (LANL), Los Alamos, NM (United States). Retrieved January 29, 2025, from https://www.osti.gov/biblio/960616

Hage, J., & Verheij, B. (1994). Reason-based logic: A logic for reasoning with rules and reasons. *Information & Communications Technology Law*, *3*(2–3), 171–209. https://doi.org/10.1080/13600834.1994.9965701

Horty, J. (2004). The result model of precedent. *Legal Theory*, *10*(1), 19–31. https://doi.org/10.1017/S1352325204000151

Horty, J. (2011). Rules and reasons in the theory of precedent. *Legal Theory*, *17*(1), 1–33. https://doi.org/10.1017/S1352325211000036

Horty, J. (2019). Reasoning with dimensions and magnitudes. *Artificial Intelligence and Law*, *27*(3), 309–345. https://doi.org/10.1007/s10506-019-09245-0

Jackson, E., & Mendoza, C. (2020). Setting the record straight: What the COMPAS core risk and need assessment is and is not. *Harvard Data Science Review*, *2*(1). https://doi.org/10.1162/99608f92.1b3dadaa

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. *Proceedings of the 34th International Conference on Machine Learning*, 1885–1894. https://proceedings.mlr.press/v70/koh17a.html

Koons, R. (2017). Defeasible reasoning. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archivES/FALL2017/entries/reasoning-defeasible

Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 166–172. https://doi.org/10.18653/v1/W19-3823

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). *How we analyzed the COMPAS recidivism algorithm*. ProPublica. Retrieved May 16, 2024, from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Liu, X., Lorini, E., Rotolo, A., & Sartor, G. (2022). Modelling and explaining legal case-based reasoners through classifiers. In E. Francesconi, G. Borges, & C. Sorge (Eds.), *Legal Knowledge and Information Systems. JURIX 2022: The Thirty-fifth Annual Conference* (pp. 83–92). IOS Press. https://doi.org/10.3233/FAIA220451

Manzano, M., & Aranda, V. (2022). Many-sorted logic. In *The Stanford Encyclopedia of Philosophy* (Winter 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2022/entries/logic-many-sorted/

Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, & Suresh Venkatasubramanian. (2021). It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, *1*. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Molnar, C. (2024). *Interpretable Machine Learning*. Independently published. https://christophm.github.io/interpretable-ml-book/

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. *The Journal of the Royal Statistical Society*, *135*(3), 370–384. https://doi.org/10.2307/2344614

New York State Division of Criminal Justice Services, Office of Probation and Correctional Alternatives. (2015). Practitioner Guidance for Probation and Community Corrections Agencies. https://apps.criminaljustice.ny.gov/opca/pdfs/2015-5-NYCOMPAS-Guidance-August-4-2015.pdf

Nisbet, R., Miner, G., & Yale, K. (2018, January 1). *Handbook of statistical analysis and data mining applications* (Second). Academic Press. https://doi.org/10.1016/C2012-0-06451-4

Northpointe. (2009). Measurement and treatment implications of COMPAS core scales. *Northpointe Institute for Public Management, Inc. A*.

Nugent, C., & Cunningham, P. (2005). A case-based explanation system for black-box systems. *Artificial Intelligence Review*, *24*(2), 163–178. https://doi.org/10.1007/s10462-005-4609-5

Odekerken, D. (2025). *Arguing with incomplete information: Formalisms, algorithms and applications in law enforcement* [Doctoral thesis]. Universiteit Utrecht. https://doi.org/10.33540/2627

Odekerken, D., & Bex, F. (2020). Towards transparent human-in-the-loop classification of fraudulent web shops. In Serena Villata, Jakub Harašta, & Petr Křemen (Eds.), *Legal Knowledge and Information Systems. JURIX 2020: The Thirty-third Annual Conference* (pp. 239–242). IOS Press. https://doi.org/10.3233/FAIA200873

Odekerken, D., Bex, F., & Prakken, H. (2023a). Precedent-based reasoning with incomplete cases. In G. Sileno, J. Spanakis, & G. van Dijck (Eds.), *Legal Knowledge and Information Systems. JURIX 2023: The Thirty-sixth Annual Conference* (pp. 33–42). IOS Press. https://doi.org/10.3233/FAIA230943

Odekerken, D., Bex, F., & Prakken, H. (2023b). Justification, stability and relevance for case-based reasoning with incomplete focus cases. *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, 177–186. https://doi.org/10.1145/3594536.3595136

OpenAI. (2024, March 4). *GPT-4 technical report*. arXiv: 2303.08774 [cs]. https://doi.org/10.48550/arXiv.2303.08774

Parcalabescu, L., & Frank, A. (2024, August). On Measuring Faithfulness or Self-consistency of Natural Language Explanations. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6048–6089). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.329

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Prakken, H. (2021). A formal analysis of some factor- and precedent-based accounts of precedential constraint. *Artificial Intelligence and Law*, *29*(4), 559–585. https://doi.org/10.1007/s10506-021-09284-6

Prakken, H., & Ratsma, R. (2022). A top-level model of case-based argumentation for explanation: Formalisation and experiments. *Argument & Computation*, *13*(2), 159–194. https://doi.org/10.3233/AAC-210009

Prakken, H., & Sartor, G. (1998). Modelling reasoning with precedents in a formal dialogue game. *Artificial Intelligence and Law*, *6*(2), 231–287. https://doi.org/10.1023/A:1008278309945

Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 97–101. https://doi.org/10.18653/v1/N16-3020

Rigoni, A. (2018). Representing dimensions within the reason model of precedent. *Artificial Intelligence and Law*, *26*(1), 1–22. https://doi.org/10.1007/s10506-017-9216-7

Rissland, E. L., & Ashley, K. D. (1987). A case-based system for trade secrets law. *Proceedings of the First International Conference on Artificial Intelligence and Law*, 60–66. https://doi.org/10.1145/41735.41743

Roth, B. (2003). *Case-based reasoning in the law: A formal theory of reasoning by case comparison* [Doctoral dissertation]. Universiteit Maastricht. https://doi.org/10.26481/dis.20031126ar

Roth, B., & Verheij, B. (2004). Cases and dialectical arguments – an approach to case-based reasoning. *On the Move to Meaningful Internet Systems 2004: OTM 2004 Workshops*, *3292*, 634–651. https://doi.org/10.1007/978-3-540-30470-8_75

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rudin, C., & Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.5a8a3a3d

Rudin, C., Wang, C., & Coker, B. (2020a). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, *2*(1). https://doi.org/10.1162/99608f92.6ed64b30

Rudin, C., Wang, C., & Coker, B. (2020b). Broader Issues Surrounding Model Transparency in Criminal Justice Risk Scoring. *Harvard Data Science Review*, *2*(1). https://doi.org/10.1162/99608f92.038c43fe

Schlimmer, J. (1981). *Mushroom dataset*. https://doi.org/10.24432/C5959T

Steging, C., Renooij, S., & Verheij, B. (2021). Discovering the rationale of decisions: Towards a method for aligning learning and reasoning. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 235–239. https://doi.org/10.1145/3462757.3466059

Steging, C., Renooij, S., Verheij, B., & Bench-Capon, T. (2023). Arguments, rules and cases in law: Resources for aligning learning and reasoning in structured domains. *Argument & Computation*, *14*(2), 235–243. https://doi.org/10.3233/AAC-220017

Strasser, C., & Antonelli, G. A. (2019). Non-monotonic logic. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2019/entries/logic-nonmonotonic

Tripathi, Y., Donakanti, R., Girhepuje, S., Kavathekar, I., Vedula, B. H., Krishnan, G. S., Goel, A., Goyal, S., Ravindran, B., & Kumaraguru, P. (2024). InSaAF: Incorporating Safety Through Accuracy and Fairness - Are LLMs Ready for the Indian Legal Domain? In Jaromir Savelka, Jakub Harasta, Tereza Novotna, & Jakub Misek (Eds.), *Legal Knowledge and Information Systems. JURIX 2024: The Thirty-seventh Annual Conference* (pp. 344–351, Vol. 395). IOS Press. https://doi.org/10.3233/FAIA241266

Turpin, M., Michael, J., Perez, E., & Bowman, S. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *Advances in Neural Information Processing Systems*, *36*, 74952–74965. Retrieved February 24, 2025, from https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html

Twining, W., & Miers, D. (2010). *How to do things with rules* (5th ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511844959

van Woerkom, W. (2025). Formal results on case-base consistency: A COMPAS case study. *Proceedings of the Twentieth International Conference on Artificial Intelligence and Law*. https://doi.org/10.1145/3769126.3769190

van Woerkom, W., Grossi, D., Prakken, H., & Verheij, B. (2022a, June). Landmarks in case-based reasoning: From theory to data. In Stefan Schlobach, María Pérez-Ortiz, & Myrthe Tielman (Eds.), *HHAI2022: Augmenting Human Intellect. Proceedings of the First International Conference on Hybrid Human-Artificial Intelligence* (pp. 212–224, Vol. 354). IOS Press. https://doi.org/10.3233/FAIA220200

van Woerkom, W., Grossi, D., Prakken, H., & Verheij, B. (2022b). Justification in case-based reasoning. *1st International Workshop on Argumentation for eXplainable AI*, *3209*. https://ceur-ws.org/Vol-3209/#5942

van Woerkom, W., Grossi, D., Prakken, H., & Verheij, B. (2023a). Hierarchical precedential constraint. *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, 333–342. https://doi.org/10.1145/3594536.3595154

van Woerkom, W., Grossi, D., Prakken, H., & Verheij, B. (2023b, December). Hierarchical *a fortiori* reasoning with dimensions. In G. Sileno, J. Spanakis, & G. van Dijck (Eds.), *Legal Knowledge and Information Systems. JURIX 2023: The Thirty-sixth Annual Conference* (pp. 43–52). IOS Press. https://doi.org/10.3233/FAIA230944

van Woerkom, W., Grossi, D., Prakken, H., & Verheij, B. (2024a). *A Fortiori* case-based reasoning: From theory to data. *Journal of Artificial Intelligence Research*, *81*, 401–441. https://doi.org/10.1613/jair.1.15178

van Woerkom, W., Grossi, D., Prakken, H., & Verheij, B. (2024b, December). A case-based-reasoning analysis of the COMPAS dataset. In Jaromir Savelka, Jakub Harasta, Tereza Novotna, & Jakub Misek (Eds.), *Legal Knowledge and Information Systems. JURIX 2024: The Thirty-seventh Annual Conference* (pp. 180–190, Vol. 395). IOS Press. https://doi.org/10.3233/FAIA241244

van Woerkom, W., Grossi, D., Prakken, H., & Verheij, B. (2025). Hierarchical models of precedential constraint [In press]. *Artificial Intelligence and Law*.

Verheij, B. (2017). Formalizing arguments, rules and cases. *Proceedings of the Sixteenth International Conference on Articial Intelligence and Law*, 199–208. https://doi.org/10.1145/3086512.3086533

Verheij, B. (2020). Artificial intelligence as law. *Artificial Intelligence and Law*, *28*(2), 181–206. https://doi.org/10.1007/s10506-020-09266-0

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2). https://doi.org/10.2139/ssrn.3063289

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

Williams, J., & Kolter, J. Z. (2021). A Bayesian model of cash bail decisions. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 827–837. https://doi.org/10.1145/3442188.3445908

Yukhnenko, D., Blackwood, N., & Fazel, S. (2020). Risk factors for recidivism in individuals receiving community sentences: A systematic review and meta-analysis. *CNS spectrums*, *25*(2), 252–263. https://doi.org/10.1017/S1092852919001056

Nederlandse Samenvatting

Casusgebaseerd redeneren (CR) houdt in dat een probleem wordt vergeleken met eerdere gevallen om besluitvorming te ondersteunen en conclusies te trekken. Dit type redenering is fundamenteel voor *common law*-rechtsstelsels, waar rechtbanken verplicht zijn eerdere uitspraken te volgen aan de hand van het *stare decisis*-principe. In deze systemen vergelijken advocaten huidige geschillen met eerdere zaken om argumenten op te bouwen, en gebruiken rechters eerdere zaken om hun conclusies te rechtvaardigen en uit te leggen.

Er zijn formele modellen ontwikkeld om juridisch CR te beschrijven. Deze berusten vaak op het representeren van zaken in termen van *factoren*: verzamelingen van feiten die vaak voorkomen in juridische zaken en die de positie van de ene of de andere partij in het geschil versterken of verzwakken. Een veelgebruikt voorbeeld van factoren heeft betrekking op het domein van bedrijfsgeheimen; het is moeilijk precies te definiëren wat een bedrijfsgeheim is, en daarom worden factoren gebruikt om aspecten van de betreffende situatie te beschrijven die de zaak sterker of zwakker maken voor de partij die beweert het bedrijfsgeheim te bezitten. Enkele voorbeelden van zulke factoren zijn: de mate waarin de informatie bekend is bij buitenstaanders; de waarde van de informatie voor de eigenaar en diens concurrenten; en de hoeveelheid moeite of geld die de eigenaar heeft besteed aan het ontwikkelen van de informatie.

Een bekend voorbeeld van een formeel model van CR is het zogenoemde *resultaatmodel* (RM) van precedentgebondenheid. Het formuleert een beknopt *a fortiori*-principe van beslissingsgebondenheid voor zaken die worden weergegeven in termen van de factoren van het domein, wat als volgt kan worden samengevat: wanneer een eerdere zaak is beslist in het voordeel van een bepaalde partij, moet elke nieuwe zaak waarin de factoren minstens even sterk in het voordeel van die partij spreken en niet sterker tegen haar pleiten, eveneens in haar voordeel worden beslist; een symmetrische regel geldt voor de andere partij. Sinds de introductie ervan zijn er vele varianten en uitbreidingen van het RM ontwikkeld.

Het RM en de varianten ervan staan centraal in dit proefschrift, met name wat betreft hun toepasbaarheid op kunstmatige intelligentie (AI). Juridische besluitvormers generaliseren eerdere beslissingen om tot een oordeel te komen in een nieuwe zaak, en dit wordt vergeleken met de manier waarop datagestuurde AI-systemen trainingsdata generaliseren om output te genereren voor nieuwe input. Deze analogie maakt het mogelijk het RM toe te passen in de context van datagestuurde besluitvorming, en dit heeft in de afgelopen jaren geleid tot de ontwikkeling van diverse toepassingen — bijvoorbeeld om AI-output te verklaren of te rechtvaardigen, of om frauduleuze online activiteiten te detecteren.

Kort samengevat zijn de doelstellingen van dit proefschrift tweevoudig. Ten eerste om het RM uit te breiden tot een algemene theorie van *a fortiori* CR, met als doel deze theorie toe te passen op het gebied van AI en recht; en ten tweede om de resulterende theorie te toetsen door haar toe te passen op de analyse van datagestuurde AI-beslissingen. Deze doelstellingen worden respectievelijk behandeld in Deel I en Deel II van dit proefschrift.

Deel I

Al geruime tijd wordt erkend dat een eenvoudige, op factoren gebaseerde representatie van zaken niet het volledige beeld weergeeft. Zo kunnen factoren bijvoorbeeld een hiërarchische structuur hebben. Neem de eerder genoemde factor die betrekking heeft op de mate waarin informatie bekend is bij buitenstaanders; deze kan worden uitgesplitst in subfactoren die beschrijven of de informatie uniek was en of werknemers een geheimhoudingsverklaring hebben moeten ondertekenen. Daarnaast kunnen factoren meerwaardige in plaats van binaire waarden aannemen; zulke factoren worden in de literatuur vaak *dimensies* genoemd. Ten slotte is het in de praktijk vaak onmogelijk om voor alle factoren te bepalen of ze van toepassing zijn, omdat sommige feiten onbekend of irrelevant zijn. Dit brengt ons bij de eerste centrale onderzoeksvraag van dit proefschrift: Kunnen we het RM van precedentgebondenheid uitbreiden tot een algemene theorie van *a fortiori* casusgebaseerd redeneren, die partiële, dimensionele en hiërarchische informatie omvat?

We beantwoorden deze vraag bevestigend door een reeks uitbreidingen van het RM te ontwikkelen. Eerst passen we het RM en diens dimensionele variant (DRM) aan om te werken met onvolledige informatie. Vervolgens introduceren we een hiërarchische versie van het RM (het HRM), waarin gebondenheid wordt gedefinieerd via recursie over een factorhiërarchie, wat meervoudige precedentredenering mogelijk maakt. Ten slotte presenteren we een dimensioneel-hiërarchische versie van het RM (het DHRM), waarin uitkomsten niet-binair zijn en gebondenheid wordt uitgedrukt met onder- en bovengrenzen. De modellen bouwen voort op elkaar, waarbij het DHRM alle voorgaande omvat.

Deel II

De in Deel I ontwikkelde uitbreidingen zijn gemotiveerd door toepassingen van het RM binnen de domeinen van AI en recht. Deze toepassingen vloeien voort uit de analogie tussen juridische besluitvorming en datagestuurde besluitvorming door AI. Dit leidt vanzelf tot de tweede centrale onderzoeksvraag die in dit proefschrift wordt behandeld: Hoe kan de in Deel I ontwikkelde theorie worden toegepast op het gebied van AI?

Om deze vraag te beantwoorden benutten we een verbinding tussen onze theorie en formele logica om een implementatie ervan te schrijven op basis van de *"Satisfiability Modulo Theories" solver* Z3. We tonen aan hoe deze implementatie gebruikt kan worden om gebondenheid uit te rekenen, maar ook voor verschillende andere doeleinden zoals het toetsen van de consistentie en volledigheid van een casusverzameling. We illustreren de werking van de implementatie aan de hand van experimenten met verschillende datasets.

Een van de voordelen van de theorie is dat zij een notie van consistentie voor datagedreven beslissingen definieert. Door de consistentie van een verzameling datagedreven beslissingen te meten, ontstaat een berekenbare maatstaf voor de interne beslisconsistentie van een AI-systeem. Ter illustratie passen we deze concepten toe op data over het COMPAS-systeem: een datagedreven AI-systeem dat in de Verenigde Staten onder meer wordt ingezet om de kans op recidive bij strafrechtelijke verdachten te voorspellen. Het gebruik van dit systeem is het onderwerp geworden van een voortdurende discussie in de literatuur over het verantwoord gebruik van AI. Onze analyse werpt nieuw licht op deze discussie en bevestigt zo, onzes inziens, het nut van de theorie.

List of SIKS-Dissertations

2016

01 Syed Saiden Abbas (RUN), *Recognition of Shapes by Humans and Machines*.
02 Michiel Christiaan Meulendijk (UU), *Optimizing medication reviews through decision support: prescribing a better pill to swallow*.
03 Maya Sappelli (RUN), *Knowledge Work in Context: User Centered Knowledge Worker Support*.
04 Laurens Rietveld (VUA), *Publishing and Consuming Linked Data*.
05 Evgeny Sherkhonov (UvA), *Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers*.
06 Michel Wilson (TUD), *Robust scheduling in an uncertain environment*.
07 Jeroen de Man (VUA), *Measuring and modeling negative emotions for virtual training*.
08 Matje van de Camp (TiU), *A Link to the Past: Constructing Historical Social Networks from Unstructured Data*.
09 Archana Nottamkandath (VUA), *Trusting Crowdsourced Information on Cultural Artefacts*.
10 George Karafotias (VUA), *Parameter Control for Evolutionary Algorithms*.
11 Anne Schuth (UvA), *Search Engines that Learn from Their Users*.
12 Max Knobbout (UU), *Logics for Modelling and Verifying Normative Multi-Agent Systems*.
13 Nana Baah Gyan (VUA), *The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach*.
14 Ravi Khadka (UU), *Revisiting Legacy Software System Modernization*.
15 Steffen Michels (RUN), *Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments*.
16 Guangliang Li (UvA), *Socially Intelligent Autonomous Agents that Learn from Human Reward*.
17 Berend Weel (VUA), *Towards Embodied Evolution of Robot Organisms*.
18 Albert Meroño Peñuela (VUA), *Refining Statistical Data on the Web*.
19 Julia Efremova (TU/e), *Mining Social Structures from Genealogical Data*.
20 Daan Odijk (UvA), *Context & Semantics in News & Web Search*.
21 Alejandro Moreno Célleri (UT), *From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground*.
22 Grace Lewis (VUA), *Software Architecture Strategies for Cyber-Foraging Systems*.
23 Fei Cai (UvA), *Query Auto Completion in Information Retrieval*.
24 Brend Wanders (UT), *Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach*.
25 Julia Kiseleva (TU/e), *Using Contextual Information to Understand Searching and Browsing Behavior*.
26 Dilhan Thilakarathne (VUA), *In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains*.
27 Wen Li (TUD), *Understanding Geo-spatial Information on Social Media*.
28 Mingxin Zhang (TUD), *Large-scale Agent-based Social Simulation - A study on epidemic prediction and control*.
29 Nicolas Höning (TUD), *Peak reduction in decentralised electricity systems - Markets and prices for flexible planning*.
30 Ruud Mattheij (TiU), *The Eyes Have It*.
31 Mohammad Khelghati (UT), *Deep web content monitoring*.
32 Eelco Vriezekolk (UT), *Assessing Telecommunication Service Availability Risks for Crisis Organisations*.
33 Peter Bloem (UvA), *Single Sample Statistics, exercises in learning from just one example*.
34 Dennis Schunselaar (TU/e), *Configurable Process Trees: Elicitation, Analysis, and Enactment*.
35 Zhaochun Ren (UvA), *Monitoring Social Media: Summarization, Classification and Recommendation*.
36 Daphne Karreman (UT), *Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies*.
37 Giovanni Sileno (UvA), *Aligning Law and Action - a conceptual and computational inquiry*.
38 Andrea Minuto (UT), *Materials that Matter - Smart Materials meet Art & Interaction Design*.
39 Merijn Bruijnes (UT), *Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect*.
40 Christian Detweiler (TUD), *Accounting for Values in Design*.
41 Thomas King (TUD), *Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance*.
42 Spyros Martzoukos (UvA), *Combinatorial and Compositional Aspects of Bilingual Aligned Corpora*.
43 Saskia Koldijk (RUN), *Context-Aware Support for Stress Self-Management: From Theory to Practice*.
44 Thibault Sellam (UvA), *Automatic Assistants for Database Exploration*.
45 Bram van de Laar (UT), *Experiencing Brain-Computer Interface Control*.
46 Jorge Gallego Perez (UT), *Robots to Make you Happy*.
47 Christina Weber (UL), *Real-time foresight - Preparedness for dynamic innovation networks*.
48 Tanja Buttler (TUD), *Collecting Lessons Learned*.
49 Gleb Polevoy (TUD), *Participation and Interaction in Projects. A Game-Theoretic Analysis*.

50 Yan Wang (TiU), *The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains*.

2017

01 Jan-Jaap Oerlemans (UL), *Investigating Cybercrime*.
02 Sjoerd Timmer (UU), *Designing and Understanding Forensic Bayesian Networks using Argumentation*.
03 Daniël Harold Telgen (UU), *Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines*.
04 Mrunal Gawade (CWI), *Multi-core Parallelism in a Column-store*.
05 Mahdieh Shadi (UvA), *Collaboration Behavior*.
06 Damir Vandic (EUR), *Intelligent Information Systems for Web Product Search*.
07 Roel Bertens (UU), *Insight in Information: from Abstract to Anomaly*.
08 Rob Konijn (VUA), *Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery*.
09 Dong Nguyen (UT), *Text as Social and Cultural Data: A Computational Perspective on Variation in Text*.
10 Robby van Delden (UT), *(Steering) Interactive Play Behavior*.
11 Florian Kunneman (RUN), *Modelling patterns of time and emotion in Twitter #anticipointment*.
12 Sander Leemans (TU/e), *Robust Process Mining with Guarantees*.
13 Gijs Huisman (UT), *Social Touch Technology - Extending the reach of social touch through haptic technology*.
14 Shoshannah Tekofsky (TiU), *You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior*.
15 Peter Berck (RUN), *Memory-Based Text Correction*.
16 Aleksandr Chuklin (UvA), *Understanding and Modeling Users of Modern Search Engines*.
17 Daniel Dimov (UL), *Crowdsourced Online Dispute Resolution*.
18 Ridho Reinanda (UvA), *Entity Associations for Search*.
19 Jeroen Vuurens (UT), *Proximity of Terms, Texts and Semantic Vectors in Information Retrieval*.
20 Mohammadbashir Sedighi (TUD), *Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility*.
21 Jeroen Linssen (UT), *Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)*.
22 Sara Magliacane (VUA), *Logics for causal inference under uncertainty*.
23 David Graus (UvA), *Entities of Interest — Discovery in Digital Traces*.
24 Chang Wang (TUD), *Use of Affordances for Efficient Robot Learning*.
25 Veruska Zamborlini (VUA), *Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search*.
26 Merel Jung (UT), *Socially intelligent robots that understand and respond to human touch*.
27 Michiel Joosse (UT), *Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors*.
28 John Klein (VUA), *Architecture Practices for Complex Contexts*.
29 Adel Alhuraibi (TiU), *From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"*.
30 Wilma Latuny (TiU), *The Power of Facial Expressions*.
31 Ben Ruijl (UL), *Advances in computational methods for QFT calculations*.
32 Thaer Samar (RUN), *Access to and Retrievability of Content in Web Archives*.
33 Brigit van Loggem (OU), *Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity*.
34 Maren Scheffel (OU), *The Evaluation Framework for Learning Analytics*.
35 Martine de Vos (VUA), *Interpreting natural science spreadsheets*.
36 Yuanhao Guo (UL), *Shape Analysis for Phenotype Characterisation from High-throughput Imaging*.
37 Alejandro Montes Garcia (TU/e), *WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy*.
38 Alex Kayal (TUD), *Normative Social Applications*.
39 Sara Ahmadi (RUN), *Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR*.
40 Altaf Hussain Abro (VUA), *Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems*.
41 Adnan Manzoor (VUA), *Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle*.
42 Elena Sokolova (RUN), *Causal discovery from mixed and missing data with applications on ADHD datasets*.
43 Maaike de Boer (RUN), *Semantic Mapping in Video Retrieval*.
44 Garm Lucassen (UU), *Understanding User Stories - Computational Linguistics in Agile Requirements Engineering*.
47 Jie Yang (TUD), *Crowd Knowledge Creation Acceleration*.
48 Angel Suarez (OU), *Collaborative inquiry-based learning*.

2018

01 Han van der Aa (VUA), *Comparing and Aligning Process Representations*.

02 Felix Mannhardt (TU/e), *Multi-perspective Process Mining*.
03 Steven Bosems (UT), *Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction*.
04 Jordan Janeiro (TUD), *Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks*.
05 Hugo Huurdeman (UvA), *Supporting the Complex Dynamics of the Information Seeking Process*.
06 Dan Ionita (UT), *Model-Driven Information Security Risk Assessment of Socio-Technical Systems*.
07 Jieting Luo (UU), *A formal account of opportunism in multi-agent systems*.
08 Rick Smetsers (RUN), *Advances in Model Learning for Software Systems*.
10 Julienka Mollee (VUA), *Moving forward: supporting physical activity behavior change through intelligent technology*.
11 Mahdi Sargolzaei (UvA), *Enabling Framework for Service-oriented Collaborative Networks*.
12 Xixi Lu (TU/e), *Using behavioral context in process mining*.
13 Seyed Amin Tabatabaei (VUA), *Computing a Sustainable Future*.
14 Bart Joosten (TiU), *Detecting Social Signals with Spatiotemporal Gabor Filters*.
15 Naser Davarzani (UM), *Biomarker discovery in heart failure*.
16 Jaebok Kim (UT), *Automatic recognition of engagement and emotion in a group of children*.
17 Jianpeng Zhang (TU/e), *On Graph Sample Clustering*.
18 Henriette Nakad (UL), *De Notaris en Private Rechtspraak*.
19 Minh Duc Pham (VUA), *Emergent relational schemas for RDF*.
20 Manxia Liu (RUN), *Time and Bayesian Networks*.
21 Aad Slootmaker (OU), *EMERGO: a generic platform for authoring and playing scenario-based serious games*.
22 Eric Fernandes de Mello Araújo (VUA), *Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks*.
23 Kim Schouten (EUR), *Semantics-driven Aspect-Based Sentiment Analysis*.
24 Jered Vroon (UT), *Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots*.
25 Riste Gligorov (VUA), *Serious Games in Audio-Visual Collections*.
27 Maikel Leemans (TU/e), *Hierarchical Process Mining for Scalable Software Analysis*.
28 Christian Willemse (UT), *Social Touch Technologies: How they feel and how they make you feel*.
29 Yu Gu (TiU), *Emotion Recognition from Mandarin Speech*.
30 Wouter Beek (VUA), *The "K" in "semantic web" stands for "knowledge": scaling semantics to the web*.

2019

02 Emmanuelle Beauxis Aussalet (CWI, UU), *Statistics and Visualizations for Assessing Class Size Uncertainty*.
03 Eduardo Gonzalez Lopez de Murillas (TU/e), *Process Mining on Databases: Extracting Event Data from Real Life Data Sources*.
04 Ridho Rahmadi (RUN), *Finding stable causal structures from clinical data*.
05 Sebastiaan van Zelst (TU/e), *Process Mining with Streaming Data*.
06 Chris Dijkshoorn (VUA), *Nichesourcing for Improving Access to Linked Cultural Heritage Datasets*.
07 Soude Fazeli (TUD), *Recommender Systems in Social Learning Platforms*.
08 Frits de Nijs (TUD), *Resource-constrained Multi-agent Markov Decision Processes*.
09 Fahimeh Alizadeh Moghaddam (UvA), *Self-adaptation for energy efficiency in software systems*.
10 Qing Chuan Ye (EUR), *Multi-objective Optimization Methods for Allocation and Prediction*.
11 Yue Zhao (TUD), *Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs*.
12 Jacqueline Heinerman (VUA), *Better Together*.
13 Guanliang Chen (TUD), *MOOC Analytics: Learner Modeling and Content Generation*.
14 Daniel Davis (TUD), *Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses*.
15 Erwin Walraven (TUD), *Planning under Uncertainty in Constrained and Partially Observable Environments*.
16 Guangming Li (TU/e), *Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models*.
18 Gerard Wagenaar (UU), *Artefacts in Agile Team Communication*.
19 Vincent Koeman (TUD), *Tools for Developing Cognitive Agents*.
20 Chide Groenouwe (UU), *Fostering technically augmented human collective intelligence*.
21 Cong Liu (TU/e), *Software Data Analytics: Architectural Model Discovery and Design Pattern Detection*.
23 Qin Liu (TUD), *Intelligent Control Systems: Learning, Interpreting, Verification*.
24 Anca Dumitrache (VUA), *Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing*.
25 Emiel van Miltenburg (VUA), *Pragmatic factors in (automatic) image description*.
26 Prince Singh (UT), *An Integration Platform for Synchromodal Transport*.
27 Alessandra Antonaci (OU), *The Gamification Design Process applied to (Massive) Open Online Courses*.
28 Esther Kuindersma (UL), *Cleared for take-off: Game-based learning to prepare airline pilots for critical situations*.
29 Daniel Formolo (VUA), *Using virtual agents for simulation and training of social skills in safety-critical circumstances*.
30 Vahid Yazdanpanah (UT), *Multiagent Industrial Symbiosis Systems*.
31 Milan Jelisavcic (VUA), *Alive and Kicking: Baby Steps in Robotics*.
32 Chiara Sironi (UM), *Monte-Carlo Tree Search for Artificial General Intelligence in Games*.
33 Anil Yaman (TU/e), *Evolution of Biologically Inspired Learning in Artificial Neural Networks*.
34 Negar Ahmadi (TU/e), *EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES*.
35 Lisa Facey-Shaw (OU), *Gamification with digital badges in learning programming*.
36 Kevin Ackermans (OU), *Designing Video-Enhanced Rubrics to Master Complex Skills*.

37 Jian Fang (TUD), *Database Acceleration on FPGAs*.
38 Akos Kadar (OU), *Learning visually grounded and multilingual representations.*

2020

01 Armon Toubman (UL), *Calculated Moves: Generating Air Combat Behaviour*.
02 Marcos de Paula Bueno (UL), *Unraveling Temporal Processes using Probabilistic Graphical Models*.
03 Mostafa Deghani (UvA), *Learning with Imperfect Supervision for Language Understanding*.
04 Maarten van Gompel (RUN), *Context as Linguistic Bridges*.
05 Yulong Pei (TU/e), *On local and global structure mining*.
06 Preethu Rose Anish (UT), *Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support*.
07 Wim van der Vegt (OU), *Towards a software architecture for reusable game components*.
09 Myriam Traub (UU), *Measuring Tool Bias and Improving Data Quality for Digital Humanities Research*.
10 Alifah Syamsiyah (TU/e), *In-database Preprocessing for Process Mining*.
11 Sepideh Mesbah (TUD), *Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models*.
12 Ward van Breda (VUA), *Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment*.
13 Marco Virgolin (CWI), *Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming*.
14 Mark Raasveldt (CWI/UL), *Integrating Analytics with Relational Databases*.
15 Konstantinos Georgiadis (OU), *Smart CAT: Machine Learning for Configurable Assessments in Serious Games*.
16 Ilona Wilmont (RUN), *Cognitive Aspects of Conceptual Modelling*.
17 Daniele Di Mitri (OU), *The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences*.
18 Georgios Methenitis (TUD), *Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems*.
19 Guido van Capelleveen (UT), *Industrial Symbiosis Recommender Systems*.
20 Albert Hankel (VUA), *Embedding Green ICT Maturity in Organisations*.
21 Karine da Silva Miras de Araujo (VUA), *Where is the robot?: Life as it could be*.
22 Maryam Masoud Khamis (RUN), *Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar*.
23 Rianne Conijn (UT), *The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging*.
24 Lenin da Nóbrega Medeiros (VUA/RUN), *How are you feeling, human? Towards emotionally supportive chatbots*.
25 Xin Du (TU/e), *The Uncertainty in Exceptional Model Mining*.
26 Krzysztof Leszek Sadowski (UU), *GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization*.
27 Ekaterina Muravyeva (TUD), *Personal data and informed consent in an educational context*.
28 Bibeg Limbu (TUD), *Multimodal interaction for deliberate practice: Training complex skills with augmented reality*.
29 Ioan Gabriel Bucur (RUN), *Being Bayesian about Causal Inference*.
30 Bob Zadok Blok (UL), *Creatief, Creatiever, Creatiefst*.
31 Gongjin Lan (VUA), *Learning better – From Baby to Better*.
32 Jason Rhuggenaath (TU/e), *Revenue management in online markets: pricing and online advertising*.
33 Rick Gilsing (TU/e), *Supporting service-dominant business model evaluation in the context of business model innovation*.
34 Anna Bon (UM), *Intervention or Collaboration? Redesigning Information and Communication Technologies for Development*.
35 Siamak Farshidi (UU), *Multi-Criteria Decision-Making in Software Production*.

2021

02 Rijk Mercuur (TUD), *Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models*.
03 Seyyed Hadi Hashemi (UvA), *Modeling Users Interacting with Smart Devices*.
04 Ioana Jivet (OU), *The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning*.
05 Davide Dell'Anna (UU), *Data-Driven Supervision of Autonomous Systems*.
06 Daniel Davison (UT), *"Hey robot, what do you think?" How children learn with a social robot*.
07 Armel Lefebvre (UU), *Research data management for open science*.
08 Nardie Fanchamps (OU), *The Influence of Sense-Reason-Act Programming on Computational Thinking*.
09 Cristina Zaga (UT), *The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play*.
10 Quinten Meertens (UvA), *Misclassification Bias in Statistical Learning*.
11 Anne van Rossum (UL), *Nonparametric Bayesian Methods in Robotic Vision*.
12 Lei Pi (UL), *External Knowledge Absorption in Chinese SMEs*.
13 Bob R. Schadenberg (UT), *Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning*.
14 Negin Samaeemofrad (UL), *Business Incubators: The Impact of Their Support*.

15 Onat Ege Adali (TU/e), *Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm*.
16 Esam A. H. Ghaleb (UM), *Bimodal emotion recognition from audio-visual cues*.
17 Dario Dotti (UM), *Human Behavior Understanding from motion and bodily cues using deep neural networks*.
18 Remi Wieten (UU), *Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks*.
19 Roberto Verdecchia (VUA), *Architectural Technical Debt: Identification and Management*.
20 Masoud Mansoury (TU/e), *Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems*.
21 Pedro Thiago Timbó Holanda (CWI), *Progressive Indexes*.
22 Sihang Qiu (TUD), *Conversational Crowdsourcing*.
23 Hugo Manuel Proença (UL), *Robust rules for prediction and description*.
24 Kaijie Zhu (TU/e), *On Efficient Temporal Subgraph Query Processing*.
25 Eoin Martino Grua (VUA), *The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications*.
26 Benno Kruit (CWI/VUA), *Reading the Grid: Extending Knowledge Bases from Human-readable Tables*.
27 Jelte van Waterschoot (UT), *Personalized and Personal Conversations: Designing Agents Who Want to Connect With You*.
28 Christoph Selig (UL), *Understanding the Heterogeneity of Corporate Entrepreneurship Programs*.

2022

01 Judith van Stegeren (UT), *Flavor text generation for role-playing video games*.
02 Paulo da Costa (TU/e), *Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey*.
03 Ali el Hassouni (VUA), *A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare*.
04 Ünal Aksu (UU), *A Cross-Organizational Process Mining Framework*.
05 Shiwei Liu (TU/e), *Sparse Neural Network Training with In-Time Over-Parameterization*.
06 Reza Refaei Afshar (TU/e), *Machine Learning for Ad Publishers in Real Time Bidding*.
07 Sambit Praharaj (OU), *Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics*.
08 Maikel L. van Eck (TU/e), *Process Mining for Smart Product Design*.
09 Oana Andreea Inel (VUA), *Understanding Events: A Diversity-driven Human-Machine Approach*.
10 Felipe Moraes Gomes (TUD), *Examining the Effectiveness of Collaborative Search Engines*.
11 Mirjam de Haas (UT), *Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring*.
12 Guanyi Chen (UU), *Computational Generation of Chinese Noun Phrases*.
13 Xander Wilcke (VUA), *Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge*.
14 Michiel Overeem (UU), *Evolution of Low-Code Platforms*.
15 Jelmer Jan Koorn (UU), *Work in Process: Unearthing Meaning using Process Mining*.
16 Pieter Gijsbers (TU/e), *Systems for AutoML Research*.
17 Laura van der Lubbe (VUA), *Empowering vulnerable people with serious games and gamification*.
18 Paris Mavromoustakos Blom (TiU), *Player Affect Modelling and Video Game Personalisation*.
19 Bilge Yigit Ozkan (UU), *Cybersecurity Maturity Assessment and Standardisation*.
20 Fakhra Jabeen (VUA), *Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media*.
21 Seethu Mariyam Christopher (UM), *Intelligent Toys for Physical and Cognitive Assessments*.
22 Alexandra Sierra Rativa (TiU), *Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations*.
23 Ilir Kola (TUD), *Enabling Social Situation Awareness in Support Agents*.
24 Samaneh Heidari (UU), *Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values*.
25 Anna L.D. Latour (UL), *Optimal decision-making under constraints and uncertainty*.
26 Anne Dirkson (UL), *Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences*.
27 Christos Athanasiadis (UM), *Emotion-aware cross-modal domain adaptation in video sequences*.
28 Onuralp Ulusoy (UU), *Privacy in Collaborative Systems*.
29 Jan Kolkmeier (UT), *From Head Transform to Mind Transplant: Social Interactions in Mixed Reality*.
30 Dean De Leo (CWI), *Analysis of Dynamic Graphs on Sparse Arrays*.
31 Konstantinos Traganos (TU/e), *Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management*.
32 Cezara Pastrav (UU), *Social simulation for socio-ecological systems*.
33 Brinn Hekkelman (CWI/TUD), *Fair Mechanisms for Smart Grid Congestion Management*.
34 Nimat Ullah (VUA), *Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change*.
35 Mike E.U. Ligthart (VUA), *Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction*.

2023

01 Bojan Simoski (VUA), *Untangling the Puzzle of Digital Health Interventions*.
02 Mariana Rachel Dias da Silva (TiU), *Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts*.
03 Shabnam Najafian (TUD), *User Modeling for Privacy-preserving Explanations in Group Recommendations*.
04 Gineke Wiggers (UL), *The Relevance of Impact: bibliometric-enhanced legal information retrieval*.
05 Anton Bouter (CWI), *Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications*.
06 António Pereira Barata (UL), *Reliable and Fair Machine Learning for Risk Assessment*.
07 Tianjin Huang (TU/e), *The Roles of Adversarial Examples on Trustworthiness of Deep Learning*.
08 Lu Yin (TU/e), *Knowledge Elicitation using Psychometric Learning*.
09 Xu Wang (VUA), *Scientific Dataset Recommendation with Semantic Techniques*.
10 Dennis J.N.J. Soemers (UM), *Learning State-Action Features for General Game Playing*.
11 Fawad Taj (VUA), *Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications*.
12 Tessel Bogaard (VUA), *Using Metadata to Understand Search Behavior in Digital Libraries*.
13 Injy Sarhan (UU), *Open Information Extraction for Knowledge Representation*.
14 Selma Čaušević (TUD), *Energy resilience through self-organization*.
15 Alvaro Henrique Chaim Correia (TU/e), *Insights on Learning Tractable Probabilistic Graphical Models*.
16 Peter Blomsma (TiU), *Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters*.
17 Meike Nauta (UT), *Explainable AI and Interpretable Computer Vision – From Oversight to Insight*.
18 Gustavo Penha (TUD), *Designing and Diagnosing Models for Conversational Search and Recommendation*.
19 George Aalbers (TiU), *Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals*.
20 Arkadiy Dushatskiy (TUD), *Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning*.
21 Gerrit Jan de Bruin (UL), *Network Analysis Methods for Smart Inspection in the Transport Domain*.
22 Alireza Shojaifar (UU), *Volitional Cybersecurity*.
23 Theo Theunissen (UU), *Documentation in Continuous Software Development*.
24 Agathe Balayn (TUD), *Practices Towards Hazardous Failure Diagnosis in Machine Learning*.
25 Jurian Baas (UU), *Entity Resolution on Historical Knowledge Graphs*.
26 Loek Tonnaer (TU/e), *Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour*.
27 Ghada Sokar (TU/e), *Learning Continually Under Changing Data Distributions*.
28 Floris den Hengst (VUA), *Learning to Behave: Reinforcement Learning in Human Contexts*.
29 Tim Draws (TUD), *Understanding Viewpoint Biases in Web Search Results*.

2024

01 Daphne Miedema (TU/e), *On Learning SQL: Disentangling concepts in data systems education*.
02 Emile van Krieken (VUA), *Optimisation in Neurosymbolic Learning Systems*.
03 Feri Wijayanto (RUN), *Automated Model Selection for Rasch and Mediation Analysis*.
04 Mike Huisman (UL), *Understanding Deep Meta-Learning*.
05 Yiyong Gou (UM), *Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair*.
06 Azqa Nadeem (TUD), *Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence*.
07 Parisa Shayan (TiU), *Modeling User Behavior in Learning Management Systems*.
08 Xin Zhou (UvA), *From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation*.
09 Giso Dal (UT), *Probabilistic Inference Using Partitioned Bayesian Networks*.
10 Cristina-Iulia Bucur (VUA), *Linkflows: Towards Genuine Semantic Publishing in Science*.
12 Peide Zhu (TUD), *Towards Robust Automatic Question Generation For Learning*.
13 Enrico Liscio (TUD), *Context-Specific Value Inference via Hybrid Intelligence*.
14 Larissa Capobianco Shimomura (TU/e), *On Graph Generating Dependencies and their Applications in Data Profiling*.
15 Ting Liu (VUA), *A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health*.
16 Arthur Barbosa Câmara (TUD), *Designing Search-as-Learning Systems*.
17 Razieh Alidoosti (VUA), *Ethics-aware Software Architecture Design*.
18 Laurens Stoop (UU), *Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations*.
19 Azadeh Mozafari Mehr (TU/e), *Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior*.
20 Ritsart Anne Plantenga (UL), *Omgang met Regels*.
21 Federica Vinella (UU), *Crowdsourcing User-Centered Teams*.
22 Zeynep Ozturk Yurt (TU/e), *Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts*.
23 Jie Luo (VUA), *Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution*.

24 Nirmal Roy (TUD), *Exploring the effects of interactive interfaces on user search behaviour*.
25 Alisa Rieger (TUD), *Striving for Responsible Opinion Formation in Web Search on Debated Topics*.
26 Tim Gubner (CWI), *Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework*.
27 Lincen Yang (UL), *Information-theoretic Partition-based Models for Interpretable Machine Learning*.
28 Leon Helwerda (UL), *Grip on Software: Understanding development progress of Scrum sprints and backlogs*.
29 David Wilson Romero Guzman (VUA), *The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases*.
30 Vijanti Ramautar (UU), *Model-Driven Sustainability Accounting*.
31 Ziyu Li (TUD), *On the Utility of Metadata to Optimize Machine Learning Workflows*.
32 Vinicius Stein Dani (UU), *The Alpha and Omega of Process Mining*.
33 Siddharth Mehrotra (TUD), *Designing for Appropriate Trust in Human-AI interaction*.
34 Robert Deckers (VUA), *From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method*.
35 Sicui Zhang (TU/e), *Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach*.
36 Thomas Mulder (TU/e), *Optimization of Recursive Queries on Graphs*.
37 James Graham Nevin (UvA), *The Ramifications of Data Handling for Computational Models*.
38 Christos Koutras (TUD), *Tabular Schema Matching for Modern Settings*.
39 Paola Lara Machado (TU/e), *The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance*.
40 Montijn van de Ven (TU/e), *Guiding the Definition of Key Performance Indicators for Business Models*.
41 Georgios Siachamis (TUD), *Adaptivity for Streaming Dataflow Engines*.
42 Emmeke Veltmeijer (VUA), *Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis*.
43 Cedric Waterschoot (KNAW Meertens Instituut), *The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms*.
44 Marcel Schmitz (OU), *Towards learning analytics-supported learning design*.
45 Sara Salimzadeh (TUD), *Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making*.
46 Georgios Stathis (Leiden University), *Preventing Disputes: Preventive Logic, Law & Technology*.
47 Daniel Daza (VUA), *Exploiting Subgraphs and Attributes for Representation Learning on Knowledge Graphs*.
48 Ioannis Petros Samiotis (TUD), *Crowd-Assisted Annotation of Classical Music Compositions*.

2025

01 Max van Haastrecht (UL), *Transdisciplinary Perspectives on Validity: Bridging the Gap Between Design and Implementation for Technology-Enhanced Learning Systems*.
02 Jurgen van den Hoogen (JADS), *Time Series Analysis Using Convolutional Neural Networks*.
03 Andra-Denis Ionescu (TUD), *Feature Discovery for Data-Centric AI*.
04 Rianne Schouten (TU/e), *Exceptional Model Mining for Hierarchical Data*.
05 Nele Albers (TUD), *Psychology-Informed Reinforcement Learning for Situated Virtual Coaching in Smoking Cessation*.
06 Daniël Vos (TUD), *Decision Tree Learning: Algorithms for Robust Prediction and Policy Optimization*.
07 Ricky Maulana Fajri (TU/e), *Towards Safer Active Learning: Dealing with Unwanted Biases, Graph-Structured Data, Adversary, and Data Imbalance*.
08 Stefan Bloemheuvel (TiU), *Spatio-Temporal Analysis Through Graphs: Predictive Modeling and Graph Construction*.
09 Fadime Kaya (VUA), *Decentralized Governance Design - A Model-Based Approach*.
10 Zhao Yang (UL), *Enhancing Autonomy and Efficiency in Goal-Conditioned Reinforcement Learning*.
11 Shahin Sharifi Noorian (TUD), *From Recognition to Understanding: Enriching Visual Models Through Multi-Modal Semantic Integration*.
12 Lijun Lyu (TUD), *Interpretability in Neural Information Retrieval*.
13 Fuda van Diggelen (VUA), *Robots Need Some Education: on the complexity of learning in evolutionary robotics*.
14 Gennaro Gala (TU/e), *Probabilistic Generative Modeling with Latent Variable Hierarchies*.
15 Michiel van der Meer (UL), *Opinion Diversity through Hybrid Intelligence*.
16 Monika Grewal (TU Delft), *Deep Learning for Landmark Detection, Segmentation, and Multi-Objective Deformable Registration in Medical Imaging*.
17 Matteo De Carlo (VUA), *Real Robot Reproduction: Towards Evolving Robotic Ecosystems*.
18 Anouk Neerincx (UU), *Robots That Care: How Social Robots Can Boost Children's Mental Wellbeing*.
19 Fang Hou (UU), *Trust in Software Ecosystems*.
20 Alexander Melchior (UU), *Modelling for Policy is More Than Policy Modelling (The Useful Application of Agent-Based Modelling in Complex Policy Processes)*.
21 Mandani Ntekouli (UM), *Bridging Individual and Group Perspectives in Psychopathology: Computational Modeling Approaches using Ecological Momentary Assessment Data*.
22 Hilde Weerts (TU/e), *Decoding Algorithmic Fairness: Towards Interdisciplinary Understanding of Fairness and Discrimination in Algorithmic Decision-Making*.
23 Roderick van der Weerdt (VUA), *IoT Measurement Knowledge Graphs: Constructing, Working and Learning with IoT Measurement Data as a Knowledge Graph*.
24 Zhong Li (UL), *Trustworthy Anomaly Detection for Smart Manufacturing*.
25 Kyana van Eijndhoven (TiU), *A Breakdown of Breakdowns: Multi-Level Team Coordination Dynamics under Stressful Conditions*.
26 Tom Pepels (UM), *Monte-Carlo Tree Search is Work in Progress*.

27 Danil Provodin (JADS, TU/e), *Sequential Decision Making Under Complex Feedback*.
28 Jinke He (TU Delft), *Exploring Learned Abstract Models for Efficient Planning and Learning*.
29 Erik van Haeringen (VUA), *Mixed Feelings: Simulating Emotion Contagion in Groups*.
30 Myrthe Reuver (VUA), *A Puzzle of Perspectives: Interdisciplinary Language Technology for Responsible News Recommendation*.
31 Gebrekirstos Gebreselassie Gebremeskel (RUN), *Spotlight on Recommender Systems: Contributions to Selected Components in the Recommendation Pipeline*.
32 Ryan Brate (UU), *Words Matter: A Computational Toolkit for Charged Terms*.
33 Merle Reimann (VUA), *Speaking the Same Language: Spoken Capability Communication in Human-Agent and Human-Robot Interaction*.
34 Eduard C. Groen (UU), *Crowd-Based Requirements Engineering*.
35 Urja Khurana (VUA), *From Concept To Impact: Toward More Robust Language Model Deployment*.
36 Anna Maria Wegmann (UU), *Say the Same but Differently: Computational Approaches to Stylistic Variation and Paraphrasing*.
37 Chris Kamphuis (RUN), *Exploring Relations and Graphs for Information Retrieval*.
38 Valentina Maccatrozzo (VUA), *Break the Bubble: Semantic Patterns for Serendipity*.
39 Dimitrios Alivanistos (VUA), *Knowledge Graphs & Transformers for Hypothesis Generation: Accelerating Scientific Discovery in the Era of Artificial Intelligence*.
40 Stefan Grafberger (UvA), *Declarative Machine Learning Pipeline Management via Logical Query Plans*.
41 Mozhgan Vazifehdoostirani (TU/e), *Leveraging Process Flexibility to Improve Process Outcome - From Descriptive Analytics to Actionable Insights*.
42 Margherita Martorana (VUA), *Semantic Interpretation of Dataless Tables: a metadata-driven approach for findable, accessible, interoperable and reusable restricted access data*.
43 Krist Shingjergji (OU), *Sense the Classroom - Using AI to Detect and Respond to Learning-Centered Affective States in Online Education*.
44 Robbert Reijnen (TU/e), *Dynamic Algorithm Configuration for Machine Scheduling Using Deep Reinforcement Learning*.
45 Anjana Mohandas Sheeladevi (VUA), *Occupant-Centric Energy Management: Balancing Privacy, Well-being and Sustainability in Smart Buildings*.
46 Ya Song (TU/e), *Graph Neural Networks for Modeling Temporal and Spatial Dimensions in Industrial Decision-making*.
47 Tom Kouwenhoven (UL), *Collaborative Meaning-Making. The Emergence of Novel Languages in Humans, Machines, and Human-Machine Interactions*.
48 Evy van Weelden (TiU), *Integrating Virtual Reality and Neurophysiology in Flight Training*.
49 Selene Báez Santamaría (VUA), *Knowledge-centered conversational agents with a drive to learn*.
50 Lea Krause (VUA), *Contextualising Conversational AI*.
51 Jiaxu Zhao (TU/e), *Understanding and Mitigating Unwanted Biases in Generative Language Models*.
52 Qiao Xiao (TU/e), *Model, Data and Communication Sparsity for Efficient Training of Neural Networks*.
53 Gaole He (TUD), *Towards Effective Human-AI Collaboration: Promoting Appropriate Reliance on AI Systems*.
54 Go Sugimoto (VUA), *Missing Links: Investigating the Quality of Linked Data and its Tools in Cultural Heritage and Digital Humanities*.
55 Sietze Kai Kuilman (TUD), *AI that Glitters is Not Gold: Requirements for Meaningful Control of AI Systems*.
56 Wijnand van Woerkom (UU), *A Fortiori Case-Based Reasoning: Formal Studies with Applications in Artificial Intelligence and Law*.

List of Scientific Publications

van Woerkom, W., Grossi, D., Prakken, H., & Verheij, B. (2022a, June). Landmarks in case-based reasoning: From theory to data. In Stefan Schlobach, María Pérez-Ortiz, & Myrthe Tielman (Eds.), *HHAI2022: Augmenting Human Intellect. Proceedings of the First International Conference on Hybrid Human-Artificial Intelligence* (pp. 212–224, Vol. 354). IOS Press. https://doi.org/10.3233/FAIA220200

van Woerkom, W., Grossi, D., Prakken, H., & Verheij, B. (2022b). Justification in case-based reasoning. *1st International Workshop on Argumentation for eXplainable AI*, *3209*. https://ceur-ws.org/Vol-3209/#5942

van Woerkom, W., Grossi, D., Prakken, H., & Verheij, B. (2023a). Hierarchical precedential constraint. *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, 333–342. https://doi.org/10.1145/3594536.3595154

van Woerkom, W., Grossi, D., Prakken, H., & Verheij, B. (2023b, December). Hierarchical *a fortiori* reasoning with dimensions. In G. Sileno, J. Spanakis, & G. van Dijck (Eds.), *Legal Knowledge and Information Systems. JURIX 2023: The Thirty-sixth Annual Conference* (pp. 43–52). IOS Press. https://doi.org/10.3233/FAIA230944

van Woerkom, W., Grossi, D., Prakken, H., & Verheij, B. (2024a). *A Fortiori* case-based reasoning: From theory to data. *Journal of Artificial Intelligence Research*, *81*, 401–441. https://doi.org/10.1613/jair.1.15178

van Woerkom, W., Grossi, D., Prakken, H., & Verheij, B. (2024b, December). A case-based-reasoning analysis of the COMPAS dataset. In Jaromir Savelka, Jakub Harasta, Tereza Novotna, & Jakub Misek (Eds.), *Legal Knowledge and Information Systems. JURIX 2024: The Thirty-seventh Annual Conference* (pp. 180–190, Vol. 395). IOS Press. https://doi.org/10.3233/FAIA241244

van Woerkom, W., Grossi, D., Prakken, H., & Verheij, B. (2025). Hierarchical models of precedential constraint [In press]. *Artificial Intelligence and Law*.

van Woerkom, W. (2025). Formal results on case-base consistency: A COMPAS case study. *Proceedings of the Twentieth International Conference on Artificial Intelligence and Law*. https://doi.org/10.1145/3769126.3769190

Curriculum Vitæ

Education

2014–2017	Bachelor's degree	in **Artificial Intelligence** at the *University of Amsterdam*.
2017–2021	Master's degree	in **Logic** (mathematics track) at the *University of Amsterdam*.
2021	Summer school	on **Human Centered Artificial Intelligence** (ACAI 2021), organized by the *Human AI Net* in Berlin, Germany.
2021–2023	SIKS courses	in various topics on Artificial Intelligence, offered by the *Netherlands Research School for Information and Knowledge Systems*.
2023	Summer school	on **Artificial Intelligence and law** at the *European University Institute* in Florence, Italy.

Academic Work

2015–2016	Academic tutor	at the *University of Amsterdam*.
2015–2019	Teaching assistant	for various courses on Artificial Intelligence, at the *University of Amsterdam*.
2021–2024	Bachelor thesis supervisor	at *Utrecht University*.
2021–2024	Workgroup lecturer	for the *Introduction to Adaptive Systems* and *Data Analysis and Retrieval* courses at *Utrecht University*.
2021–2025	Doctoral researcher	at *Utrecht University*.
2025–	Postdoctoral researcher	at the *Max Planck Institute for Comparative and International Private Law* in Hamburg, Germany.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisors Henry, Bart, and Davide for their continued support over the past four years. I am incredibly lucky to have had three supervisors who were all so engaged with my work, which manifested itself in many fruitful discussions and extensive feedback on drafts. It is a testament to your qualities as a supervisory team that the trajectory went rather smoothly, despite my best efforts to complicate matters through stubbornness and working too close to deadlines. To the members of my committee—Floris Bex, Jan Broersen, John Horty, Giovanni Sartor, and Annette ten Teije—thank you for putting in the time and effort to read and evaluate this work.

I would also like to take this opportunity to thank Alban Ponse and Benno van den Berg, who have supervised me during my time studying at the University of Amsterdam. Doing research projects as part of my studies, under such excellent supervision, was an invaluable experience and this dissertation would have looked very differently without it.

During my time at Utrecht University I've met many great colleagues. I enjoyed the many dinners and lunches that I have had with Changxi, Hui and Shuai over the years. From the Hybrid Intelligence Center, my friends and colleagues Cor and Ludi have been a constant presence throughout my PhD trajectory; we really did travel all over the world went through (or should I say, *survived*) a lot together—here's to many more to come.

Most of this dissertation was written in the study room at the ILLC. I would like to thank the other dwellers of that room—Hugh Mee, Paul, the other Paul, Wessel, Giacomo, Swapnil and Shin—for their excellent company, the lively discussions in the "silent" room, the lunches, and all the trips to the coffee machine. To my fellow masters of logic with an appetite for holidays: Bas, Freddy, Lukas, and Simon—thank you for the many laughs, the endless nights working at the ILLC and later, our meet-ups across Europe. Thanks to the members of the dinner gang—Paul, Swapnil, Jeremy, Shin, Hugh Mee—for the meals that we have shared and the ludic post-dinner challenges. Lydia, thank you for the countless games of foosball and the sunny BBQs—I look forward to many more. Nathan, bedankt voor het stappen, het haringhappen, en het delen van je aanstekelijke voorliefde voor taal.

Traveling back in time once more, I would like to thank some of the friends that I made during my time in the Bachelor's program. Max, it is always a pleasure to speak with you and I'm glad we keep in touch, even in the absence of external pressures to do so. The same goes for Cor; our friendship has been a profoundly enjoyable constant since the moment we first met.

Special thanks go out to my paranymphs—thank you for sharing the organizational burden of my defense, but most of all for your friendship over the years. Paul, thank you for the many category-theoretic discussions and all the sporting activities: the cycling, the swimming, and let us not forget our table tennis rivalry that culminated in a winner-takes-all match that I won gloriously. Wessel, thank you for all the jest, gezelligheid and the dinners. I would be a sitting duck before the committee if it weren't for the resilience you've instilled in me.

I am also deeply grateful to my friends outside academia who have always been there for me. Tim, I will always treasure our nights of playing pool and never-ending banter. Ray, thank you for providing laughter, companionship, and the best cocktails. Life would be dull without the two of you.

Since handing in this work, I've moved to Hamburg, and I was very warmly welcomed there by a group of people whom I'd like to take this opportunity to thank. Katharina, Oskar, Vera, thank you for being such wonderful and inspiring colleagues. Sebastian, Sofia, Robin, Michael, and Freddy—I'm blessed to have met you all.

Hugh Mee, my dear girlfriend, thank you for always being by my side. Whether it was on one of our trips to beautiful remote places, or during late night working sessions, you've supported me through thick and thin and it means the world to me.

Finally, to my family. Rix, thank you for the beautiful cover of this thesis, and for all those late night discussions we've had over the years. Thank you to my parents Adri and Annemarie, and to my siblings Wouter, Willemijn, and Wibout, for believing in me and encouraging me to live up to my potential—I could not have done this without you.