Explaining machine learning outputs to humans: a case-based reasoning approach

Elisa Friscione

Supervisor: Henry Prakken
Second examiner: Floris J. Bex

Utrecht University

May 14th, 2019

Abstract

The problem of interpretability, in other words the problem of explaining machine learning outputs in terms that are understandable for a human has become a widely debated topic in the field of AI. In particular, this work is concerned with explanations of machine learning outputs in the legal domain. HYPO was chosen as the blueprint implementation for the current model of explanation, which builds on HYPO while also attempting to improve it. The resulting model was tested on two case studies, and the yielded outputs were compared against the ML outputs.

Contents

1 Introduction

The problem of interpretability, that is, the problem of explaining an algorithm's output in terms that a human can understand has recently become an intensely debated topic, especially after a law on data protection and privacy, the General Data Protection Regulation (GDPR), was introduced. This regulation includes the so-called *right to explanation* which, as the name suggests, is a person's right to be given an explanation for an algorithm's output. Although the right to explanation existed prior to the enactment of the GDPR, it sparked the debate anew. The reason is that the right to explanation is, in itself, quite controversial: does it refer to the explanation of a *particular* output, or the explanation of the whole model? What is certain beyond this dispute is that the need to provide additional information concerning automated decisions has been acknowledged.

There is, however, a first important issue. Given the fact that some machine learning algorithms are black boxes, how can such explanations be formulated in the first place? A first answer might be that algorithms ought to be transparent: the assumption is that, by revealing the algorithm's source code, then the algorithm's behaviour can be explained. Although appealing, there are limitations to this solution. For instance, transparency may not be a viable option: for an algorithm to be transparent all of its inputs, parameters, and computations would have to be divulged to the user, but it is often the case that the data involved is sensitive (e.g., a person's medical record), and cannot be legally disclosed. Furthermore, preventing people from knowing the specifics of an algorithm also deters them from trying to game the system (Kroll J.A. et al., 2016). Finally, it can be debated whether access to the code provides an appropriate explanation of an algorithm's output. In fact, "Transparency advocates often claim that by reviewing a program's disclosed source code, an analyst will be able to determine how a program behaves." (Kroll J.A. et al., 2016). This view is naive in the sense that it does not take into account how difficult such an analysis can be (Kroll J.A. et al., 2016). Moreover, it is unlikely that the average person would be capable of understanding the algorithm's source code. The nature of the user's expertise appears to be of vital importance for the design of interpretable algorithms (Guidotti R. et al., 2018), reason for which transparency alone cannot solve the problem of interpretability.

There is at least one alternative. If the objective is the explanation of an algorithm's decision, then it is not necessary for the whole process to be explained: the rationale that yielded a particular output is the question of interest in the current work. By clarifying the rationale it becomes possible to question whether a certain rule was applied fairly, and so on.

Such explanations should be in some understandable format in order to allow people to both comprehend the output, and to eventually challenge it. This is not different from the way humans provide explanations, too: when explaining an action or a decision, a person will not give a detailed account of the neurological and psychological processes that lead her to that choice, but she will explain it in terms of reasons and beliefs. For these reasons, human under-

standing and how humans understand and make explanations are good starting points for this research field. Thus it is necessary to define what makes a good explanation (a question which answer might depend on the field of application), as well as look at theories from psychology and philosophy in order to grasp the workings of human understanding.

The nature of this problem is very general, and there might be general solutions to it. However, the contexts that would benefit from explanations of machine learning outputs are so diverse that, as a first step, pursuing a tailor-made solution is probably the best option. In this research legal applications will be considered.

The question concerning human understanding is going to be addressed in the following section of this introduction. Once an informal, working definition of explanation is given, the problem of providing explainable outputs is explored from the perspective of interpretable AI, and the kind of interpretability sought in this work is determined. Afterwards, the scope of this research is restricted to the area of AI and law, and relevant literature on the subject is going to be presented in the last section of this introduction. Finally, the objectives and methodology are introduced.

The remainder of this paper is organised as follows: in section 2 the basic components for the current model of explanation are introduced, and the limitations of the chosen approach, as well as possible solutions to such issues are investigated. In section 3 the model is more systematically presented, and examples are provided in section 4. After addressing some questions of interest for future research, the conclusions drawn from this research are reported in section 5.

1.1 Understanding and explanation

Understanding human understanding and what makes a good explanation are key aspects in this research field. To say that there is an intrinsic relationship between the two might seem a banal statement: in fact, an explanation's purpose is to provide understanding. However, whereas epistemology and philosophy of science have been long concerned with defining what an explanation is (especially a scientific explanation), only recently has understanding become a central topic in these fields.

Different accounts of what understanding is exist, such as operational or behaviourist theories that link understanding to a set of abilities, but a more general, intuitive definition of understanding will be used in this work. According to Van Camp (Van Camp W., 2014) it is possible to distinguish between *genuine understanding*, the *belief of understanding*, and the *feeling of understanding*. Clearly, explanation has to provide genuine understanding (Van Camp W., 2014): similarly to knowledge, which can be defined as justified true belief, genuine understanding occurs when our beliefs are confirmed by a certain state of the world. Thus, what is genuine understanding?

Van Camp's definition of *genuine understanding* is based on the psychological concept of understanding, according to which "its key characteristic is that

it involves assimilation of the information to be understood into an interconnected framework of knowledge." (Van Camp W., 2014). Van Camp highlights how understanding is similar to knowledge, and yet different: whereas knowledge can be isolated, "To understand something is not simply to know it, but to know of its relation to other knowledge and to be aware of those relationships, to know that knowledge structure." (Van Camp W., 2014). Hence, he concludes, "Understanding is about making connections between various pieces of information. Thus explanation is about describing the various connections between facts." (Van Camp W., 2014).

If these definitions are accepted, then it is apparent how a user's expertise, that is, the user's knowledge, is important in the design of explanations in the field of AI. If understanding depends on "what our body of knowledge is, and contingent facts about our cognitive capacity for forming knowledge structures," (Van Camp W., 2014), then they ought to be taken into account when explaining a whole model, or a model's output. Indeed, "knowing the user's experience in the task is a key aspect of the perception of the interpretability of a model." (Guidotti R. et al., 2018).

How do explanations provide understanding? According to (Van Camp W., 2014), there are two different ways in which an explanation can occur. In the first case, explanations increase our body of knowledge; in the second case, explanations provide an account for those facts or events that do not fit with an existent body of beliefs, thus "An explanation would bring about understanding by accounting for apparent conflicts, or by removing them." (Van Camp W., 2014). Similarly, according to some explanations are *contrastive*: it is argued that "people do not ask why event P happened, but rather why event P happened instead of some event Q." (Miller T., 2017). Explanations seem to have more qualities that are relevant for this work. For instance, the fact that explanations are selected, that is, cognitive biases influence how humans pick an explanation over a great number of other explanations (Miller T., 2017); secondly, that probabilities and statistical generalisations do not always provide the best explanations for people, "unless accompanied by an underlying *causal* explanation for the generalisation itself." (Miller T., 2017); lastly, there is the fact that explanations are social, in that they are a transfer of knowledge that occurs through conversation or other interactions (Miller T., 2017). These points will be particularly relevant in the later sections of this paper.

As stated before, this work is concerned with legal applications. How people understand judicial decisions in particular, and which kinds of explanations provide better understanding are both questions that require empirical testing. Surely such testing has to account for the different levels of expertise of the people who engage with the law, from a common person to a supreme court judge. Whereas the first might prefer a simpler, more intuitive explanation, the latter might prefer a detailed, larger account explaining a certain decision, a point which is also suggested in (Guidotti R. et al., 2018). Due to time constraints, however, this work uses the following informal definition of explanation. This notion, based on the several contributions from the literature discussed in this section, should be viewed as a principle that a model of explanation must fulfil.

Definition 1. [Explanation] A good explanation is one that takes into account the cognitive biases a person brings to understanding, as well as her background knowledge, cognitive abilities, and possible time constraints. A good explanation does not only increase a person's knowledge, but it accounts for cases that seem to contradict her set of beliefs.

Now that the standards and requirements for machine learning explanations have been explored from the human side, it is possible to look at the field of interpretable AI. Two main questions have to be answered: firstly, what kind of interpretability is sought in this research? Secondly, which standards of interpretability should be pursued based on the literature on human understanding?

1.2 Interpretability and explainability

In the context of machine learning interpretability is meant as the ability to explain an output in terms that are understandable to humans. However, beyond this general definition there is little agreement on the use of the terms "interpretability" and "interpretable," as well as on which models can be considered interpretable, or which features should be taken into account by research on interpretability (Lipton Z.C., 2016). Here two main papers will be considered: Lipton's *The mythos of model interpretability* (Lipton Z.C., 2016) and Guidotti and colleagues' *A survey of methods for explaining black box models* (Guidotti R. et al., 2018). Whereas Lipton's work attempts to give a proper definition of interpretability, the research carried out by Guidotti and colleagues is concerned with investigating and collecting solutions to the various issues of interpretability. Even though there are similarities between the two papers, there are still divergences that show how a standard has yet to be reached in this research field.

Guidotti and colleagues characterise interpretability in terms of dimensions (Guidotti R. et al., 2018). The first one is the difference between global and local interpretability: either the whole logic of the model can be made understandable, or just an algorithm's particular output can be explained. Secondly, they mention time limitations as another important feature of interpretability: the amount of time a user is allowed to understand an explanation is quite important, as the user might be in a situation that requires immediate action. Finally, they highlight how the user's expertise has to be taken into account by research on interpretability: for example, a more expert user might prefer a detailed explanation compared to someone with a more superficial knowledge of the issue at hand. Lipton does not mention time limitations and user's expertise, but Guidotti et al.'s distinction between global and local interpretability can be translated into transparency and post-hoc interpretability in Lipton's terms. In fact, the aim of transparency is to explain how the model works (either at the level of the whole model, at the level of its parameters, or at the level of the learning algorithm itself), whereas post-hoc interpretability provides information "after-the-fact" (Lipton Z.C., 2016). Even though Lipton's paper does not state clearly whether post-hoc interpretability concerns a particular decision or

whether it can be used to explain the whole model, post-hoc interpretability is here considered to be aimed at explaining single decisions (otherwise post-hoc interpretability would just turn out to be another transparency issue).

Furthermore, unlike Guidotti and colleagues Lipton appears to have a stronger stance regarding which aspect of interpretability should be pursued. His argument that humans exhibit none of the different forms of transparency (i.e., they are not globally interpretable), and therefore such a strict standard should not be applied to algorithms as well is quite persuasive. This view is reasonable, above all in light of one of the features of interpretability discussed by Guidotti et al., the user's expertise. Transparency, i.e. the exposition of some (source) code, would not provide a good explanation to those users who have no knowledge about machine learning, programming, and so on. Moreover, whereas a trade-off between transparency and performance is necessary, if transparency is given up and post-hoc interpretability is pursued, then such a trade-off would no longer be an issue (Lipton Z.C., 2016).

The two papers address very similar desiderata of interpretability research, such as, among others, trust, fairness, usability, and reliability. A point of disagreement appears to be the relationship between complexity and interpretability. Guidotti et al. state that the complexity of the predictive model, i.e. the model's size, is a component for measuring interpretability: simpler models are more interpretable than bigger ones. In his work, Lipton asserts that simpler models such as linear regression and decision trees are not intrinsically more interpretable. His argument is once again based on one of the features of interpretability mentioned by Guidotti et al., time limitations. A model can be said to be transparent if the user is capable of carrying out its computations, taken its input and parameters, in a reasonable amount of time. However, human cognitive abilities are limited, reason for which it is not possible to quantify how much time is a reasonable amount of time (Lipton Z.C., 2016). Hence, even though the dimensions of interpretability Guidotti and colleagues suggest appear to be relevant and reasonable, it would seem that they do not see their full consequences in their own work.

For the purposes of this research post-hoc interpretability should be considered while taking into account the two dimensions from Guidotti et al.'s work, time limitations and user's expertise. Both of these factors play an important role in human understanding and the design of explanations. Then, since this work is concerned with applications in the legal field, it should be asked in which manner explanations are delivered in this context. Perhaps obviously, arguments are one of the main tools in the field. Can legal arguments be modelled through a logic of argumentation in order to provide explanations of the same kind? Even though logic is not mentioned in Guidotti et al.'s survey of methods for explaining a black box's output (Guidotti R. et al., 2018), Grabmair's VJAP shows how argumentation can be used both to make predictions, and then to explain said predictions (for more details about this model, we refer to (Grabmair M., 2017)). Given a black box and the type of data it processes, it might be possible to formalise the problem in some commonsense reasoning logic.

At this point a question can still be asked: do we really need interpretable AI? Is it not enough that machine learning models perform well? Can we not rely on a model that fits the data? Although persuasive, arguments about a model's fitness are not necessarily good arguments, as explained in (Roberts S. & Pashler H., 2000). Even though their work is mostly concerned with the use of a model's fitness as an argument to support psychological theories, there is criticism that applies to the use of this argument in general. For instance there is the problem of overfitting, that is, the problem that arises whenever a model performs extremely well with certain data, but fails to fit additional or unseen data. An unjustified number of parameters may cause overfitting. Secondly, it is possible for two equally flexible theories to fit the same data while making very different assumptions about it. In other words, a model's performance does not say much about how accurate the model's assumptions are. Indeed, sometimes the problem is determining the criteria used by the model in order to classify the data: is this criteria reasonable to the human modeller, or did the model catch a pattern in the data that is irrelevant to humans?

In other words, how can we assess an algorithm's rationale? Again, how humans reason should be a source of inspiration if we want algorithms to deliver explanations in a similar fashion. Legal argumentation models how humans reason in the legal domain; this knowledge has already been applied in the field of AI and law, which lead to the creation of systems such as HYPO and CATO. Both are discussed in the following section.

1.3 Legal argumentation

A logic of argumentation would seem to suit particularly well legal applications. Argumentation is a central notion in law, as there are several aspects of it that involve "appeals to precedent, principle, policy, and purpose, and involves the *attack* as well as the *construction* of arguments." (Prakken H. & Sartor G., 2004).

Legal reasoning is characterised by several types of inferences, each reflecting the different aspects law can deal with. For example, when determining the content of laws and legal concepts the "prevailing modes of reasoning are analogy, appeals to precedent or policy, and the balancing of interests." (Prakken H. & Sartor G., 2004). When inferring legal consequences deductive inference is used, although enriched with nonmonotonic reasoning in order to deal with conflicting rules or exceptions (Prakken H. & Sartor G., 2004). This means that legal arguments can be translated in different ways, such as using syllogisms or modus ponens (Feteris E. & Kloosterhuis H., 2011), or through the use of argumentation schemes (Gordon T.F. & Walton D., 2009). Argumentation schemes, in general, are prototypical kinds of inferences which can be attacked by so-called critical questions (Prakken H. & Sartor G., 2004). Witness testimony is an example of such an argument scheme, which can be expressed as *If W says P, and W was in the position to observe P, therefore (presumably) P.* A way of attacking this scheme is by pointing to the fact that the witness is proved to be unreliable. Lastly, reasoning by analogy seems to be characteristic espe-

cially of common law, in which "similarities between the facts are advanced as reasons for recommending or justifying the same results." (Twining W., 1999), Similarly, precedents can be cited in the form of competing analogies (Twining W., 1999).

Arguments can be said to be the way through which people explain and justify why a certain decision should be taken. Hence, if the objective is building a model of explanation, looking at the arguments humans themselves use in the legal domain to explain certain outcomes appears to be a sensible choice. In other words, argumentation is both a tool actually used in real life situations, as well as a logical tool which is suggested to suit a model of explanation. Indeed, systems that can reason with and about cases in order to explain outcomes are well-established in the legal domain, such as HYPO and CATO. Among the inference strategies mentioned above, these systems especially rely on reasoning by analogy. At first glance such an approach might seem limited, but as it will be shown, HYPO and CATO provide a variety of tools that successfully capture the domain they are applied in. These tools are suggested to be useful in a model of explanation, too.

1.3.1 HYPO

In HYPO legal arguments are generated by citing precedents; these previous cases provide justifications for legal decisions about whether plaintiff or defendant should win a certain dispute. According to Ashley, what these justifications state is that "1) in the precedent, a prior court resolved the competing factors in favour of a particular side; 2) the current situation is analogous to the precedent because it involves the same competing factors; 3) therefore, the current dispute should be decided the same way." (Ashley K.D., 1991). HYPO compares the current problem situation with previous cases in its case knowledge base (CKB), and it uses such comparisons for its inferences. In (Ashley K.D., 1991) HYPO's process is explained to go through these nine steps:

1. "Analyse the problem situation dimensionally;"
2. "Retrieve relevant cases from the CKB;"
3. "Select relevant cases that are the most on point on the problem;"
4. "Select most on point cases that are best for each side to cite;"
5. "Compute distinctions between the best cases and the problem;"
6. "Identify counter-examples to the best cases;"
7. "Evaluate and summarise overall argument;"
8. "Generate 3-Ply arguments citing the best cases, distinction, and counter-examples;"
9. "Generate hypotheticals to strengthen or weaken arguments."

Dimensions are one of the main characteristic of HYPO. There is a total of 13 dimensions, which correspond to 13 factors relevant in the trade secrets domain. An example is the dimension *Competitive-Advantage*, in which plaintiff's position is strengthened "the greater the competitive advantage gained by the defendant." (Ashley K.D., 1991). Another example is the dimension *Disclosure-In-Negotiations*, in which plaintiff's case is stronger to the extent that the trade secret was not disclosed to the defendant during negotiations. Dimensions can have different types of ranges (Ashley K.D., 1991), but the main point is that they capture how strong or weak a party's position is, allowing to compare the magnitude of competing factors. This means that the cases comprising the CKB are also analysed by HYPO in order to determine the relevant dimensions.

For HYPO to select the most on-point cases it is necessary to search the shared dimensions between cases, that is, the relevant similarities: "The set of relevant similarities, $S(c_1, c_2)$, between two cases c_1 and c_2 is the intersection of the set of dimensions that apply to c_1 and the set of dimensions that apply to c_2." (Ashley K.D., 1991). A case is on point if the set of shared dimensions is not empty. Among the on-point cases it is possible to compute the most on-point cases (MOP): "saying that a case, c_i, is more on point than another case, c_k, means that the set of relevant similarities between the problem p and c_k is a proper subset of $S(p, c_i)$." Ashley K.D. (1991). Comparing which cases are more on-point does not mean the *number* of shared dimensions is taken into account, but rather the "overlaps of the sets of dimensions they share with the problem." (Ashley K.D., 1991).

Most on-point cases are candidates as best cases to cite for each side. Intuitively, a best case should share with the current problem situation at least one dimension that favours the side citing it. A most on-point case that shares dimensions that are favourable only to the opponent might become a counter-example. Once such cases are identified, the relevant differences between them can be computed. Ashley makes the example of a cited case that favours the plaintiff; then the set of relevant differences is given by the union of three sets, namely "(1) the pro-defendant dimensions that apply only to the problem; (2) the pro-plaintiff dimensions that apply only to the cited case; and (3) the shared dimensions that favour the plaintiff more strongly in the cited case than the problem." Ashley K.D. (1991). A counter-example is a precedent which outcome is contrary to the cited case. Since HYPO computes each sides' best arguments, it also looks for each sides' counter-examples.

Then, HYPO evaluates and reports which side has stronger arguments by listing each side's untrumped cases, and it outputs 3-Ply arguments: HYPO first makes a point for one side, drawing analogies between the current problem situation and the precedent, to which it responds by citing counter-examples, which in turn can be rebutted. Finally, HYPO "suggests hypothetical modifications of the problem situation in which a side's argument would be strengthened or weakened," that is, modifications are applied so that "new cases become relevant opening up new possible points and responses." Ashley K.D. (1991). Such modifications involve most on-point near miss cases and potentially more on-point counter-examples, but for a more detailed description of HYPO's process

we refer to Ashley K.D. (1991).

1.3.2 CATO

Similarly to HYPO, in CATO an account of the arguments used by the two sides arguing for a favourable outcome is provided, and the current problem situation is compared to the relevant past cases. Again, the domain is trade secret law. The main new feature introduced in CATO is the representation and application of *middle-level normative background knowledge*, which is called *factor hierarchy*: this factor hierarchy "covers the basic requirements of a claim for trade secret misappropriation," however it "is not meant to state necessary or sufficient conditions for winning a claim of trade secret misappropriation." (Aleven V., 2003). At the bottom of the hierarchy 26 factors used to represent cases are found, and each is linked to 11 *intermediate legal concerns*, which in turn are linked to 5 legal issues at the top of the hierarchy. Intermediate legal concerns and legal issues are also called "abstract factors" or "high-level factors." The links can be either positive or negative, that is, a factor either supports or not a certain conclusion. An excerpt of CATO's factor hierarchy can be found in (Aleven V., 2003).

The second innovation of the CATO's model is its ability to use the factor hierarchy for generating multi-case arguments organised by issues. As a first step, CATO identifies the issues raised by the problem, which depend on the factors present in the case. In fact, "to identify issues in a problem or case, CATO collects all Legal Issues in the Factor Hierarchy that are linked to the case's applicable factors." (Aleven V., 2003). For example, if a case contains the conflicting factors F15 Unique-Product, and F16 Info-Reverse-Engineerable, then these factors raise issue F101, which concerns the question whether plaintiff's information actually constitutes a trade secret. Once the issues have been identified, CATO generates arguments to address each of them. Again, the model relies on its background knowledge in several ways (Aleven V., 2003):

1. "to identify issues in a problem;"
2. "in the discussion of an issue, to focus on the factual strengths and weaknesses (factors) that are related to the issue and to cite cases to emphasise strengths and downplay weaknesses;"
3. "to give reasons why particular factual strengths matter to an issue being discussed;"
4. "to find strengths that are closely related to weaknesses and therefore compensate for those weaknesses."

For a more detailed description of this process, we refer to (Aleven V., 2003).

Another feature of the CATO model is that it employs background knowledge in order to reason about the significance of distinctions: it is argued in the paper that "much legal argument [...] involves debating whether a case is really the same as the problem or not." (Aleven V., 2003). In order to emphasise the

significance of a distinction, the model points out how the problem and the case differ with respect to some abstract factor(s), which shows how the two issues are different on a deeper level. When the significance of a distinction is downplayed, CATO shows that a parallel exists at a more abstract level, arguing that the distinction is a matter of details, such as different factors that nonetheless point to the same, more abstract conclusion. Furthermore, a method for downplaying a distinction is by showing an opposite interpretation: CATO shows that "the case in which the distinction occurs has additional factors that support an interpretation of that case that is opposite to an interpretation suggested by the presence of the distinction." (Aleven V., 2003).

Finally, CATO uses both background knowledge and reasoning about the significance of distinctions to select the most relevant cases. Both methods, which build on HYPO's own criterion for selecting the best untrumped cases, are described in detail in (Aleven V., 2003).

1.4 Research questions

The research aims to answer the following questions:

1. Which among the tools provided by case-based reasoning systems best suit methods of explanation for machine learning outputs?
    (a) Does the chosen approach provide the sufficient tools for generating explanations for ML outputs?
2. How often do the resulting model's outputs correspond to the original machine learning outputs?

In order to answer these questions a literature review was carried out, and two case studies were chosen in order to test the current work's approach. Bench-Capon provided the datasets from the experiments (Bench-Capon T.J., 1993) described in the following section, and ProPublica made available the datasets reporting recidivism scores computed by the COMPAS algorithm (ProPublica, 2016).

The first step was determining which case-based reasoning tools were the best options for a method of explanation. Once defined, the basic components have been adapted to the current problem situation, and a model was created through a bottom-up approach.

Once the model is obtained, it is going to be tested on the two case studies more thoroughly described in the following sections.

1.5 Case studies

Here two case studies are introduced. The first builds on Bench-Capon's social welfare benefit problem (Bench-Capon T.J., 1993), whereas the second is concerned with an algorithm used for computing recidivism scores (ProPublica, 2016).

1.5.1 The social welfare benefit problem

In (Bench-Capon T.J., 1993) an artificial problem was created in order to assess a neural network's performance, as well as to test the rationale for the classifications. The problem was a fictional welfare benefit paid to pensioners in order to contribute to the expenses when visiting a spouse in a hospital. There were a total of six conditions (Bench-Capon T.J., 1993):

- The person should be of pensionable age (age 65 for men, age 60 for women);
- Out of the last five contribution years, the person should have paid four of them;
- The person must be a spouse of the patient;
- The person should not be absent from the UK;
- The person should not have capital resources that amount to more than £3000;
- If the relative is an in-patient, the hospital should be within a certain distance; if an out-patient, beyond that distance.

The conditions were used to generate the data, and more importantly, to assess the model's performance in terms of the rationale employed: because the rules are not known by the model, it was possible to compare the rules the model drew from the training set against the experimenter's conditions.

Three neural networks (one with one hidden layer, one with two hidden layers, and one with three hidden layers) were trained on a dataset consisting of 50% satisfying cases, and the remaining half failing on each of the conditions. The three neural networks all converged, showing above-chance performance levels (the one hidden layer net reaching 99.25% of correctly classified cases), but upon closer inspection the analysis of the neural networks proved dissatisfying. For example, even though sex was an important condition, only the three hidden layers network took it into consideration, and even then it was highly inaccurate. Thus those networks could achieve an acceptable performance while ignoring relevant conditions for the problem at hand. Moreover, as Bench-Capon notes (Bench-Capon T.J., 1993), it was possible to identify that the conditions were ignored simply because the problem was an artificial one, in other words he possessed prior knowledge about it. In most cases, however, we do not possess such knowledge. In a second experiment, in which 50% of cases failed on only *one* condition, the rationale behind the neural network's performance appeared more sensible. Yet, "That the rationale is acceptable can again, however, be seen only from a standpoint of knowing what the rationale should be." (Bench-Capon T.J., 1993).

Being an artificial scenario with few, explicit conditions, this problem seems suited to be modelled through a CBR approach.

1.5.2 The recidivism score problem [1]

The algorithms employed in the United States for predicting the likelihood that a convict will recidivate are an example of how machine learning can have a significant impact on individuals. Several such algorithms exist: one of them is COMPAS, an algorithm which risk scores have been analysed in the work carried out by ProPublica in (ProPublica, 2016), and which is the main inspiration for this case study. These algorithms provide scores to determine the risk that a person will offend again in the future; in COMPAS' case, such a score is a number from 1 (low risk) to 10 (high risk).

By relinquishing control to the algorithm, with no grounds to support or challenge the score, a judge or an officer's decision could be liable to being mislead. This situation is an example of a likely scenario in which the user would benefit from being provided an explanation with the algorithm's output: it would be easier to determine the reliability of the score if the judge or the officer knew the arguments in support of it.

Unlike the social welfare benefit problem, in this case the rationale employed by these algorithms is unknown. Thus, the first challenge is defining this problem in terms of likely dimensions. In order to do so, a logistic regression model provided by doctor Ad Feelders (from Utrecht University) was used.

1.6 Conclusions

In this introduction a literature overview, as well as this research's objectives were provided. It was suggested that, in order to solve the problem of interpretable ML outputs, the first step is to gain some insight into human understanding. As explained beforehand, understanding is the ability to put together different pieces of knowledge, and explanations can ease that process. The way explanations provide understanding is by either increasing the person's body of knowledge, or by accounting for conflicts or contradictions with an existing knowledge base. This means that explanations are cognitively biased, and an optimal explanation takes into account the person's knowledge, or her expertise, as well as the context in which the explanation is provided. Both these aspects are fundamental for the issue of interpretability on the "computer side", too.

In general, it can be said that an output is interpretable if it is expressed in some way understandable to humans. For this reason it was asserted that transparency is not the solution for making interpretable outputs: transparency requires to reveal an algorithm's parameters, computations, as well as the data it processes, and not only it is problematic because the data can be legally sensitive, but such knowledge would hardly provide an explanation. Indeed, most people are not familiar with machine learning. Furthermore, humans are not transparent either, since explanations do not include a description of the neurological processes that yielded a decision; therefore, it does not seem reasonable to require such a standard from algorithms (Lipton Z.C., 2016).

[1] This case study was first presented in the internship paper (Friscione E., 2018). The original case study was modified and extended where needed.

Optimal explanations for ML outputs may vary from field to field, and here legal applications were considered. In particular, the question became how humans explain a legal decision. The use of arguments permeates the field of the law, and several examples of different types of reasoning were presented. Moreover, the field of AI applied to law is quite thriving, therefore case-based reasoning systems such as HYPO and CATO were investigated to understand whether they could provide a good starting point for creating a method of explanation for ML outputs.

Finally it was possible to formulate the current work's research questions, as well as the method pursued for answering them.

2 Notation, input facts, and dimensions

In this section the basic components for a model of explanation are going to be introduced. First of all the first research question is going to be answered, namely which tools provided by HYPO and CATO can be employed in generating explanations for ML outputs. Once this question is answered, the notation is going to be discussed more in detail, as well some limitations of the model chosen as a blueprint, and possible solutions to such issues.

2.1 Dimensions vs Factors

The first question that ought to be answered is which, between HYPO and CATO, offers the best set of tools for a model of explanation.

Although factors may be easier to implement computationally due to their boolean nature, it is acknowledged in the literature that factors are not sufficient to represent legal situations properly (Bench-Capon T.J. & Rissland E.L., 2001; Bench-Capon T.J., 2017). On the contrary, dimensions allow for a more nuanced representation. Factors are either present or absent, whereas dimensions support one side to different degrees. In (Bench-Capon T.J. & Rissland E.L., 2001) an example is made for the domain of trade-secret law: if the fact that some product information was divulged is treated as a dimension, then the higher the number of disclosures, the stronger the case will be for the defendant; if it is treated as a factor, a line must be drawn in order to determine whether the factor applies to the case (i.e., is one disclosure enough for the factor to apply?). This means that the case in which plaintiff made one disclosure will be treated like an instance in which fifty or hundreds of disclosures occurred. Furthermore, it is suggested that in an approach relying on dimensions factors are not necessarily lost. Indeed, factors can be seen as the extremes in a scale. Thus, whereas representing a situation with factors leads to a less detailed representation, dimensions are both fine-grained and can replace factors in those cases where intermediate positions are absent.

For these reasons, HYPO is chosen as the main inspiration for a model of explanation. In particular, this research's model will build on HYPO's dimensions, on its definition of on-pointness, as well as on its argumentation system.

2.2 Notation

Dimensions are here based on HYPO dimensions as introduced in (Ashley K.D., 1991). However, because the original notation seems to be unnecessarily complicated a simplified version is presented.

In the original HYPO a legal frame is used to represent the relevant facts of a given case, as well as aspects of the judicial decision if the case has been decided. The legal subframes are concerned with the details of a given case (e.g., information about a product or an employment agreement (Ashley K.D., 1991)). Then HYPO uses factual predicates, which are "generalized factual statements that confirm whether certain legally significant relationships are true in the case" (Ashley K.D., 1991). In order to confirm such legally significant relationships a retrieval method that checks the information contained in the legal (sub-)frames is used. A more thorough explanation can be found in (Ashley K.D., 1991).

The case studies here presented are quite simple in nature, therefore they do not require multiple levels of representation. There does not seem to be any advantage to using a main frame and subframe(s). Thus, a single, all-encompassing frame replaces the legal frame and its related subframes. Factual predicates are removed altogether, as it is unclear whether they are necessary to retrieve the relevant information contained in the legal frame. Subsequently, dimensions are adjusted in order to accommodate these changes.

An example that shows how cases are represented in the current framework is now in order. Let us start from the social welfare benefit problem.

```
Case-Number: 000000001
Citation: Decided application
Benefit-Awarded: No
Name: Mario-Rossi
Date-Of-Birth: 21/03/1946
Age: 72
Sex: Male
Marital-Status: Married
Domicile: XX, YY, UK
Capital-Amount: 5000£
Paid-Contributions: Yes
Patient-Status: In-patient
Hospital-Distance: 5 km
```

Figure 1. Representation of a case from the social welfare benefit dataset.

A submitted application is represented similarly. The frame includes all the information relevant for awarding the benefit, that is the person's age, sex, marital status, current address, capital resources, whether she paid four out of five of the last contribution years, the patient status of the spouse, and the hospital distance.

Cases from the COMPAS dataset are represented using the same logic.

```
Case number: 13011352CF10A
Citation: Decided score
Date: 13/01/2013
Recidivism score: 8 (high)
Name: Marcu-Brown
Age-Category: Less than 25
Sex: Male
Race: African-American
Juvenile-Felonies: No
Prior-Felonies: Yes
Date-Of-Arrest: 12/01/2013
Length of stay: 1 day
Charge-Degree: Felony
Type-Of-Charge: Possession of cannabis
```

Figure 2. Representation of a case from the COMPAS dataset.

The facts to include in the frame were decided in two steps. The COMPAS dataset contains the following facts that were deemed to be the most relevant in determining the recidivism risk: sex (`sex`), age (`age`), race (`race`), juvenile felony counts (`juv_fel_count`), juvenile misdemeanor counts (`juv_mis_count`), other juvenile counts (`juv_other_count`), priors count (`priors_count`), charge degree (`c_charge_degree`), type of charge (`c_charge_desc`). These elements are assigned either boolean values, numeric values, or string values.

Afterwards this list of relevant facts was compared against the list of factors that have a significant impact in predicting recidivism within two years in Ad Feelders' model. It would appear that among the most significant factors there are the charge degree (`c_charge_degree`), race (`race`), age category (`age_cat`), sex (`sex`), priors count (`priors_count`), and length of stay (`length_of_stay`). A few examples that show how this model works can be found in appendix A.

Hence prior (juvenile) misdemeanours were removed from the list, as well as other (juvenile) counts, and age was replaced with age category. `length_of_stay` was also introduced in the list.

This is the same information used to determine the dimensions' behaviour. Thus, rather than using factual predicates, for a dimension to apply it should directly query the main frame's slots through a retrieval method.

2.3 Dimensions for the social welfare benefit problem

The dimensions for the social welfare benefit problem can now be introduced. As stated beforehand, the same facts used to build the case frames were used to create dimensions. In some cases there is a one-on-one relation between input fact and the corresponding dimension; in other instances one or more input facts were used as sources to infer another dimension, as shown in the table below.

INPUT	DIMENSION
age sex	PensionableAge
paid	PaidContributions
married	MaritalStatus
absent	Absent
capital	CapitalResources
in-patient distance	HospitalDistance

Table 1. Relations between input facts and dimensions in the social welfare benefit problem.

Now it is possible to look into the details of the dimensions. It can be noted that, unlike HYPO dimensions as presented in (Ashley K.D., 1991), the following dimensions do not contain a pro or con direction (in HYPO's case, a pro-defendant or a pro-plaintiff direction). This issue will be discussed in a later section. Furthermore, the comparison type slot is also left out.

Dimension name: PensionableAge
Prerequisites: Applicant is of pensionable age.
Pensionable age is 65 for men.
Pensionable age is 60 for women.
Focal slot(s): Sex, Age
Range: [0, ..., 100]

Dimension name: Absent
Prerequisites: Applicant lives in the UK.
Focal slot(s): Domicile
Range: Yes/No

Dimension name: MaritalStatus
Prerequisites: The patient is the spouse of the applicant.
Focal slot(s): Marital-Status
Range: Yes/No

Dimension name: HospitalDistance
Prerequisites: Spouse is a in- or out-patient.
Focal slot(s): Patient-Status
Range: $x \geq 0$

Dimension name: CapitalResources
Prerequisites: Applicant's capital resources tantamount to no more than 3000£
Focal slot(s): Capital-Amount
Range: $x \geq 0$

Dimension name: PaidContributions
Prerequisites: Applicant has paid four out of five of the last contribution years.
Focal slot(s): Paid-Contributions
Range: Yes/No

Figure 3. Dimensions for the social welfare benefit problem.

The *prerequisites* are the conditions necessary for the dimension to apply. If the dimension applies, the *focal slot(s)* report where in the case frame the relevant information can be found. Something about the prerequisites should be noted for the current case studies: whereas in the original HYPO some dimensions can apply to a case while others do not, in the current case studies every fact situation will contain the inputs relevant for every dimension to apply. In both the social welfare benefit problem and the recidivism score problem every case is complete, that is, there is no missing information, and every case contains the same information as the other ones. As such, the use of prerequisites in the current case studies is not strictly necessary: clearly every dimension will apply to the case. Nevertheless, prerequisites will be kept for future studies in which cases either lack information or the information varies from case to case.

2.4 Dimensions for the recidivism score problem

The same reasoning was applied to the recidivism score problem. The table below shows how the dimensions relate to the input facts.

INPUT	**DIMENSION**
`age_cat`	AgeCat
`sex`	Sex
`juv_fel_count`	JuvenilePriors
`priors_count`	PriorsCount
`c_charge_degree`	ChargeDegree
`c_charge_desc`	Offence
`length_of_stay`	LengthOfStay
`race`	Race

Table 2. Relations between input facts and dimensions in the recidivism score problem.

The dimension `Offence` accounts mostly for felonies, but a separate dimension could be made for misdemeanors. `PriorsCount` and `JuvenilePriors` account for the defendant's prior counts, although Feelders' model seems to point out that juvenile felonies are not as important in determining the likelihood an individual will recidivate; the model only uses priors count to predict the likelihood.

Dimension name: AgeCat
Prerequisites: There is a defendant
Defendant was arrested
Defendant was charged
Focal slot: Age-Category
Range: [0, ..., 100]

Dimension name: Sex
Prerequisites: There is a defendant
Defendant was arrested
Defendant was charged
Focal slot: Sex
Range: [M;F]

Dimension name: Offence
Prerequisites: There is a defendant
Defendant was arrested
Defendant was charged
Focal slot: Type-Of-Charge
Range: [Murder, ..., Theft,..., Failure to appear in court]

Dimension name: PriorsCount
Prerequisites: Defendant committed an offence in the past
Focal slot: Prior-Felonies
Range: [0, ..., 38]

Dimension name: LengthOfStay
Prerequisites: Defendant was jailed
Focal slot: Length-Of-Stay
Range: [0, ..., 800]

Dimension name: JuvenilePriors
Prerequisites: Defendant has juvenile prior counts
Focal slot: Juvenile-felonies
Range: Yes/No

Dimension name: ChargeDegree
Prerequisites: Defendant was charged
Focal slot: Charge-Degree
Range: [M;F]

Dimension name: Race
Prerequisites: There is a defendant
Defendant was arrested
Defendant was charged
Focal slot: Race
Range: [Caucasian, African-American, Native American, Hispanic, Asian, Other]

Figure 4. Dimensions for the recidivism score problem.

The range of the dimension `Offence` includes a list of crimes. The main issue would be ordering (or partially ordering) this list; whereas failure to appear in court is clearly less severe than murder, comparing felonies such as arson and robbery might be more complicated. Perhaps a first criteria would be dividing between crimes against the person and crimes against property. For now this dimension is mainly used to retrieve cases in which the same felony was committed: it appears a sensible choice to compare cases with the same type of offence. Finally, the dimensions `PriorsCount` and `JuvenilePriors` must be assigned a numeric range and a boolean range respectively, because the dataset only reports the defendant's number of priors and whether she has some juvenile prior.

2.5 Pro, con, and neutral dimensions

In the dimensions listed thus far the direction has not been determined yet. As explained in (Ashley K.D., 1991), a dimension's direction shows which side of the scale favours the plaintiff, versus which side favours the defendant. The

reason why direction has been left out is because it was not always clear, in the current case studies, when and how the dimension favoured either the social benefit applicant or the defendant in the recidivism score problem. This is possibly due to the fact that HYPO has a few limitations, one of which, namely the fact that there is no interaction between dimensions, is discussed in this work.

Thus, how can a dimension's direction be chosen in these case studies? For the social welfare benefit problem the dataset was downsized and inspected in order to determine the behaviour of the artificial neural network (ANN) employed in the original experiments. Furthermore, the criteria for applying to the benefit were used. For what concerns the recidivism score problem Ad Feelders' model, as well as ProPublica's analysis, were used to pinpoint the various dimensions' direction.

2.5.1 Dimensions' direction in the social welfare benefit problem

In inspecting the social welfare benefit dataset, the following steps were first taken in order to obtain a more manageable dataset: first the variables of interest were isolated, as the dataset contained several noise variables that were irrelevant for the CBR system (but that might be relevant for the ANN's output, as explained in a later section of this paper). Secondly, all the cases in which the variable "married" was assigned value 0 were removed, since the benefit is supposedly awarded only to the patients' spouses. Finally, any application in which the age was below 60 was removed as well, since the benefit is meant for pensioners.

Once downsized, the following facts were noted:

- The conjunction of paid = 0 and capital > 3000 always yielded "not qualified" as an output;
- The conjunction of paid = 1 and capital > 3000 always yielded "not qualified" as an output;
- Absence = 1 always yields "not qualified" as an output;
- The conjunction of in-patient = 1 and distance > 49 always yields "not qualified" as an output;
- If paid = 0 and capital < 3000, then the ANN can yield "qualified" as an output.

This means that the cap on capital resources is a hard requirement, and no case exists in which someone whose capital amounts to more than 3000£ was awarded the benefit. Absence is also such a hard requirement: for instance, if a person's capital is below 3000£ but she does not currently live in the UK, she will automatically be denied the benefit. The same applies to the relation between the patient's status and hospital's distance: if someone whose resources are below 3000£ does not meet the correct criteria, then they will not be eligible for the benefit.

DIMENSION	DIRECTION		
	Pro	Neutral	Con
PensionableAge	x $\geq$ 65 (M)	-	x $<$ 65
	x $\geq$ 60 (F)	-	x $<$ 60
PaidContributions	Yes	No	
MaritalStatus	Yes	-	No
Absent	No	-	Yes
CapitalResources	x $\leq$ 3000£	-	x $>$ 3000£
HospitalDistance	x $<$ 50 (IN)	-	x $\geq$ 50
	x $\geq$ 50 (OUT)	-	x $<$ 50

Table 3. Direction of dimensions in the social welfare benefit problem.

The pro direction intuitively represents the cases in which the applicant satisfies the welfare benefit requirements; likewise, the con direction accounts for those cases in which the candidate fails to meet the criteria. The neutral direction roughly corresponds to those thresholds within a dimension that are neither pro-plaintiff nor pro-defendant, as explained in (Ashley K.D., 1991). Here, however, a case will be made to introduce entirely neutral dimensions which never have a pro or con direction per se, but rather they may favour one side over another depending on the context given by either other dimensions, or additional information.

A more obvious case for neutral dimensions is found in the recidivism score problem.

2.5.2 Dimensions' direction in the recidivism score problem

In order to determine the dimensions' direction in the recidivism score problem, the model provided by Ad Feelders was used. In this model logistic regression is used to determine the likelihood the defendant will recidivate. Thus, coefficients were examined to determine the contribution of each input fact to the prediction.

- c_charge_degreeM yields lower scores compared to c_charge_degreeF (-0.1997430).
- Compared to raceAfrican American, raceAsian (-0.6835894), raceCaucasian (-0.0908918), raceHispanic (-0.2634037), raceNative American (-0.33386200), and raceOther (-0.2434707) all yield lower scores.
- Compared to age_cat 25-45, age_cat Greater than 45 yields lower scores (-0.6774594), whereas age_cat Less than 25 yields higher scores (0.7306911).
- Males receive higher scores than females (0.3408185).
- Having priors count contributes higher scores (0.1613732).
- lenght_of_stay also contributes to higher scores (0.0026240).

However, this model slightly differs from the analysis by ProPublica. According to ProPublica (ProPublica, 2016), women are 19.4% more likely to receive higher scores than men. For this reason, I placed both Male and Female in the neutral range for the dimension `Sex`. Furthermore I proposed some dimensions, such as `JuvenilePriors` and `Offence`, that do not figure in Ad Feelders' model but that intuitively should appear in an explanation. Although intuitions might differ, if asked why someone is a high recidivism risk it seems sensible that a possible given reason is the fact that this person already committed felonies in the past.

It must also be noted that the thresholds reported in the following table were chosen arbitrarily, and different design choices were possible.

	DIRECTION		
DIMENSION	Pro	Neutral	Con
AgeCat	$x \geq 45$	$25 < x < 45$	$x \leq 25$
Sex	-	M/F	-
JuvenilePriors	No	-	Yes
PriorsCount	0	[1, 2, 3]	$x \geq 4$
ChargeDegree	-	M	F
Offence	-	-	[Range list]
LengthOfStay	$0 < x \leq 50$	$50 < x \leq 500$	$x > 500$
Race	-	Asian, Native American Hispanic, Other, Caucasian	African-American

Table 4. Direction of dimensions in the recidivism score problem.

The pro direction indicates that the defendant should get a low recidivism score, whereas the con direction shows she should get a high one. In this case study we have instances of boolean dimensions, that is, dimensions that only have extreme ranges (e.g. `JuvenilePriors`), as well neutral dimensions (e.g. `Sex`). Indeed, this case study highlights how the concept of dimension as introduced in (Ashley K.D., 1991) may not suffice and thus requires refinement. The case of boolean dimensions seem pretty straightforward. In the following section the notion of neutral dimensions is introduced.

2.5.3 Extending HYPO with neutral dimensions

It was mentioned beforehand that, aside from the classic pro and con directions (or pro-plaintiff and pro-defendant), a "neutral" instance was included. This neutral area might to some extent correspond to those ranges within a dimension in which neither side is clearly favoured. However, the dimensions for the recidivism score problem open up a new possibility. Indeed, in this case study it cannot be said that any of those dimensions obviously favour the defendant, so much so that a few of them lack of a pro direction altogether (i.e. `Offence`). However, there are dimensions such as `Sex` that are neither pro, nor con the defendant. This means that, aside from the classic dimensions with a pro and con range, the HYPO model for these case studies should be extended to include:

- **Neutral dimensions:** dimensions which range does not span from pro-defendant to con-defendant, but rather offer ulterior reasons why an outcome was decided in a certain way. Whereas the facts covered by these dimensions ultimately favour or not the defendant might depend on the context, or on some other information.

- **Con dimensions:** the recidivism score problem shows how there are situations in which several facts are used against the defendant no matter what. What changes is the *gravity* of the fact. Thus having a criminal record cannot result in a fact situation that favours the defendant, however less serious crimes should not weigh as heavily on the score as crimes such as murder or kidnapping.

- **Pro dimensions:** although this is not the case with the current case studies, if con dimensions are possible then it is likely that there might be fact situations that favour one side no matter what, and the strength of this support depends on where the case's facts fall within the dimension's range.

Thus it has to be defined what is meant with "context" or "ulterior information". This aspect is likely tied to a research questioned addressed in this work's original research proposal, that is, *given these tools* [the ones provided by the HYPO model] *is the data obtained from machine learning algorithms enough, or should information be added to it?* If some dimensions are context-dependent, then at first glance it would appear that we need more information: the datasets used in this research mostly provide quantitative values, which are, at times, already non satisfactory (such as the case of the juvenile prior counts providing only a number, rather than a more detailed list of such priors).

If neutral aspects in a dimension are context-dependent then a hypothesis can be made that the mere data is not enough in order to produce a satisfactory explanation. This context may be given by the values that this dimension might promote or demote (and these values have to be added to the data). Alternatively, the context may be determined on the basis of the other dimensions. Let us take the social welfare benefit problem as an example. If a person paid only 3 of the past contribution years, but their capital resources are well below 3000£ in total, then ultimately it can be argued that the person should still be awarded the benefit because of a lack of resources to pay the four contribution years required by the application. This means that dimensions are not separate, monad-like entities, but rather it should be possible to infer information about a dimension on the basis of the other dimensions.

2.6 The problem of context

In the previous section the issue of context was introduced; in other words, the fact that extra information may be required in order to determine where a case's fact fall within a dimension's range. It was suggested that a dimension might depend on other dimensions, that is, that dimensions are potentially (co-)

dependent. In the original HYPO each dimension is a separate entity, therefore the tweaked dimensions introduced beforehand need further adjustments in order to account for dependence. It must be noted that at present the relationship between dimensions could only be inferred for the social welfare benefit problem. Due to the nature of the model employed for the recidivism score problem it is not possible to establish how the input facts, and therefore the corresponding dimensions, interact with one another. The logistic regression model only showed how each input fact contributed to the score, but it was not as informative for what concerned the relationship between the different inputs themselves.

2.6.1 Dimensions' dependencies in the social welfare benefit problem

In the social welfare benefit problem the issue of context can be solved through knowledge engineering by making explicit the possible relationships between dimensions. Let us use again the very simple example of the relationship between the dimensions `CapitalResources` and `PaidContributions`: if a person has not paid the required number of contributions it might be due to the fact that her capital resources are too scarce. A rule to express this dependence can be semi-formally expressed as:

If `PaidContributions` = No,
and `CapitalResources` $\leq 3000£$,
then `PaidContributions` *is pro applicant.*

The idea is that, if in a case situation a fact falls within the neutral direction, then further information can be sought in the relevant, related dimensions.

Hence the dimensions for the social welfare benefit problem can be changed as follows:

Dimension name: PensionableAge
Prerequisites: Applicant is of pensionable age.
Pensionable age is 65 for men.
Pensionable age is 60 for women.
Focal slot(s): Sex, Age
Range: [0, ..., 100]
Direction: pro [x $\geq$ 65 (M); x $\geq$ 60 (F)]; con [x $<$ 65 (M); x $<$ 60 (F)]
Related dimensions: N/A

Dimension name: Absent
Prerequisites: Applicant lives in the UK.
Focal slot(s): Domicile
Range: Yes/No
Direction: pro: Yes; con: No
Related dimensions: N/A

Dimension name: MaritalStatus
Prerequisites: The patient is the spouse of the applicant.
Focal slot(s): Marital-Status
Range: Yes/No
Direction: pro: Yes, con: No
Related dimensions: N/A

Dimension name: HospitalDistance
Prerequisites: Spouse is a in- or out-patient.
Focal slot(s): Patient-Status
Range: $x \geq 0$
Direction: (IN) pro: x $<$ 50 x; con: x $\geq$ 50
OUT pro: x $\geq$ 50; con: x $<$ 50
Related dimensions: N/A

Dimension name: CapitalResources
Prerequisites: Applicant's capital resources tantamount to no more than 3000£
Focal slot(s): Capital-Amount
Range: $x \geq 0$
Direction: pro: x $\leq$ 3000£; con: x $>$ 3000£
Related dimensions: N/A

Dimension name: PaidContributions
Prerequisites: Applicant has paid four out of five of the last contribution years.
Focal slot(s): Paid-Contributions
Range: Yes/No
Direction: pro: Yes; neutral: No.
Related dimensions: CapitalResources
Rule: *If* `PaidContributions` = No,
and `CapitalResources` $\leq$ 3000£,
then `PaidContributions` *is pro applicant.*

Figure 5. New dimensions for the social welfare benefit problem.

However, it can be seen that the problem of context cannot be solved by only making explicit the relation between dimensions. For example, it can be imagined that other dimensions, such as `HospitalDistance` and `CapitalResources` may not be hard requirements and therefore have neutral ranges. However, the information required to determine how they'll weigh on the final decision must be determined differently. For example, in the case of `HospitalDistance` the number of hospitals and how they are allocated throughout the territory must be known: a case might exist in which an out-patient is staying at a hospital which is not in the required range. This may be due to the fact that the next closest hospital is too far to be a feasible option.

2.7 Conclusions

In this section the first research question was answered, namely which among the tools provided by case-based reasoning systems best suit a method of explanation. HYPO was chosen over CATO because dimensions seem more apt at representing problem situations in finer details. Hence, using HYPO as a starting point a simplified notation was introduced, and the social welfare benefit and

COMPAS datasets were used to infer dimensions, the most basic components of the new model.

In inferring the new dimensions, however, a few limitations of the original HYPO were also brought to light. For instance, it was shown that a few dimensions' directions were not trivially for and against one of two parties, but the possibility arose that some dimensions may favour (or not favour) a side no matter what, or that it does not favour any side at all unless additional information is provided. For this reason it was suggested that the original concept of dimension should be expanded. Lastly, it was also suggested that for so-called neutral dimensions it should be possible to query other dimensions in order to obtain further information.

3 The model

In the following sections a model for generating explanations for ML outputs is more systematically presented. Firstly an overview of the model is introduced; then the case retrieval and argument generation processes are going to be explored more in detail.

3.1 Model overview

Now that all the main components have been introduced, a model can be defined. As before, HYPO was used as the blueprint for the simplest model. Given some input facts, the model analyses them dimensionally and compares the current case with the cases in its case knowledge base (CKB).

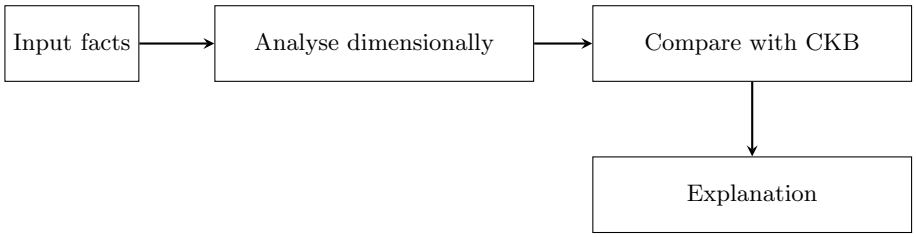

Figure 6. Representation of the basic model.

The final decision should be a binary output: either awarded or not awarded for the social welfare benefit problem, and low score or high score for the recidivism score problem. In the latter case, the scores that were originally organised in low, medium, and high categories are reduced to just two categories, low (1 to 5) and high (6 to 10).

The output must be provided with an explanation, which should take into account the cognitive biases a person brings to understanding, as well as her background knowledge, cognitive abilities, and possible time constraints.

For these reasons the provided explanation can be of two types: one for average users, which outlines the main points of the current case and explains

the decision with natural language; a second one for more expert users in the form of arguments, which not only highlights the main issues, but it also points to similar (and different) cases contained in the case knowledge base. The first type is called Natural Language Explanation (NLX), the second type Argument Explanation (AX).

The model can be further refined as follows:

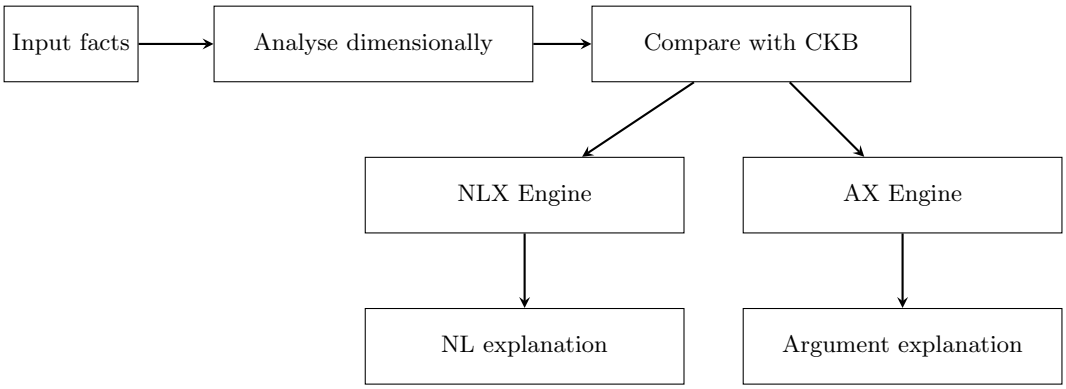

Figure 7. Representation of the basic model.

3.2 Generating explanations

In order to generate explanations a problem situation's input facts are analysed dimensionally, and then compared to the cases contained in the CKB. Among these cases, the model selects the ones similar to the current fact situation *cfs* on the basis of a definition of on-pointness introduced in the following sections. The model also retrieves the remaining similar cases, and generates a claim lattice which root node contains the cfs and similar cases, and the leaves nodes contain the remaining cases. The further from the root node a case is, the less similar it is to cfs. Afterwards, NLX and AX are generated. NLX are based on templates in which the relevant information has to be filled in by the NLX engine. However, this work mostly focuses on AX. An AX consists of an argument game inspired by 3-ply arguments as introduced in (Ashley K.D., 1991). NLX are natural language outputs, whereas an AX is a graph representing the arguments each side can make, and how these argument attack each other.

3.2.1 Case retrieval

Cases are indexed and retrieved through dimensions, which are also the basis for comparison between the current fact situation *cfs* and the cases in the case knowledge base. Until now an intuitive definition of dimension has been used. A formal definition is here given.

Definition 2. [Dimension] A **dimension** d is a triple $(\mathcal{V}, \leqslant, \mathcal{R})$ where:

- $\mathcal{V}$ is a dimension's range of values;
- $\leqslant$ is a (partial) ordering of these values;
- $\mathcal{R}$ is a function $\mathcal{R} : \mathcal{V} \longrightarrow \{pro, neutral, con\}$ that, for every value, returns either a direction or no direction.

Given a set of dimension-value pairs DV and a set of possible outcomes, $\mathcal{O} = \{\pi, \delta, undecided\}$, a formal definition of *case* is obtained.

Definition 3. [Case] A **case** c is a pair $c = (DV, o)$, where DV is a set of dimension-value pairs, and $o \in \mathcal{O}$. *O(c)* denotes a case's outcome. In the current fact situation *O(c) = undecided*.

A dimension applies to a case when certain input facts are present in a case. In HYPO a retrieval method is used to check whether a dimension's prerequisites are satisfied. As explained in section 2.3, in the current case studies every case is complete, meaning that all the cases contain a value for each dimension, therefore every dimension will always apply to every case. However, in less well-defined scenarios not every dimension applies to every case.

Hence a definition of *similarity* or *on-pointness* is in order. Let us denote the dimensions that apply to a case as *D(c)*. When a dimension applies to a case it assumes a specific value, denoted as $v(d, c)$, and direction, denoted as $r(d, c)$.

Definition 4. [On-pointness (HYPO)] A case c_1 is similar to *cfs* if $D(cfs) \cap D(c_1) \neq \emptyset$. Let us denote this set of similarities as $S_1(cfs, c_1)$. If $S_2(cfs, c_2) \subseteq S_1(cfs, c_1)$, then we say that c_1 is more similar, or more on point than c_2, to *cfs*.

Here, however, this definition has to be further refined. Let us denote with DV(c) the set of dimension-value pairs that apply to a case, and with DR(c) the set of dimension-direction pairs that apply to a case.

Definition 5. [On-pointness] A case c_1 is similar to *cfs* if $DR(cfs) \cap DR(c_1) \neq \emptyset$. Let us denote this set of similarities as $S_1(cfs, c_1)$. If $S_2(cfs, c_2) \subseteq S_1(cfs, c_1)$, then we say that c_1 is more similar, or more on point than c_2, to *cfs*.

The new definition of on-pointness is an improvement of the original HYPO definition, as it imposes further constraints based on the assumption that a dimension's direction is vital in determining how a certain fact favours (if at all) one side of the dispute.

Let us now assume A stands for the welfare's applicant or the defendant in the recidivism score problem, and let us denote the set of dimension-direction pairs that apply to A in the *cfs* as $DR_A(cfs)$.

Definition 6. [Citable cases] A precedent c_1 is citable for A if $O(c_1) = A$ and $DR_A(cfs) \cap DR(c_1) \neq \emptyset$, where $DR_A(cfs)$ and $DR(c_1)$ are the set of dimension-direction pairs that apply to the current fact situation and the cited case respectively, and that are pro A.

Definition 7. [Best case to cite] A precedent c_1 is a best case to cite for A if $O(c_1) = A$, and c_1 is at least as similar to *cfs* as any other case that is citable for A.

Let us now illustrate the notion of on-pointness employed in this work with an example. Dimensions $\mathcal{D} = (d_1, d_2, d_3)$ apply to the problem situation, and the following cases are given:

- $c_a = ([d_1, pro, v_a], \pi)$
- $c_b = ([d_1, pro, v_b], [d_2, pro, v_{b2}], \pi)$
- $c_d = ([d_2, con, v_d], [d_3, pro, v_{d2}], \delta)$
- $cfs = ([d_2, pro, v_e], ?)$

The most similar case to the current fact situation is c_b: the two cases share one dimension, namely d_2, as well as the dimension's direction (pro); it could also be the case that the dimension's specific value is the same in both instances. The second most similar case is c_d: it shares dimension d_2 with cfs, however they do not share the same direction. Finally, c_a has nothing in common with the current fact situation.

Once the most similar cases are retrieved, it is possible to represent them with a Claim Lattice, as in (Ashley K.D., 1991).

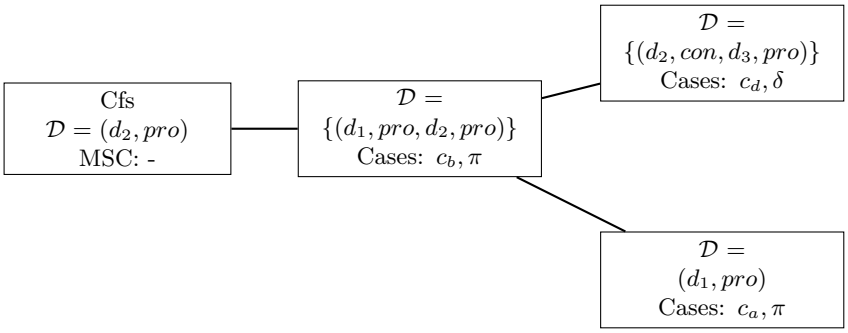

Figure 8. Cases in the root node are the most similar cases to cfs, i.e., those that share dimensions and direction. Cases in leaf nodes are those with shared dimensions and different directions. Next to each case is reported which side was favoured for that case. The further away from the root node, the less similar is the case to cfs.

3.2.2 Argument generation

Claim lattices are the basis for generating both NLX and AX. This work is going to focus on AX, and a general outline is given for NLX.

The arguments generated for the AX are based on 3-ply arguments as presented in (Ashley K.D., 1991) as well as argument games as introduced in (Prakken H., 2017). First the kinds of arguments that can be generated, as well as how arguments can defeat each other will be defined.

Definition 8. [Arguments and attack relations] An argument can be either a citation or a distinction. A citation points to a case in the CKB that is citable for the side who cites it, whereas distinctions point to differences between a cited case and the cfs that make the cited case weaker for the side who cites it. In this model distinctions point to single differences, but in a mode complex model they could point to set of differences. Given a current fact situation *cfs*:

- A citation attacks another by citing another similar case with an opposite outcome;
- A distinction points to DR pairs which are present in the cited case and that favours the side who cites it, but are absent in the *cfs*;
- A distinction points to DR pairs that favour the other side which are present in the *cfs*, but are absent in the cited case;
- A distinction points to differences between DV pairs in the cfs and DV pairs in cited case that favours the side who cites it, arguing that the cited case's specific value yielded a determined outcome.

More formally, a distinction can be:

- A pair (C_π, DR_π), where C_π is a citable case for the side that favours it, and DR_π is a dimension-direction pair that is present in C_π, but it is absent in the *cfs*;
- A pair (C_π, DR_δ), where C_π is a citable case for the side that favours it, and DR_δ is a dimension-direction pair that is absent in C_π, but it is present in the *cfs*;
- A pair $\{v_1(d_1, c_1), v_2(d_1, cfs)\}$, where $v_1(d_1, c_1)$ is the value of the dimension d_1 in the cited case, and $v_2(d_1, cfs)$ is the value of the dimension d_1 in *cfs*. If, taken the direction of d_1 into account, v_1 is "more pro" a certain side compared to v_2, then that dimension favours less the side who cited the case.

Furthermore, the notion of downplaying a distinction is introduced. This notion, however, is proposed here informally and it is suggested it should be developed for future research.

- A distinction can be downplayed by pointing to DV pairs in a case c from the CKB that weaken the distinction made by either player. This is achieved by showing that differences between values are not relevant, as cases with different values still yield the same outcome.
- Likewise, a distinction can be downplayed by pointing to DR pairs in a case c from the CKB that show how differences between directions are not relevant, as cases with different directions still yield the same outcome.

Now a notion of argument ordering must be introduced.

Definition 9. [Argument Ordering] Given a partial ordering on arguments $\leqslant$, the notation A $\leqslant$ B means that argument A is preferred over argument B; in other words, argument A is more similar to *cfs* than argument B. This holds for citation-type arguments. In this work, given two equally similar citations A and B, A can be weakly preferred over B. Given a citation A and a distinction B, the distinction B is strictly preferred over the citation. Finally, an argument that downplays a distinction is also strictly preferred over the distinction.

Thus the following defeat relations can be defined:

Definition 10. [Defeat relations] Given a partial ordering on arguments, an argument A defeats an argument B if and only if A attacks B and B is not strictly preferred over A.

Hence the argument game can be described (Modgil S. & Caminada M., 2009). In an argument game there is a proponent, P, who starts by moving a citation, and an opponent, O, who is allowed to attack said argument using either a citation or a distinction according to the game's rules.

Definition 11.[Argument game] An argument game is a tree which branches are disputes, i.e. sequences of legal moves where the possible moves are citations and distinctions. The game is played by two players, a proponent P and an opponent O. Then, given an $AF = \{Args,\ defeat\}$, the set M of moves consists of all pairs *(p, A)* such that $p \in \{P, O\}$ and $A \in Args$.

Moves can be made according to the rules in Definition 12. Here the simplest game is going to be used, namely the argument game for grounded semantics (G-game). If an argument is defeasibly provable in the G-game, then the CBR and ML outputs agree: given a ML output, P makes the first move by citing a case in the CKB to support said outcome. The games' rules are the following (Prakken H., 2017):

Definition 12. [G-game rules]

- If a player cites a case, then it must be a best case to cite for the player;

- P's first move is a citation;
- Each move must defeat the previous move;
- P cannot repeat her moves;
- P only moves strict defeaters;
- Each move must be relevant to the previous one, i.e. only the directions or the values used in the previous argument should be attacked;
- A player wins if and only if the other cannot move.

Definition 13. [Grounded extension] In order to determine whether an argument is justified, overruled, or defensible a labelling approach will be used. In a grounded extension an argument is justified if it is labelled *in*, which means that all the arguments defeating it are labelled *out*, if any. An argument is overruled (*out*) if it is defeated by a justified argument; an argument is defensible if it is neither justified nor overruled.

3.3 Conclusions

In the previous section a model for generating explanations in the form of arguments, based on HYPO, was presented. It was proposed that the model should yield two kinds of explanations, one in natural language (NLX), and one as an argument game (AX). The first is meant for non-expert users, whereas the latter, which was the main focus of this research, is meant for users possessing prior knowledge of the problem. Formal definitions describing these processes were also introduced.

4 Examples

In the following two sections examples of how the current model works are presented. Justified arguments are coloured green, overruled arguments are coloured red, whereas defeasible arguments are left white.

4.1 The social welfare benefit problem

4.1.1 Methods

As explained in section 4.1, the original dataset was downsized in order to make it manageable for a human. Cases of interest were manually chosen to be included in the social welfare benefit CKB. In particular, attention was given to those cases in which the conjunction of PaidContributions = No and CapitalResources < 3000 applied and that were decided against the applicant by the ANN, as well as cases in which the condition for HospitalDistance was satisfied, but that were nonetheless decided against the applicant by the ANN. A total of 31 cases was selected.

4.1.2 Examples

The first is case 5 from appendix B. First of all, the problem is analysed dimensionally, and then the similar cases are organised in a claim lattice.

ANALYSIS		
`PensionableAge`	x $\geq$ 65	pro
`Absent`	No	pro
`PaidContributions`	No	Neutral $\rightarrow$ CapitalResources $\leq$ 3000£ $\rightarrow$ pro
`CapitalResources`	x $\leq$ 3000£	pro
`HospitalDistance`	x $<$ 50 km	pro

Table 5. Dimensional analysis of the input facts from case 5.

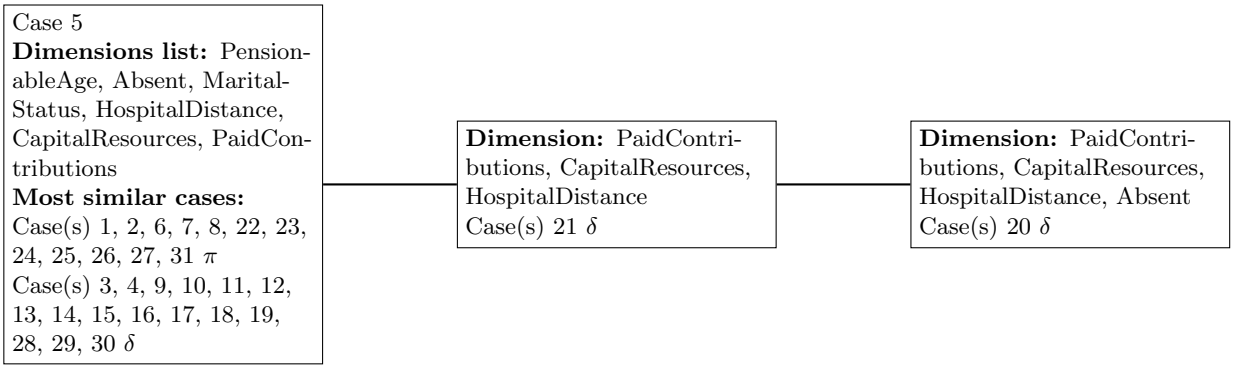

Figure 9. Claim lattice for case 5. It must be noted that this claim lattice holds for the other examples as well.

The AX engine creates an argument game for the claim.

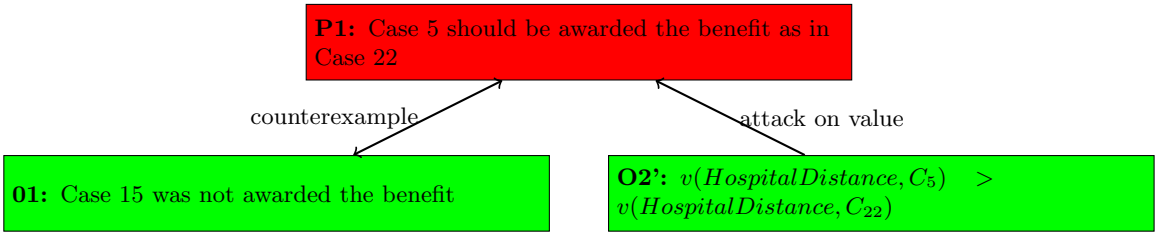

Figure 10. Argument game for case 5.

It can be seen that the argument in favour of awarding the benefit in case 5 is not justified in the G-game, meaning that the CBR system disagrees with the ML output. However, this only holds for the partial database that was selected as a case knowledge base; if the whole dataset was considered, then the

argument game could potentially proceed, as well as generate more branches. Furthermore, due the inconsistencies in the dataset these results are in no way conclusive, they are only illustrations of how the model should work.

Let us pick another case, namely case 3 from appendix B.

ANALYSIS		
`PensionableAge`	$x \geq 65$	pro
`Absent`	No	pro
`PaidContributions`	No	Neutral $\rightarrow$ CapitalResources $\leq 3000£$ $\rightarrow$ pro
`CapitalResources`	$x \leq 3000£$	pro
`HospitalDistance`	$x < 50$ km	pro

Table 6. Dimensional analysis of input facts from case 3.

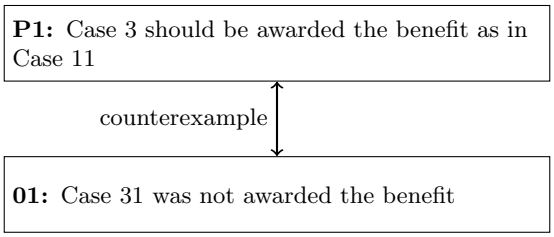

Figure 11. Argument game for case 3.

Interestingly, the argument for awarding the benefit in case 3 is undecided. In a grounded extension, this means that the argument is only defensible. This raises the issue of which outcome should be preferred, that is, whether the CBR should follow the ANN's outcome, or whether it should challenge it. A preference ordering might be necessary in order to sort out these issues: either a preference between arguments, or even a preference between the ML algorithm and the CBR system. For example, if the rule that "applicants whose capital resources are below 3000£ and have not paid the last four contribution years should be assigned the benefit if other conditions apply" was explicitly required to be enforced, then the ANN outcome in this case should be overruled.

A couple of things are worth noting: first of all, the citation P uses in her first move is a *most similar case*, namely a case that is the closest to *cfs* both for dimension directions and dimension values. This choice is based on the assumption that choosing the most similar case that favours the side who cites it is, to put it plainly, the smartest move to avoid unnecessary attacks. Secondly, for what concerns the attack themselves, in order to prevent games from becoming too long an intuitive notion of relevance was employed. For example, if O attacks a cited case's value, then P should counter-attack that specific value.

Let us consider another interesting example, such as case 9 from appendix B. Again, the same claim lattice holds in this case.

ANALYSIS		
PensionableAge	x $\geq$ 65	pro
Absent	No	pro
PaidContributions	No	Neutral $\rightarrow$ CapitalResources $\leq$ 3000£ $\rightarrow$ pro
CapitalResources	x $\leq$ 3000£	pro
HospitalDistance	x $>$ 50 km	pro

Table 7. Dimensional analysis of input facts from case 9.

And the resulting argument.

P1: Case 9 should not be awarded the benefit as in Case 30

Figure 12. Argument game for case 9.

This is an interesting case because the argument against awarding the benefit in this case cannot be attacked in any way. The reason is that, for what concerns case 9 and case 30, the value $v(CapitalResources, C_9)$ is identical to value $v(CapitalResources, C_{30})$, and $v(HospitalDistance, C_9)$ is nearly identical but lower than $v(HospitalDistance, C_{30})$. No other identical case with opposite outcome was present in the dataset's excerpt. Thus no specific values or directions can be attacked in this argument game.

Finally let us consider a case that flagrantly fails to satisfy the criteria for receiving the benefit, such as case 20 from appendix B.

ANALYSIS		
PensionableAge	x $\geq$ 65	pro
Absent	Yes	con
PaidContributions	No	Neutral $\rightarrow$ CapitalResources $>$ 3000£ $\rightarrow$ con
CapitalResources	x $>$ 3000£	con
HospitalDistance	x $\geq$ 50 km	con

Table 8. Dimensional analysis of input facts from case 20.

When retrieving similar cases, the AX engine will be unable to find any case that, given these dimensions, directions, and values, were decided in favour of the applicant. In this situation the generation of a claim lattice and argument game seems pointless: the only argument that can be made is that the applicant was denied the benefit because he fails to meet the benefit's requirements.

What about cases that meet all the relevant criteria? Similarly, it would seem that only an argument could be made that the applicant should be awarded the benefit on the basis that she satisfies all requirements. However, cases such as case 28, case 29, and case 30 from appendix B are decided against the applicant despite the fact that these people should be awarded the benefit. This means that, ideally, an argument game could be made against a "perfect" case such as case 29. The issue could be solved by making the dataset consistent and removing those cases that should have clearly been decided in favour of awarding the benefit. However this highlights an important issue, that is, the differences between the definition of similarity used in a CBR system and that used in a ML algorithm such as ANNs. This problem is further elaborated in the discussion section.

4.1.3 Results

Out of the 31 cases selected for the social welfare benefit CKB, 20 of them were used to generate argument games and compare the CBR output against the ANN output.

Case	ANN output	CBR output
1	Qualified	Undecided
2	Qualified	Not Qualified
3	Not Qualified	Undecided
4	Not Qualified	Undecided
5	Qualified	Not Qualified
6	Qualified	Undecided
7	Qualified	Not Qualified
8	Qualified	Not Qualified
9	Not Qualified	Not Qualified
10	Not Qualified	Qualified
11	Not Qualified	Undecided
12	Not Qualified	Qualified
13	Not Qualified	Qualified
14	Not Qualified	Undecided
15	Not Qualified	Qualified
16	Not Qualified	Qualified
17	Not Qualified	Not Qualified
18	Not Qualified	Qualified
19	Not Qualified	Not Qualified
20	Not Qualified	Not Qualified

Table 9. List of ANN outputs and CBR outputs.

Out of 20 cases, in only four cases the CBR output corresponds to the ANN output. It must be noted that, in a previous trial run in which a non-formal notion of downplay was used, half of the CBR outputs corresponded to the ANN outputs. The cases in which the CBR yielded "Undecided" as an output can

be viewed as cases in which the applicant failed to meet her burden of proof. Thus, from a legal standpoint such cases could be considered as "Not qualified." Under this assumption, then the CBR output and the ML output for cases 3, 4, 11, 14, and 18 would be the same.

4.2 The recidivism score problem

4.2.1 Methods

Cases were hand-picked from the original dataset in order to create a smaller, manageable case knowledge base. For the sake of simplicity, cases with the same charge description are compared in the following examples: hence from case 1 to case 15 are cases of aggravated assault with firearm, and from case 16 to case 31 are arrest cases with no charges (appendix C). Furthermore, more input facts were present in the original dataset, but only the input facts relevant for the CBR system were kept in the CKB.

4.2.2 Examples

Let us start with case 3 from appendix C.

	ANALYSIS	
`AgeCat`	x 25-45	neutral
`Sex`	M	neutral
`JuvenilePriors`	No	pro
`PriorsCount`	$x \geq 4$	con
`ChargeDegree`	F	con
`LenghtOfStay`	x < 50	pro
`Race`	other	neutral

Table 10. Dimensional analysis for case 3, appendix C.

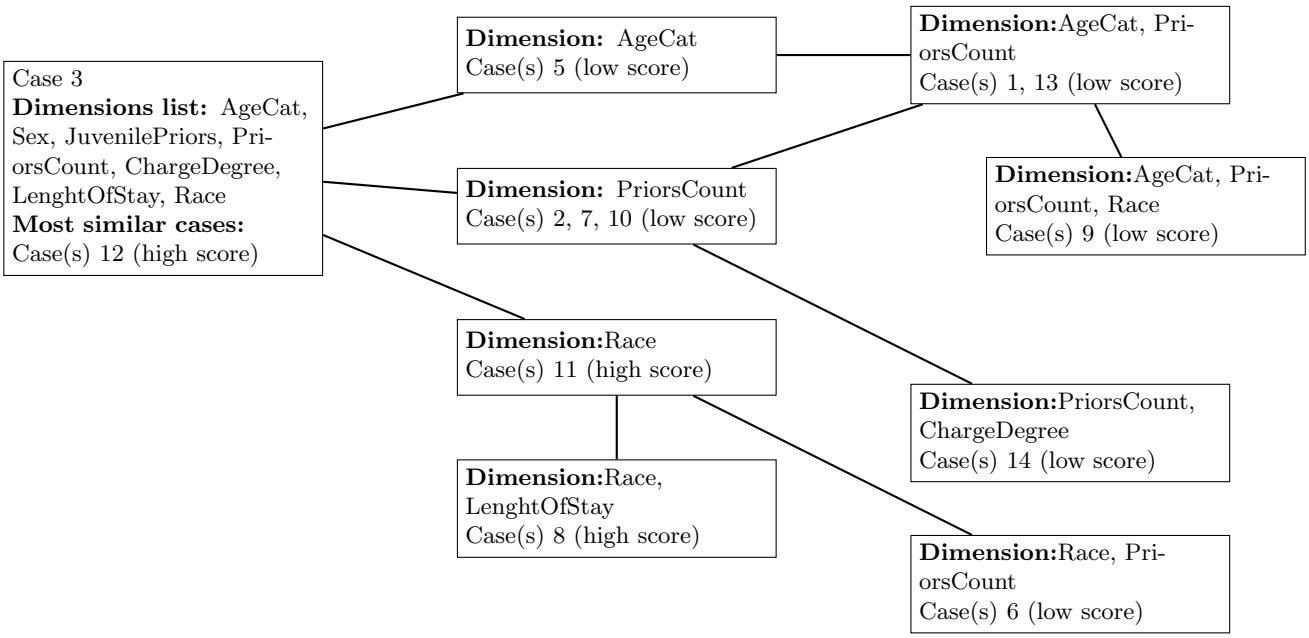

Figure 13. Claim lattice for case 3, appendix C.

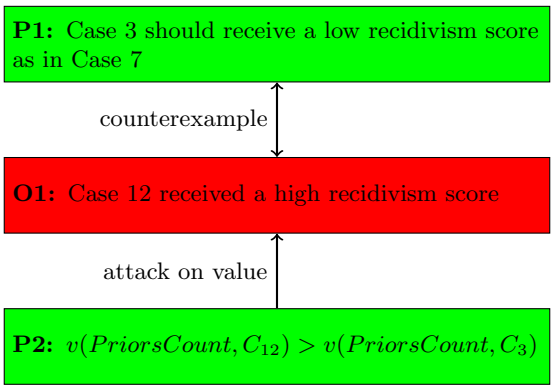

Figure 14. Argument game for case 3, appendix C.

It was mentioned in the previous examples that these argument games were originally developed using downplaying moves as well. It is worth noting that in this case and a few others removing downplaying moves ultimately did not change the CBR output, thus the presented game can be seen as a subset of the original game. In other cases, however, removing downplaying moves lead to different, quite unintuitive outcomes for the CBR model. Let us consider case 23 from appendix C.

	ANALYSIS	
`AgeCat`	25-45	neutral
`Sex`	M	neutral
`JuvenilePriors`	No	pro
`PriorsCount`	x > 4	con
`ChargeDegree`	F	con
`LenghtOfStay`	x < 50	pro
`Race`	African-American	con

Table 11. Dimensional Analysis for case 23.

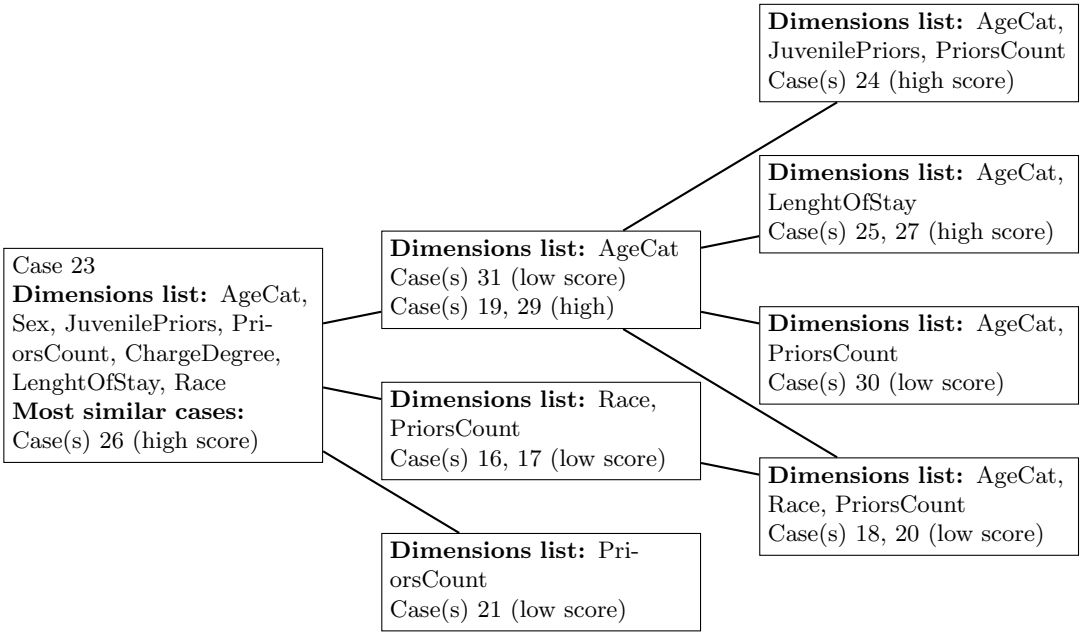

Figure 15. Claim Lattice for case 23.

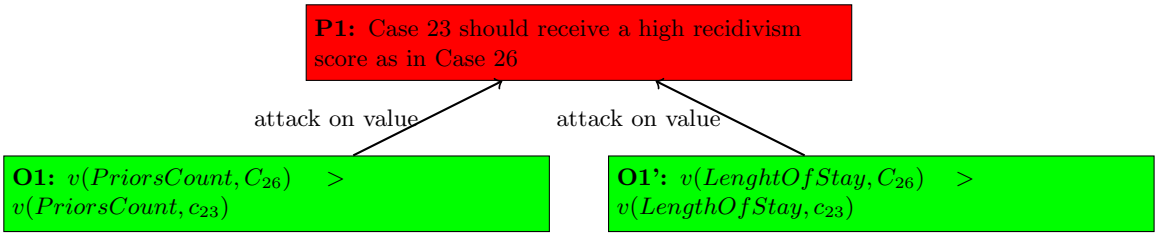

Figure 16. Argument Game for case 23.

In the original game the CBR output was the same as the ML output, i.e. case 23 should receive a high recidivism score. However, in this game this outcome is overruled. Intuitively case 23 should receive a high recidivism score based on what we know from the dimensional analysis, as there are several dimensions that appear to be against the defendant. This result seems to suggest that, even though including downplaying moves may result in longer games, they may also help providing better outputs.

Let us now look at case 16 and case 17. Case 16 is case 17's most similar case, and vice versa.

ANALYSIS		
AgeCat	25-45	neutral
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	2	neutral
ChargeDegree	F	con
LenghtOfStay	x < 50	pro
Race	other	neutral

Table 12. Dimensional analysis for case 16, appendix C.

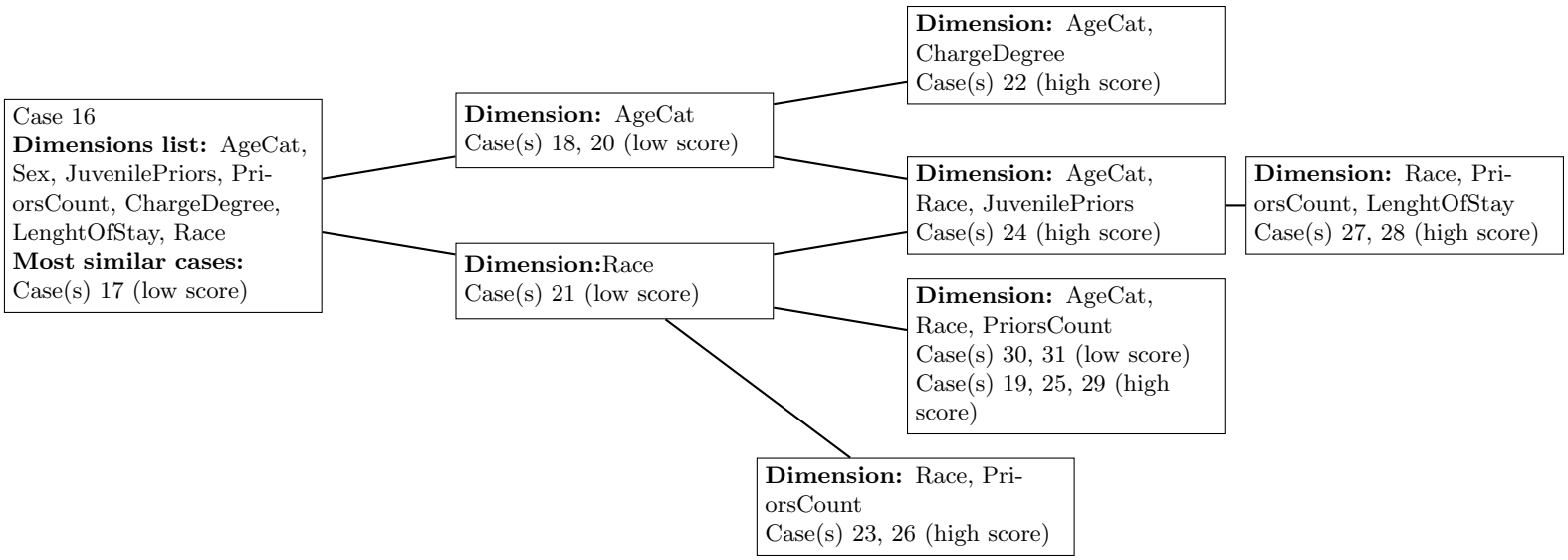

Figure 17. Claim lattice for case 16, appendix C.

P1: Case 16 should receive a low recidivism score as in Case 17

Figure 18. Argument game for case 16, appendix C.

ANALYSIS		
AgeCat	25-45	neutral
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	3	neutral
ChargeDegree	F	con
LenghtOfStay	x < 50	pro
Race	other	neutral

Table 13. Dimensional Analysis for case 17, appendix C.

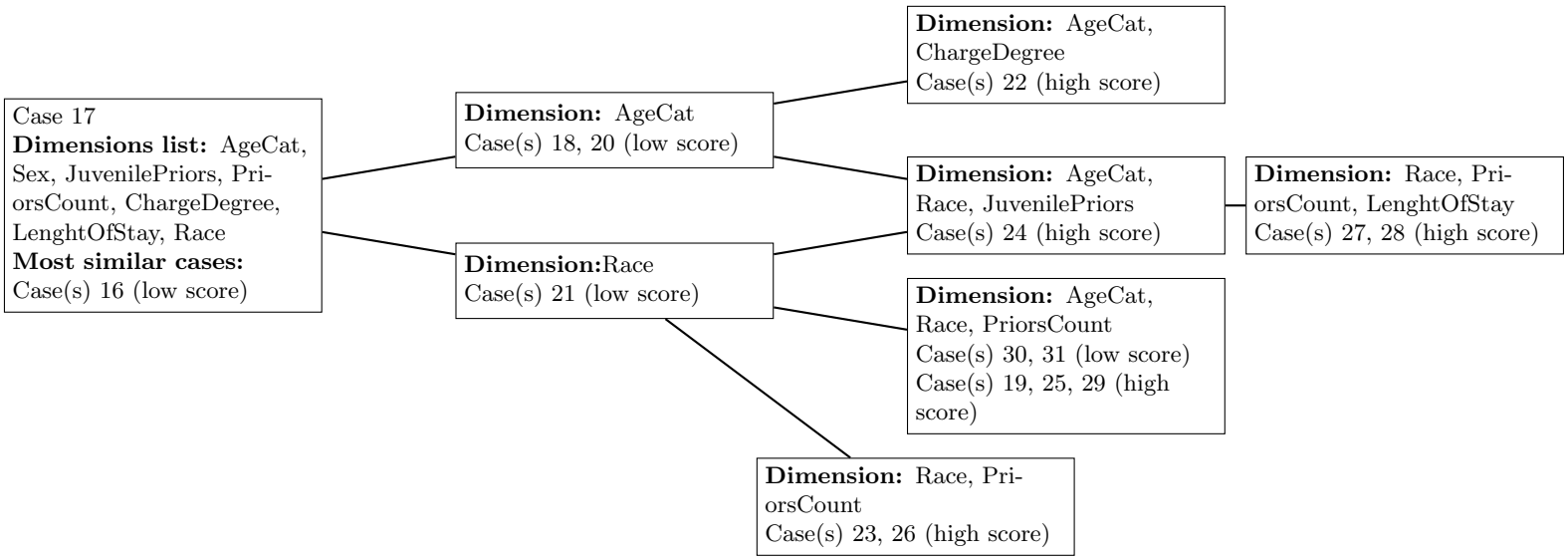

Figure 19. Claim lattice for case 17, appendix C.

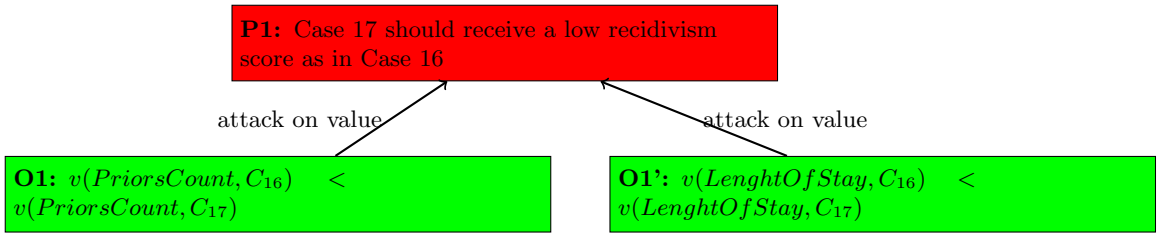

Figure 20. Argument game for case 17, appendix C.

If the specific values are considered, it can be seen that case 16 is a subset of case 17. Thus, if case 17 was given a low score, there are no grounds for attacking case 16 based on DV pairs. The opposite, however, does not hold, although it is interesting to note that, if downplaying moves are included, then the argument for giving case 17 a low recidivism score resulted as justified. Case 1 and case 13 are similar in that regard (the complete argument games can be found in appendix E).

4.2.3 Results

Similarly to the previous case study, out of the 31 cases chosen for the recidivism CKB 20 were used to generate argument games, and to compare the result of these games against the machine learning outputs. The results are reported in the table below. Nine of these cases the CBR model yielded the same outputs. It

is noteworthy that there is always a discrepancies with scores that are originally classified as medium. A binary output might not be the best option in this case study. Finally, similarly to the social welfare benefit problem, a previous trial run relying on downplaying moves provided better results.

Case	ML output	CBR output
1	low	low
2	medium (low)	high
3	low	low
4	low	low
5	medium (low)	high
6	low	low
7	medium (low)	high
10	low	low
11	high	low
12	high	high
13	low	high
16	low	low
17	low	low
19	medium (high)	low
22	high	low
23	high	low
25	medium (high)	low
26	high	low
29	high	high
31	low	high

Table 14. List of ML outputs and CBR outputs.

4.3 Natural language explanations

As stated beforehand, this work mostly focused on AX explanations. However, the rough idea for natural language explanations is now presented.

The simplest approach would be to create templates that the NLX engine can automatically fill in with the relevant information. The templates can be of two types: for the social welfare benefit problem, one for an awarded benefit and one for a denied benefit; for the recidivism score problem, one for a high score, and one for a low score. Unlike AX explanations, NLX explanations do not contain argument games; they do, however, point to past cases that are relevant to the current fact situation. For case 14 from the social welfare benefit problem a NLX explanation could look like the following:

The benefit was awarded because the applicant has reached the pensionable age and currently lives in the UK. **Although** the applicant did not pay the last four contribution years, the benefit was awarded in past cases (e.g. Case 3) in which the applicant did not pay the last four contribution years but the capital resources are below 3000£. **Lastly,** the hospital is within the required distance.

A similar template can be used for the recidivism score problem, although it may require a more thoughtful design. Indeed, the recidivism scores are based on statistical relationships, and it was said how such relationships bear little meaning to humans when they figure in an explanation (Miller T., 2017). Nevertheless the concept is the same: the relationship between the current decision, the input facts, and past cases should be expressed in a simple, yet effective language.

4.4 Discussion

In the experiments introduced in this paper an issue arose with cases which outcome was decided differently from the ML outputs. As stated beforehand, the original datasets were downsized in order to make them manageable for a human; in particular, for what concerns the social welfare benefit dataset this meant getting rid of the noise variables. This case will be discussed first.

An argument could be made that these noise variables were deemed relevant by the ANN. Such noise features would likely be absent in a real life situation, thus it is possible that the ANN would perform differently in such a scenario. This is a speculation, but in (Bench-Capon T.J., 1993) it is also admitted that the neural networks performed better the closer the data resembled likely submissions. However, the divergences may depend on yet another issue, that is, the ANN in this case and the CBR system use different notions of "on-pointness." In artificial neural networks "on-pointness" can be defined as the weights' activation patterns: if a new input triggers the same pattern as a known input, then the network will classify it accordingly. In the CBR system presented here, "on-pointness" is determined on the basis of some shared features (i.e., dimensions). A different definition of "on-pointness" may require design changes for the CBR system. For example, a system like VJAP (Grabmair M., 2017) can more likely accommodate a definition of "on-pointness" that resembles a neural network's classification method, and that is because VJAP makes use of weight parameters to predict and explain legal cases as well.

Aside from the influence of noise variables and a different notion of on-pointness, it is also possible that the dimensions' directions as determined in this work do not, in fact, properly model the problem situation.

Nevertheless it can be debated whether the explanation method must accommodate the machine learning algorithm which output it tries to explain. This is not necessarily a weakness of the system; on the contrary, cases of interest are precisely where the ML algorithm and the CBR system disagree. This may point to issues existing in the algorithm or the dataset employed, or alternatively these cases may offer the ground for challenging the machine learning output. Indeed, if the ANN in this case and the CBR system disagree it seems a person would be justified in questioning the decision.

For instance, let us take as an example any one of the cases that the CBR would decide differently from the ANN. Perhaps it should be part of the CBR output itself that the algorithm originally yielded a different decision, and provide reasons why the CBR reached the opposite conclusion. After all, the current

CBR is assuming a rule that was not implied in the original dataset: although the ANN indeed awarded the benefit in cases in which the conjunction of paid $= 0$ and capital < 3000 holds, it was not stated explicitly anywhere that, given a certain threshold, then the benefit should still be awarded. In the current case studies these difference may be due a different logic employed in solving the problem situation, but the same kind of issue may arise where the same logic is used in both the ANN and the CBR. Surely the ANN seems to imply that it is possible to not have paid the past four contribution years and still receive the benefit, but whether it was awarded or not may again depend on noise variables.

For what concerns the recidivism score case study there are further issues that need to be considered. For instance it must be noted that arbitrary thresholds were chosen when determining the dimensions' directions: this means that entirely different choices were also possible, and they may have lead to more or less successful results. Moreover, even though dimensions in this case study have neutral directions it was not possible to determine the relationship between inputs from the logistic regression model, therefore it was not possible to use the current approach to its full potential. Lastly, it is doubtful whether simplifying the original output in order to make it binary was the best approach.

In conclusion, the last research question addressed in this work can be answered. In both the recidivism score problem and the social welfare benefit problem the CBR output corresponds to the ML output for only four and nine of the examples respectively. As suggested in the previous sections, it appears that downplaying moves should be fully formalised and included in a model of explanation.

5 Conclusions and future research

In this work the problem of generating explanations for ML outputs was addressed. The issue of interpretability, that is, the ability to explain a certain output in terms understandable to a human, has become a well-acknowledged matter in the field of AI. Thus this research attempted to create a model for generating explanations that could serve such purpose.

After presenting an overview of the literature on human understanding, AI interpretability, legal argumentation, and case-based reasoning systems the main questions addressed in this work were introduced. The first problem was choosing between two main CBR systems, HYPO and CATO. In other words, the issue was which one between dimensions and factors could be the most optimal tool for the purposes of this research. The decision to pick dimensions over factors was informed by the literature: even though factors have the advantage of being easily implemented, dimensions allow for a more fine-grained representation of a problem situation. Thus HYPO was used as the main blueprint for this research.

By using the HYPO framework in order to represent the chosen case studies it was possible to conclude that, even though HYPO does indeed provide useful

tools, they are not sufficient for a model which purpose is generating explanations. Indeed, in this work two main limitations of the original HYPO were addressed: firstly, the fact that dimensions always have polar opposites, and do not account for situations in which a dimension may favour only one side, or no side at all; secondly, the fact that dimensions are not allowed to interact with each other to seek additional information. The solution proposed here was to categorise dimensions in pro, con, and neutral dimensions; in particular, for neutral dimensions it was suggested that rules could be employed in order for a dimension to query another.

The model resulting from such changes was then presented. Taking the literature into account, the model is expected to deliver explanations in two different formats: the first, a natural language explanation (NLX), is meant for non-expert users; the second, the argument explanation (AX), is meant for expert users and provides graphs representing argument games. The purpose of these argument games is to show whether, according to the current model, the ML decision is justified or not. In order to illustrate how the model works, the framework was tested on the social welfare benefit and recidivism score case studies. In both examples, 20 cases were chosen and argument games were generated. In the social welfare benefit problem, the CBR output was the same as the ANN output in only four cases; for what concerns the recidivism score problem, the ML output resulted as justified according to the CBR system in only nine cases.

Although promising, different issues have yet to be addressed if a CBR approach to ML explanations is chosen. For instance, the argument games presented here may be too simple to prove optimal for more complex problem situations. Moreover, the design choices made in this work were not unique: for example a different notion of on-pointness could be employed, which also raises the problem of the different notion of similarity used in CBR systems and ML algorithms. It would be interesting to explore how different definitions affect the CBR performance.

Nevertheless, I believe that this work showed how an approach relying on CBR might be a viable option in generating explanations for ML outputs. Originally, the following research questions were also addressed in the research proposal:

- Given these tools, is the data obtained from machine learning algorithms enough, or should information be added to it?
    1. If we need further information, should we add it through knowledge engineering, or through a dialogue approach?
        (a) If a dialogue approach is chosen, which are the conversational rules and speech acts for a method of explanation?

Due to time constraints it was only possible to answer the questions introduced in the early sections of this work. However, here a few ideas are suggested to serve as pointers to future research in order to answer the remaining research questions.

A first hypothesis is that the answer to these questions may depend on what information is being added to the system. For example, in the social welfare benefit problem it was mentioned how, for a dimension such as `HospitalDistance`, knowledge about the distribution of hospitals in the area could be useful: for instance, a spouse can be an out-patient, but the structure can fail to be in the required distance; however, the next hospital is too far to be a viable option. In this scenario, if other facts applied then it might be decided that the dimension `HospitalDistance` favours the applicant, even though she fails to meet the requirement. It is apparent how this kind of information can be knowledge-engineered into the system, because it is factual knowledge. However, if non-factual knowledge is used in taking a decision, then the issue might be more complicated. It should no longer be controversial to say that there is no such thing as a neutral algorithm – or alternatively it can be said that there is no such thing such as a neutral dataset. For example, the COMPAS algorithm appears to be racially biased as overall it assigns higher recidivism scores to black defendants compared to non-black defendants. Although it could be the case that the algorithm is structurally biased, the outcomes can be explained if the data these algorithms are trained on is taken into account: because human decisions thus far were driven by race differences, and because black people are more likely to encounter social and economical difficulties, the algorithm learnt that indeed, statistically, a black defendant is more likely to recidivate than a non-black defendant. In other words, it is the social system which is structurally biased, and the algorithms learn and reinforce the same patterns.

Then, what is the solution in such a case? Clearly there is not a quick fix to this problem, however it might be desirable in this case that, whatever extra knowledge is implemented on the system, it is not knowledge-engineered by a few people. Rather, it may be possible for the system to interact with, and learn from its human users. Although human users are just as biased as human knowledge engineers, there is at least the advantage that this knowledge is not fabricated on the basis of some more or less accurate assumptions, but rather it is inferred from several interactions with several people from several backgrounds.

This, I believe, is a step beyond creating methods of explanation for machine learning outputs. Indeed, there is no non-factual knowledge in the COMPAS dataset (or in the social welfare benefit dataset), and if there is it cannot be inferred at this stage. Rather than a tool for explaining machine learning outputs, a dialogue-based approach can be a tool for creating explainable machine learning outputs instead. Thus, whereas until now the question has been if machine learning outputs can be made somewhat interpretable using argumentation, now the question is if it is possible to use argumentation in order to create interpretable algorithms. The possibilities and limitations of a dialogue-based approach should be investigated in future research. This would be an effort requiring input from different disciplines, from argumentation in order to determine the kinds of dialogues allowed and the dialogues' rules, to human-computer interaction to figure out how people would interact with and react to such a system.

In conclusion, I believe this research has highlighted how case-based reasoning can be used for generating explanations that are both easily understandable and accessible. As our daily lives become more and more automated, the pursuit of interpretable AI is now of vital importance. As such, this work aimed at contributing to this problem by presenting a model of explanation rooted in existing CBR tools.

Appendices

A Examples of Ad Feelders' model outputs

```
> predict(compas.logreg,data.frame(c_charge_degree = "F",race
= "Asian",age_cat = "25 - 45",sex = "Male",priors_count =
3,length_of_stay = 10),type="response")
1
0.3275991
```

```
> predict(compas.logreg,data.frame(c_charge_degree = "F",race
= "Hispanic",age_cat = "25 - 45",sex = "Male",priors_count =
3,length_of_stay = 10),type="response")
1
0.4258317
```

```
> predict(compas.logreg,data.frame(c_charge_degree =
"F",race = "African-American",age_cat = "25 - 45",sex = "Fe-
male",priors_count = 3,length_of_stay = 10),type="response")
1
0.4070213
```

```
> predict(compas.logreg,data.frame(c_charge_degree = "F",race
= "Caucasian",age_cat = "25 - 45",sex = "Male",priors_count =
5,length_of_stay = 10),type="response")
1
0.5489378
```

```
> predict(compas.logreg,data.frame(c_charge_degree = "F",race =
"African-American",age_cat = "25 - 45",sex = "Male",priors_count
= 5,length_of_stay = 10),type="response")
1
0.5713281
```

```
> predict(compas.logreg,data.frame(c_charge_degree = "F",race =
"African-American",age_cat = "25 - 45",sex = "Male",priors_count
= 5,length_of_stay = 17),type="response")
1
0.5758207
```

B Social welfare benefit case knowledge base

	age	sex	paid	married	absent	capital	distance	in-patient	qualified
1	83	M	0	1	0	1100	2	1	1
2	69	F	0	1	0	300	31	1	1
3	75	F	0	1	0	500	15	1	0
4	92	M	0	1	0	900	6	1	0
5	76	M	0	1	0	2300	64	0	1
6	79	M	1	1	0	300	19	1	1
7	73	F	0	1	0	400	60	0	1
8	73	M	1	1	0	500	45	1	1
9	99	F	0	1	0	800	67	0	0
10	80	M	0	1	0	1200	82	0	0
11	95	M	1	1	0	300	5	1	0
12	80	F	1	1	0	200	11	1	0
13	71	F	1	1	0	2900	12	1	0
14	80	F	0	1	0	300	26	1	0
15	86	M	1	1	0	2200	47	1	0
16	69	F	0	1	0	2600	49	1	0
27	84	F	0	1	0	1300	88	0	0
18	95	F	0	1	0	1600	64	0	0
19	86	M	0	1	0	1900	70	0	0
20	62	M	0	1	0	6300	62	0	0
21	76	M	0	1	0	4400	69	0	0
22	86	M	0	1	0	1600	0	1	1
23	70	M	0	1	0	1500	1	1	1
24	69	F	0	1	0	1400	2	1	1
25	88	F	1	1	0	0	70	0	1
26	80	F	1	1	0	0	52	0	1
27	77	M	1	1	0	0	6	1	1
28	87	M	1	1	0	0	90	0	0
29	95	M	1	1	0	300	5	1	0
30	67	F	1	1	0	800	68	0	0
31	66	M	0	1	0	300	5	1	1

Table 15. Excerpt from the artificial social welfare benefit dataset. "Capital" reports amounts in £, whereas distance is in km.

C Recidivism score case knowledge base

	sex	age cat	race	juv priors	priors	los	c degree	charge	score	text
1	Male	> 45	Other	0	0	1	F	Agg Ass w/ firearm	1	low
2	Male	25-45	Caucasian	0	0	16	F	Agg Ass w/ firearm	5	medium
3	Male	25-45	Other	0	4	3	F	Agg Ass w/ firearm	4	low
4	Male	25-45	African-American	0	0	1	F	Agg Ass w/ firearm	1	low
5	Male	> 45	Other	0	12	1	F	Agg Ass w/ firearm	5	medium
6	Male	25-45	African-American	0	0	1	F	Agg Ass w/ firearm	2	low
7	Male	25-45	Caucasian	0	1	1	F	Agg Ass w/ firearm	5	medium
8	Male	25-45	African-American	0	4	83	F	Agg Ass w/ firearm	7	medium
9	Female	> 45	African-American	0	0	2	F	Agg Ass w/ firearm	1	low
10	Male	25-45	Caucasian	0	0	1	F	Agg Ass w/ firearm	2	low
11	Male	25-45	African-American	0	11	1	F	Agg Ass w/ firearm	8	high
12	Male	25-45	Caucasian	0	8	0	F	Agg Ass w/ firearm	8	high
13	Male	> 45	Asian	0	0	2	F	Agg Ass w/ firearm	1	low
14	Male	< 25	Other	0	0	4	M	Agg Ass w/ firearm	4	low
15	Male	< 25	African-American	0	2	8	F	Agg Ass w/ firearm	5	medium
16	Male	25-45	Other	0	2	0	F	Arrest case no charge	1	low
17	Male	25-45	Other	0	3	1	F	Arrest case no charge	4	low
18	Female	> 45	Caucasian	0	1	14	F	Arrest case no charge	1	low
19	Male	> 45	African-American	0	13	1	F	Arrest case no charge	6	medium
20	Male	> 45	Caucasian	0	0	28	F	Arrest case no charge	2	low
21	Male	15-45	African-American	0	0	1	F	Arrest case no charge	2	low
22	M	< 25	Caucasian	0	1	13	M	Arrest case no charge	10	high
23	M	25-45	African-American	0	19	1	F	Arrest case no charge	10	high
24	F	< 25	African-American	1	1	3	F	Arrest case no charge	10	high
25	M	> 45	African-American	0	3	129	F	Arrest case no charge	6	medium
26	M	25-45	African-American	0	25	26	F	Arrest case no charge	8	high
27	M	> 45	African-American	0	14	110	F	Arrest case no charge	9	high
28	M	> 45	African-American	0	21	59	M	Arrest case no charge	7	medium
29	M	> 45	African-American	0	17	6	F	Arrest case no charge	9	high
30	M	> 45	African-American	0	1	0	F	Arrest case no charge	1	low
31	F	> 45	African-American	0	5	1	F	Arrest case no charge	1	low

Table 16. Excerpt from the COMPAS dataset. Originally the scores are organised in low, medium, and high scores. For the sake of simplicity, in the current work scores from 1 to 5 are considered medium, while scores from 6 to 10 are classified as high.

D Argument games for the social welfare benefit problem

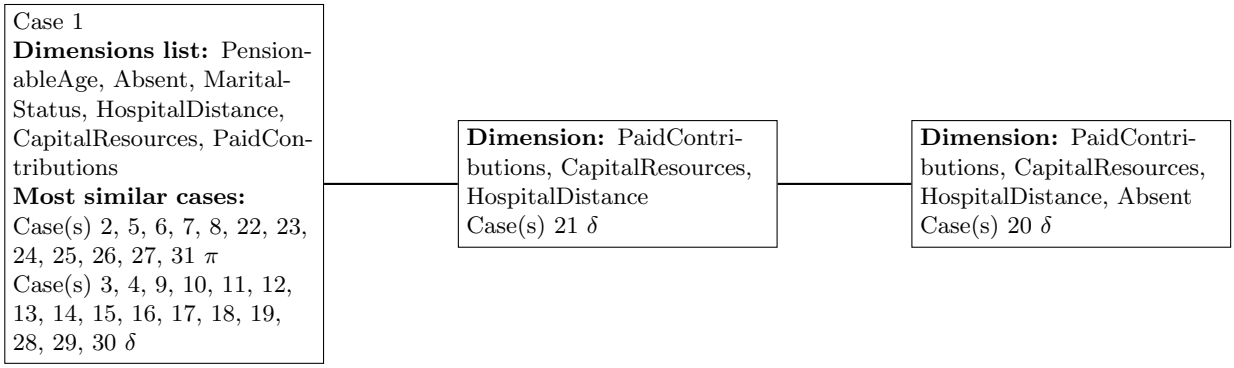

Figure 21. Claim Lattice for case(s) 1 to 19.

ANALYSIS		
`PensionableAge`	x $\geq$ 65	pro
`Absent`	No	pro
`PaidContributions`	No	Neutral $\rightarrow$ CapitalResources $\leq$ 3000£ $\rightarrow$ pro
`CapitalResources`	x $\leq$ 3000£	pro
`HospitalDistance`	x $<$ 50 km	pro

Table 17. Dimensional Analysis for case 1.

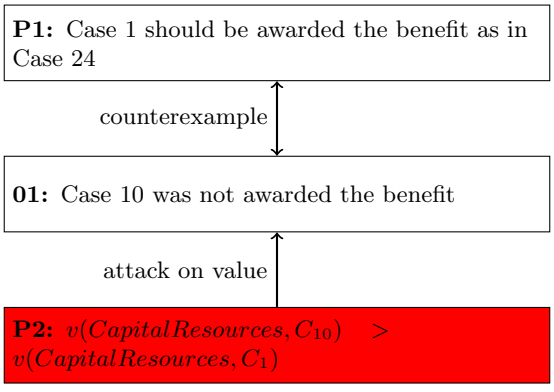

Figure 22. Argument game for case 1.

ANALYSIS		
PensionableAge	x $\geq$ 65	pro
Absent	No	pro
PaidContributions	No	Neutral $\rightarrow$ CapitalResources $\leq$ 3000£ $\rightarrow$ pro
CapitalResources	x $\leq$ 3000£	pro
HospitalDistance	x $<$ 50 km	pro

Table 18. Dimensional Analysis for case 2.

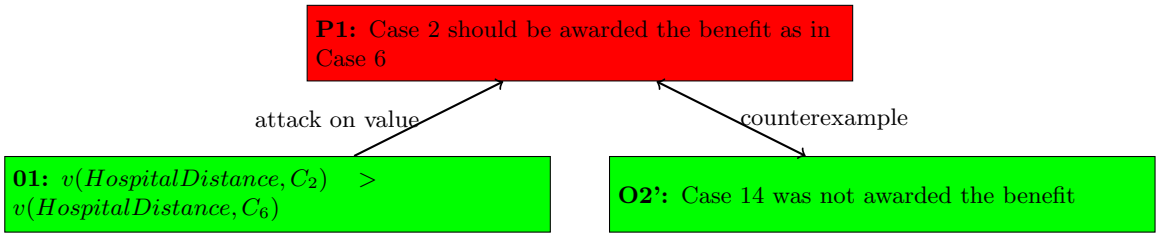

Figure 23. Argument game for case 2.

ANALYSIS		
PensionableAge	x $\geq$ 65	pro
Absent	No	pro
PaidContributions	No	Neutral $\rightarrow$ CapitalResources $\leq$ 3000£ $\rightarrow$ pro
CapitalResources	x $\leq$ 3000£	pro
HospitalDistance	x $<$ 50 km	pro

Table 19. Dimensional Analysis for case 3.

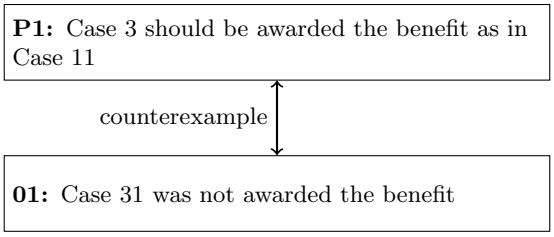

Figure 24. Argument game for case 3.

ANALYSIS		
PensionableAge	x $\geq$ 65	pro
Absent	No	pro
PaidContributions	No	Neutral $\rightarrow$ CapitalResources $\leq$ 3000£ $\rightarrow$ pro
CapitalResources	x $\leq$ 3000£	pro
HospitalDistance	x $<$ 50 km	pro

Table 20. Dimensional Analysis for case 4.

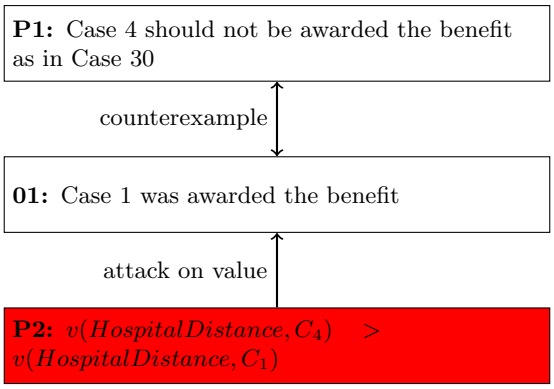

Figure 25. Argument game for case 4.

ANALYSIS		
PensionableAge	x $\geq$ 65	pro
Absent	No	pro
PaidContributions	No	Neutral $\rightarrow$ CapitalResources $\leq$ 3000£ $\rightarrow$ pro
CapitalResources	x $\leq$ 3000£	pro
HospitalDistance	x $>$ 50 km	pro

Table 21. Dimensional Analysis for case 5.

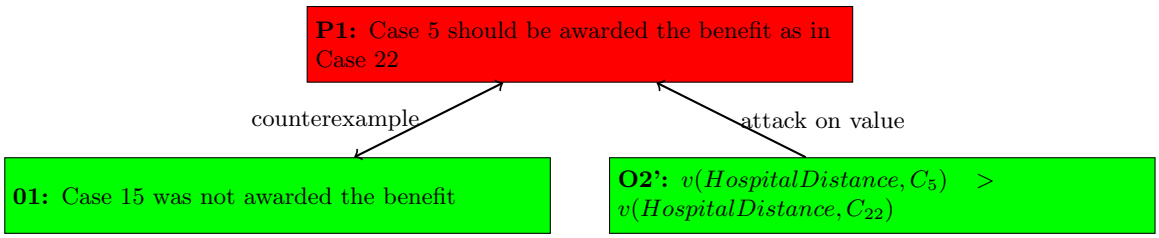

Figure 26. Argument game for case 5.

	ANALYSIS	
`PensionableAge`	x $\geq$ 65	pro
`Absent`	No	pro
`PaidContributions`	Yes	Pro
`CapitalResources`	x $\leq$ 3000£	pro
`HospitalDistance`	x $<$ 50 km	pro

Table 22. Dimensional Analysis for case 6.

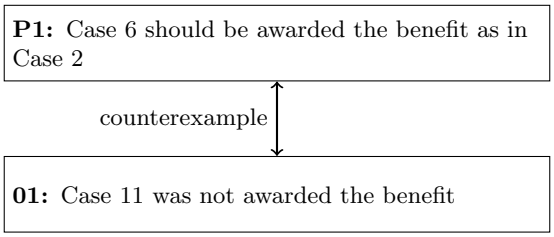

Figure 27. Argument game for case 6.

ANALYSIS		
PensionableAge	x $\geq$ 65	pro
Absent	No	pro
PaidContributions	No	Neutral $\rightarrow$ CapitalResources $\leq$ 3000£ $\rightarrow$ pro
CapitalResources	x $\leq$ 3000£	pro
HospitalDistance	x $>$ 50 km	pro

Table 23. Dimensional Analysis for case 7.

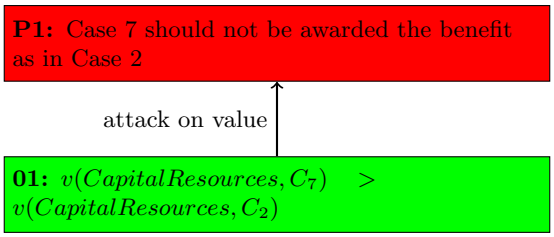

Figure 28. Argument game for case 7.

ANALYSIS		
PensionableAge	x $\geq$ 65	pro
Absent	No	pro
PaidContributions	Yes	Pro
CapitalResources	x $\leq$ 3000£	pro
HospitalDistance	x $<$ 50 km	pro

Table 24. Dimensional Analysis for case 8.

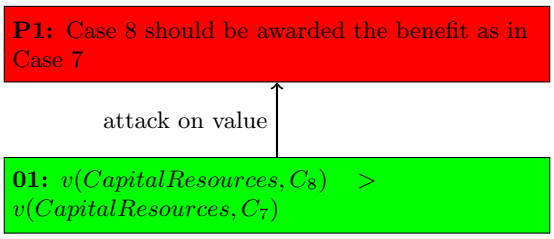

Figure 29. Argument game for case 8.

ANALYSIS		
PensionableAge	x $\geq$ 65	pro
Absent	No	pro
PaidContributions	No	Neutral $\rightarrow$ CapitalResources $\leq$ 3000£ $\rightarrow$ pro
CapitalResources	x $\leq$ 3000£	pro
HospitalDistance	x $>$ 50 km	pro

Table 25. Dimensional Analysis for case 9.

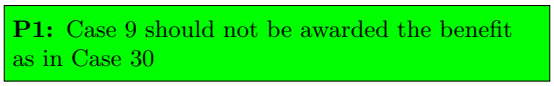

Figure 30. Argument game for case 9.

ANALYSIS		
PensionableAge	x $\geq$ 65	pro
Absent	No	pro
PaidContributions	No	Neutral $\rightarrow$ CapitalResources $\leq$ 3000£ $\rightarrow$ pro
CapitalResources	x $\leq$ 3000£	pro
HospitalDistance	x $>$ 50 km	pro

Table 26. Dimensional Analysis for case 10.

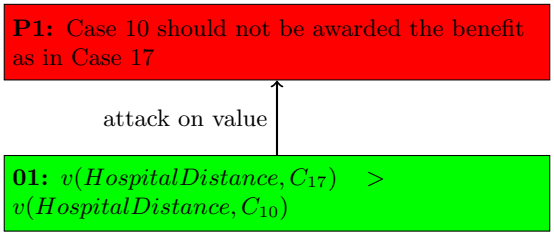

Figure 31. Argument game for case 10.

ANALYSIS		
PensionableAge	x $\geq$ 65	pro
Absent	No	pro
PaidContributions	No	Neutral $\rightarrow$ CapitalResources $\leq$ 3000£ $\rightarrow$ pro
CapitalResources	x $\leq$ 3000£	pro
HospitalDistance	x $>$ 50 km	pro

Table 27. Dimensional Analysis for case 11.

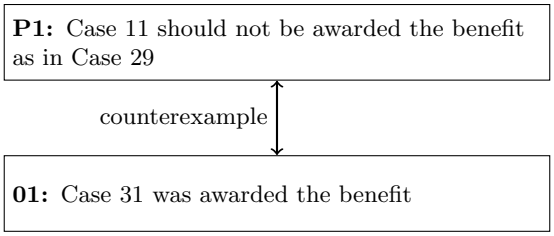

Figure 32. Argument game for case 11.

	ANALYSIS	
PensionableAge	x $\geq$ 65	pro
Absent	No	pro
PaidContributions	Yes	Pro
CapitalResources	x $\leq$ 3000£	pro
HospitalDistance	x $<$ 50 km	pro

Table 28. Dimensional Analysis for case 12.

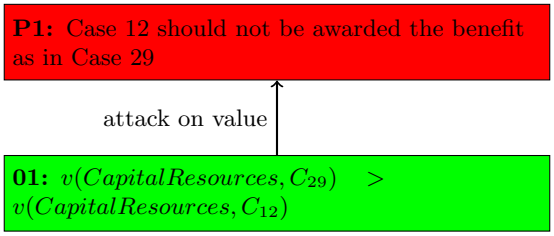

Figure 33. Argument game for case 12.

	ANALYSIS	
PensionableAge	x $\geq$ 65	pro
Absent	No	pro
PaidContributions	Yes	Pro
CapitalResources	x $\leq$ 3000£	pro
HospitalDistance	x $<$ 50 km	pro

Table 29. Dimensional Analysis for case 13.

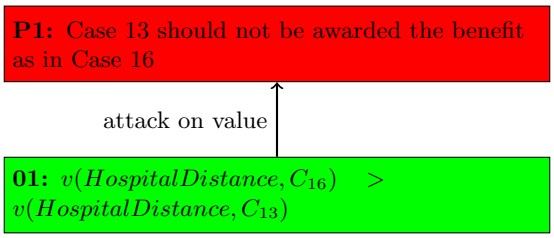

Figure 34. Argument game for case 13.

ANALYSIS		
PensionableAge	x $\geq$ 65	pro
Absent	No	pro
PaidContributions	No	Neutral $\rightarrow$ CapitalResources $\leq$ 3000£ $\rightarrow$ pro
CapitalResources	x $\leq$ 3000£	pro
HospitalDistance	x $<$ 50 km	pro

Table 30. Dimensional Analysis for case 14.

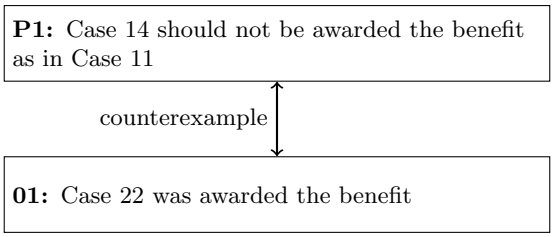

Figure 35. Argument game for case 14.

	ANALYSIS	
PensionableAge	x $\geq$ 65	pro
Absent	No	pro
PaidContributions	Yes	pro
CapitalResources	x $\leq$ 3000£	pro
HospitalDistance	x $<$ 50 km	pro

Table 31. Dimensional Analysis for case 15.

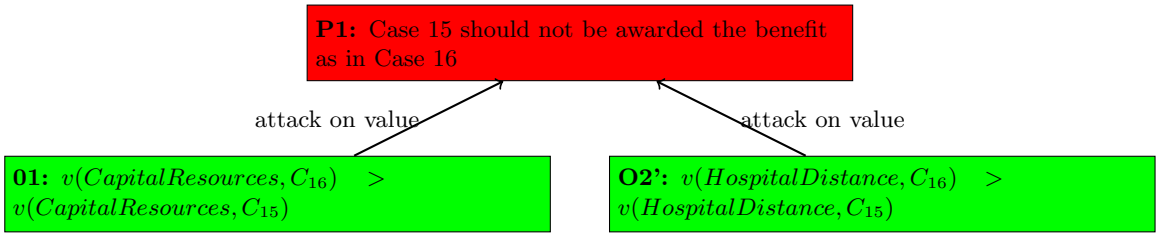

Figure 36. Argument game for case 15.

ANALYSIS		
PensionableAge	x $\geq$ 65	pro
Absent	No	pro
PaidContributions	No	Neutral $\rightarrow$ CapitalResources $\leq$ 3000£ $\rightarrow$ pro
CapitalResources	x $\leq$ 3000£	pro
HospitalDistance	x $<$ 50 km	pro

Table 32. Dimensional Analysis for case 16.

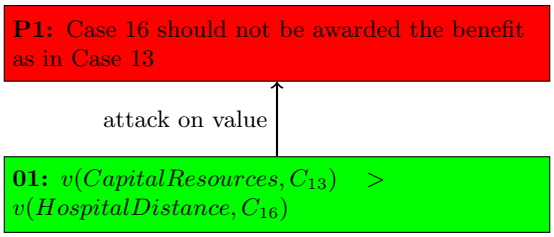

Figure 37. Argument game for case 16.

ANALYSIS		
PensionableAge	x $\geq$ 65	pro
Absent	No	pro
PaidContributions	No	Neutral $\rightarrow$ CapitalResources $\leq$ 3000£ $\rightarrow$ pro
CapitalResources	x $\leq$ 3000£	pro
HospitalDistance	x $<$ 50 km	pro

Table 33. Dimensional Analysis for case 17.

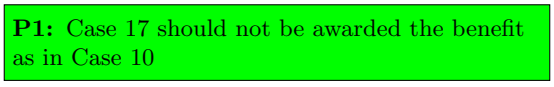

Figure 38. Argument game for case 17.

ANALYSIS		
PensionableAge	x $\geq$ 65	pro
Absent	No	pro
PaidContributions	No	Neutral $\rightarrow$ CapitalResources $\leq$ 3000£ $\rightarrow$ pro
CapitalResources	x $\leq$ 3000£	pro
HospitalDistance	x $>$ 50 km	pro

Table 34. Dimensional Analysis for case 18.

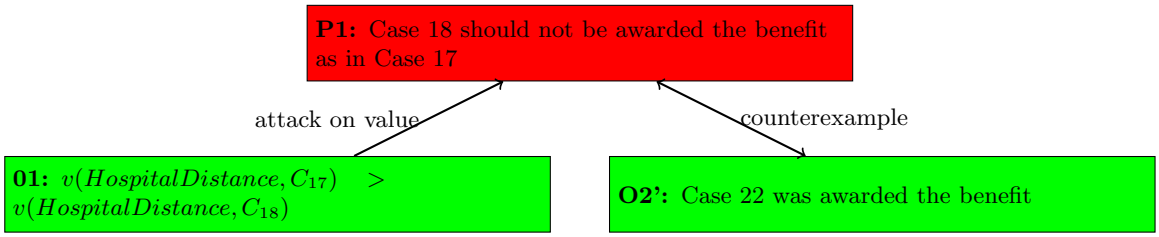

Figure 39. Argument game for case 18.

ANALYSIS		
PensionableAge	x $\geq$ 65	pro
Absent	No	pro
PaidContributions	No	Neutral $\rightarrow$ CapitalResources $\leq$ 3000£ $\rightarrow$ pro
CapitalResources	x $\leq$ 3000£	pro
HospitalDistance	x $>$ 50 km	pro

Table 35. Dimensional Analysis for case 19.

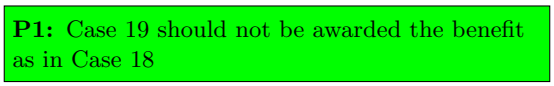

Figure 40. Argument game for case 19.

ANALYSIS		
PensionableAge	x $\leq$ 65	con
Absent	Yes	con
PaidContributions	No	Neutral $\rightarrow$ CapitalResources $\geq$ 3000£ $\rightarrow$ con
CapitalResources	x $\geq$ 3000£	con
HospitalDistance	x $>$ 50 km	con

Table 36. Dimensional Analysis for case 20.

P1: Case 20 should not be awarded the benefit as in Case 21

Figure 41. Argument game for case 20.

E Argument games for the recidivism score problem

ANALYSIS		
AgeCat	x > 45	pro
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	0	pro
ChargeDegree	F	con
LenghtOfStay	x < 50	pro
Race	other	neutral

Table 37. Dimensional Analysis for case 1.

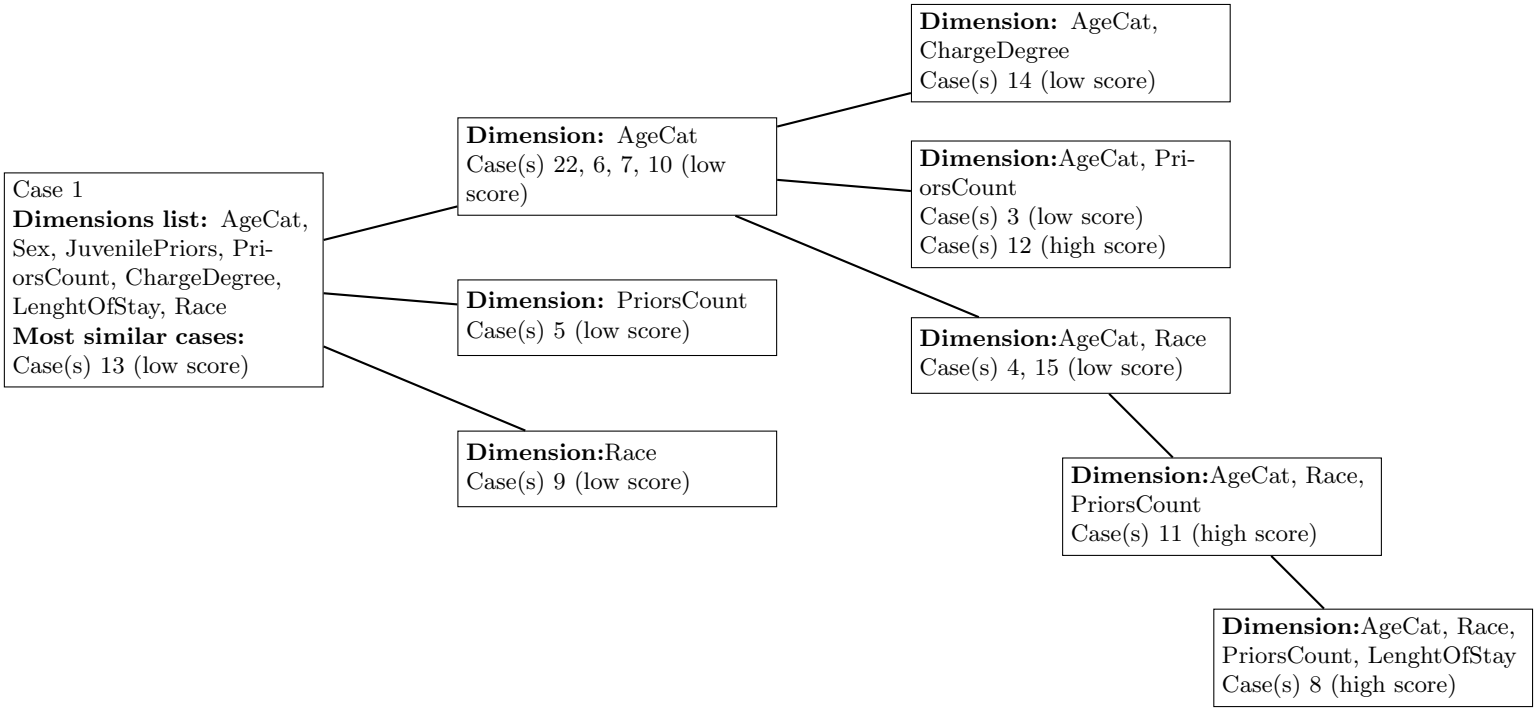

Figure 42. Claim Lattice for case 1.

P1: Case 1 should receive a low recidivism score as in Case 13

Figure 43. Argument Game for case 1.

ANALYSIS		
AgeCat	x 25-45	neutral
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	0	pro
ChargeDegree	F	con
LenghtOfStay	x < 50	pro
Race	other	neutral

Table 38. Dimensional Analysis for case 2.

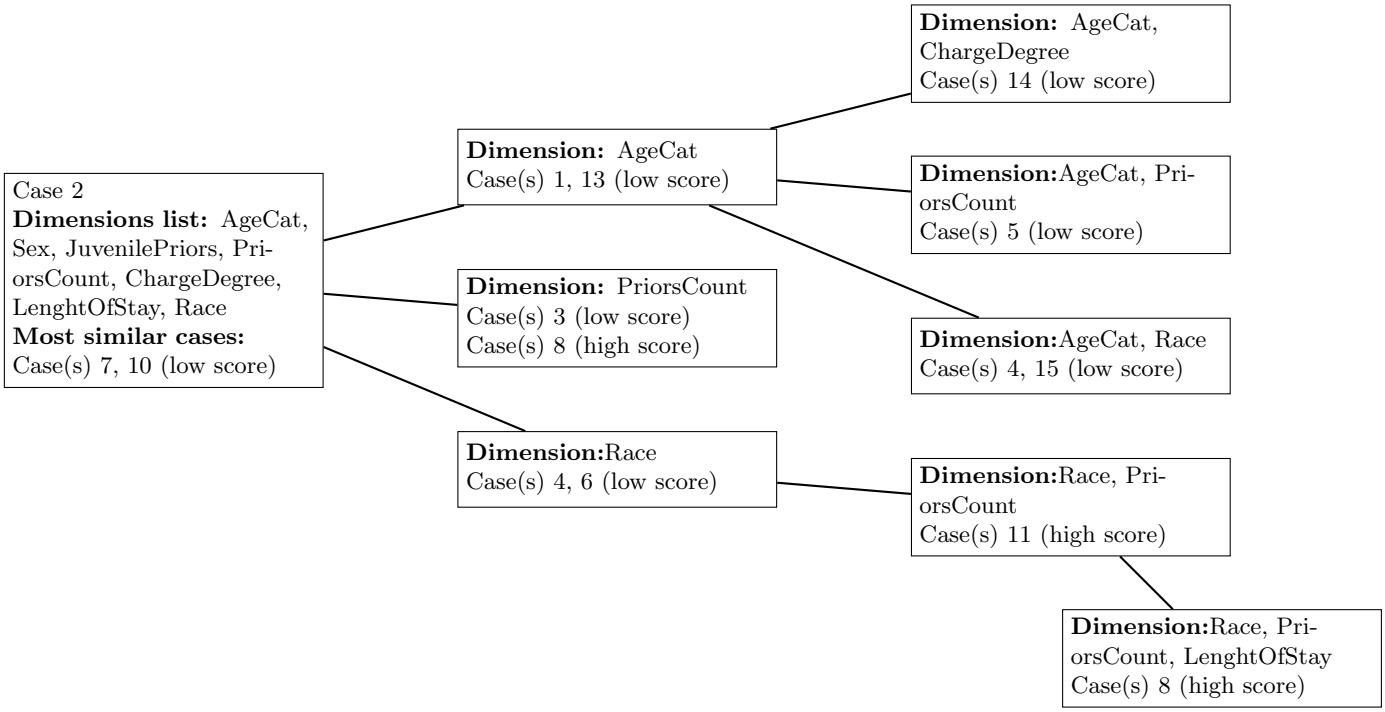

Figure 44. Claim Lattice for case 2.

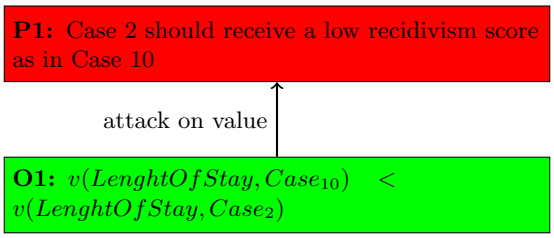

Figure 45. Argument Game for case 2.

ANALYSIS		
AgeCat	x 25-45	neutral
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	$x \geq 4$	con
ChargeDegree	F	con
LenghtOfStay	x < 50	pro
Race	other	neutral

Table 39. Dimensional Analysis for case 3.

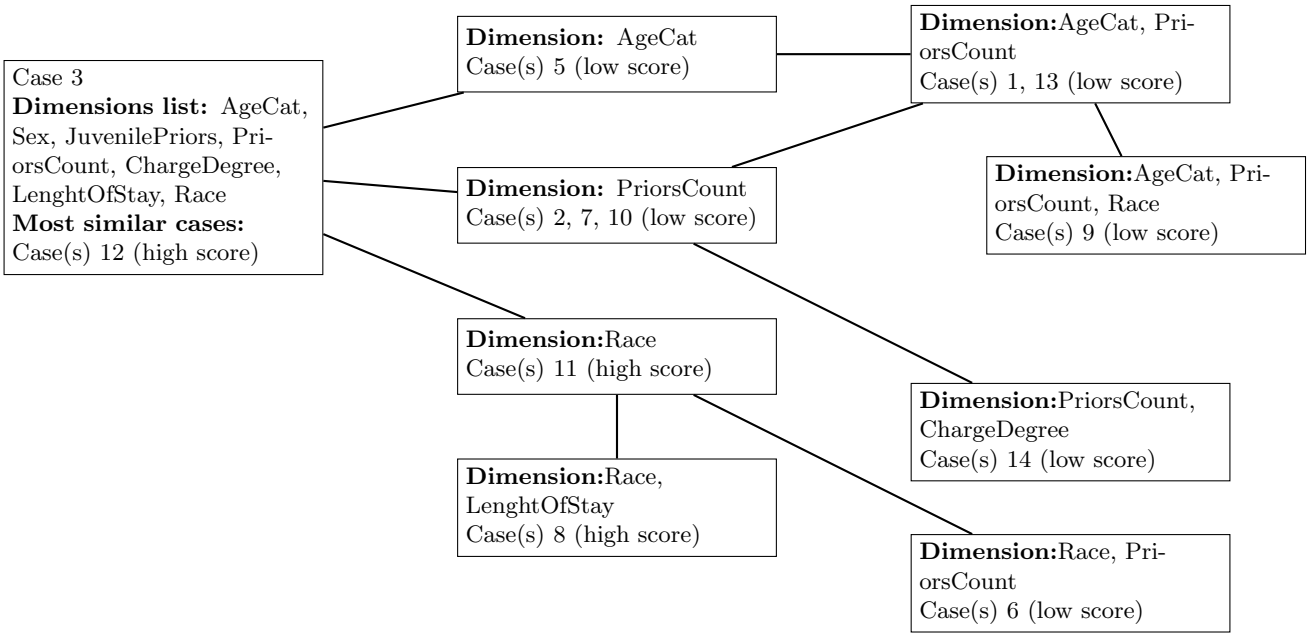

Figure 46. Claim Lattice for case 3.

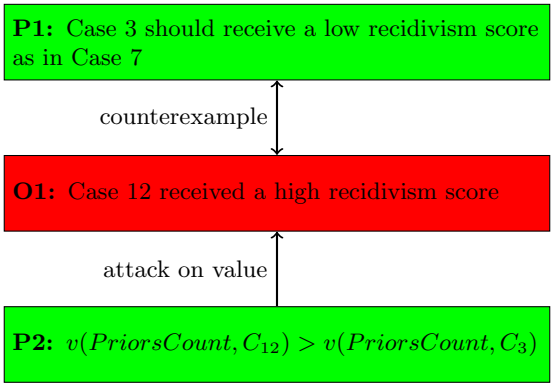

Figure 47. Argument Game for case 3.

ANALYSIS		
AgeCat	25-45	neutral
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	0	pro
ChargeDegree	F	con
LenghtOfStay	x < 50	pro
Race	African-American	con

Table 40. Dimensional Analysis for case 4.

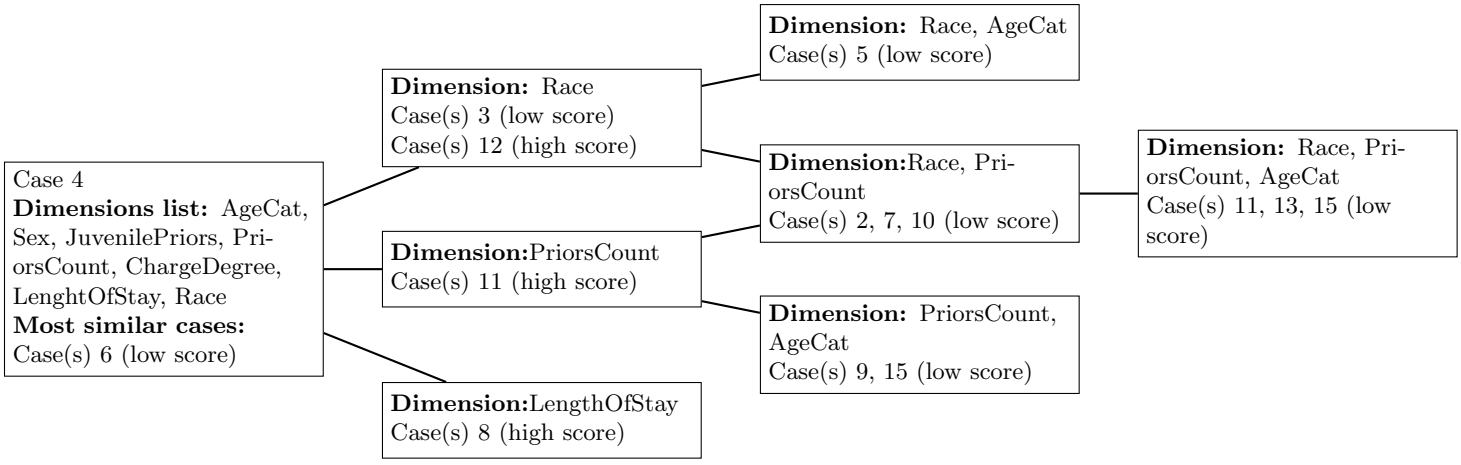

Figure 48. Claim Lattice for case 4.

P1: Case 4 should receive a low recidivism score as in Case 6

Figure 49. Argument Game for case 4.

	ANALYSIS	
AgeCat	x > 45	neutral
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	$x \geq 4$	con
ChargeDegree	F	con
LenghtOfStay	x < 50	pro
Race	Other	neutral

Table 41. Dimensional Analysis for case 5.

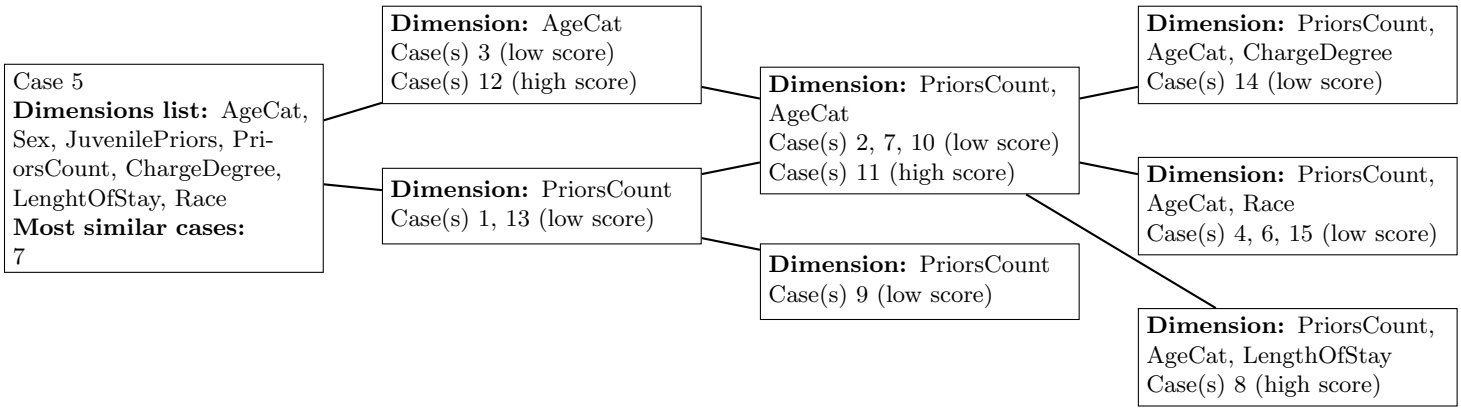

Figure 50. Claim lattice for case 5.

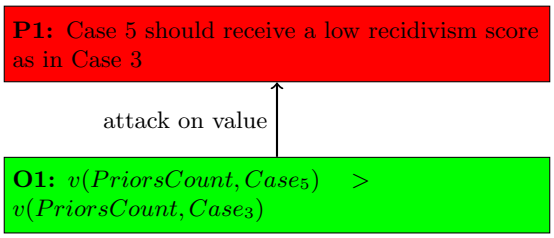

Figure 51. Argument Game for case 5.

ANALYSIS		
AgeCat	25-45	neutral
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	0	pro
ChargeDegree	F	con
LenghtOfStay	x < 50	pro
Race	African-American	con

Table 42. Dimensional Analysis for case 6.

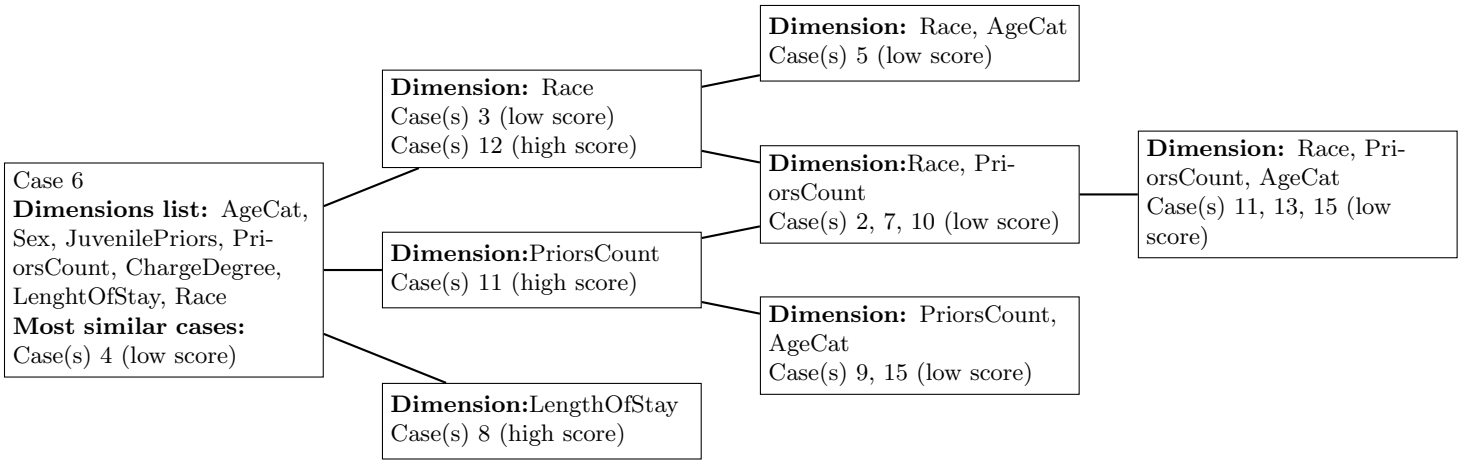

Figure 52. Claim Lattice for case 6.

P1: Case 6 should receive a low recidivism score as in Case 4

Figure 53. Argument Game for case 6.

ANALYSIS		
AgeCat	25-45	neutral
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	1	pro
ChargeDegree	F	con
LenghtOfStay	x < 50	pro
Race	Caucasian	neutral

Table 43. Dimensional Analysis for case 7.

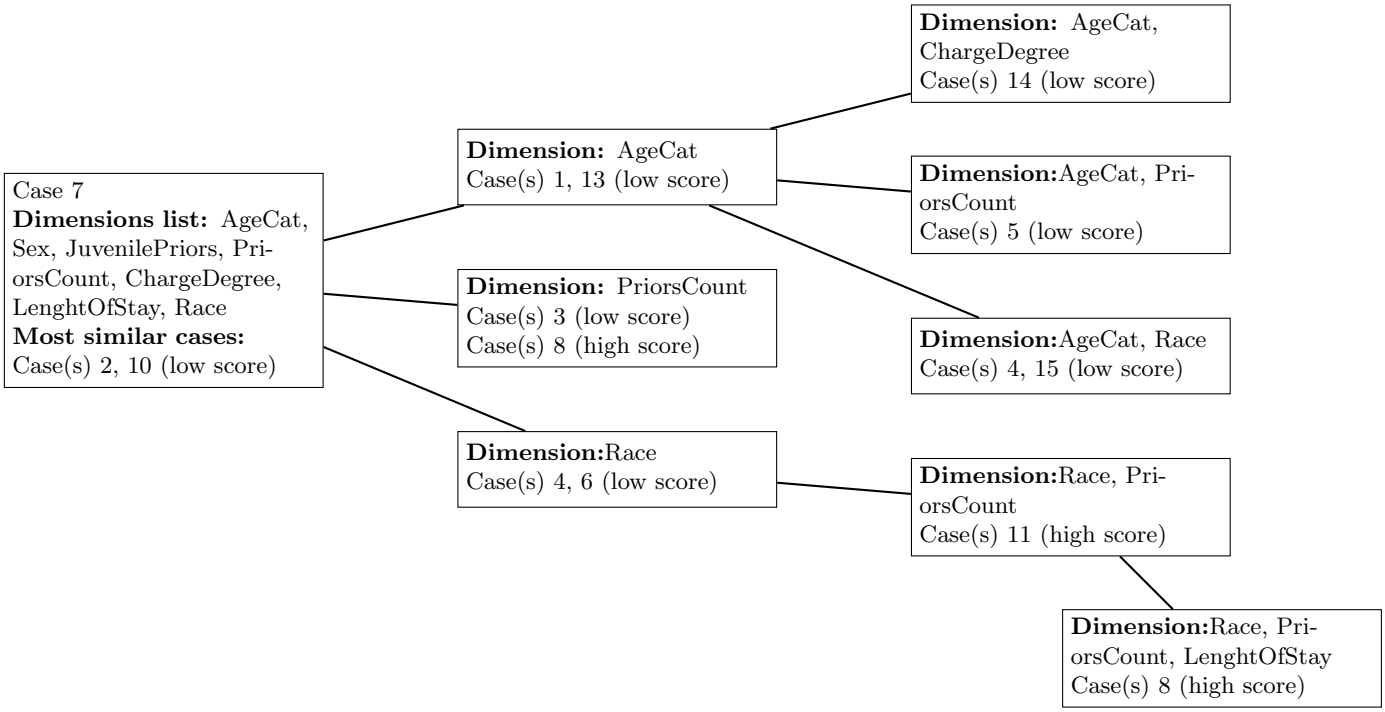

Figure 54. Claim Lattice for case 7.

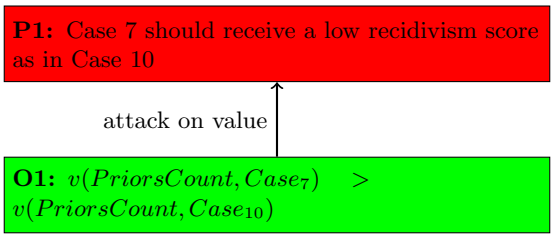

Figure 55. Argument Game for case 7.

	ANALYSIS	
AgeCat	25-45	neutral
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	0	pro
ChargeDegree	F	con
LenghtOfStay	x < 50	pro
Race	Caucasian	neutral

Table 44. Dimensional Analysis for case 10.

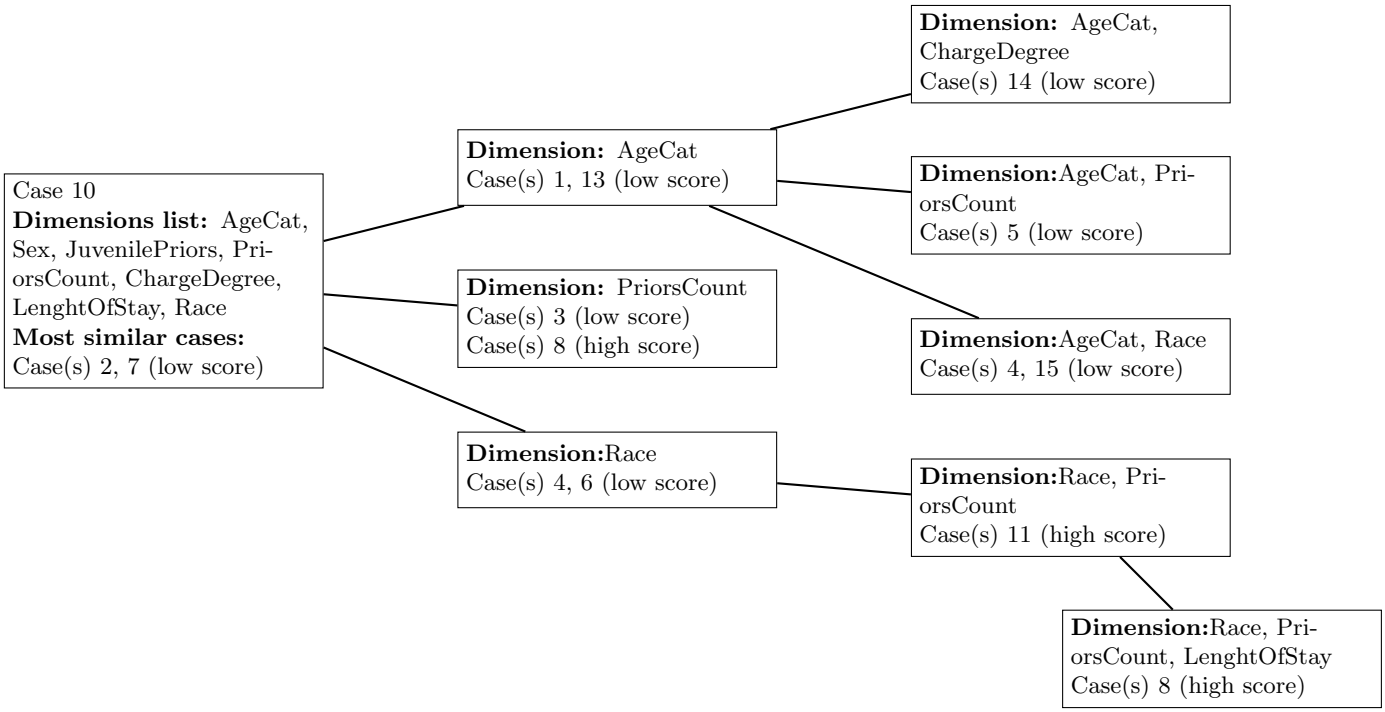

Figure 56. Claim Lattice for case 10.

P1: Case 10 should receive a low recidivism score as in Case 7

Figure 57. Argument Game for case 10.

ANALYSIS		
AgeCat	25-45	neutral
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	11	con
ChargeDegree	F	con
LenghtOfStay	x < 50	pro
Race	African-American	con

Table 45. Dimensional Analysis for case 11.

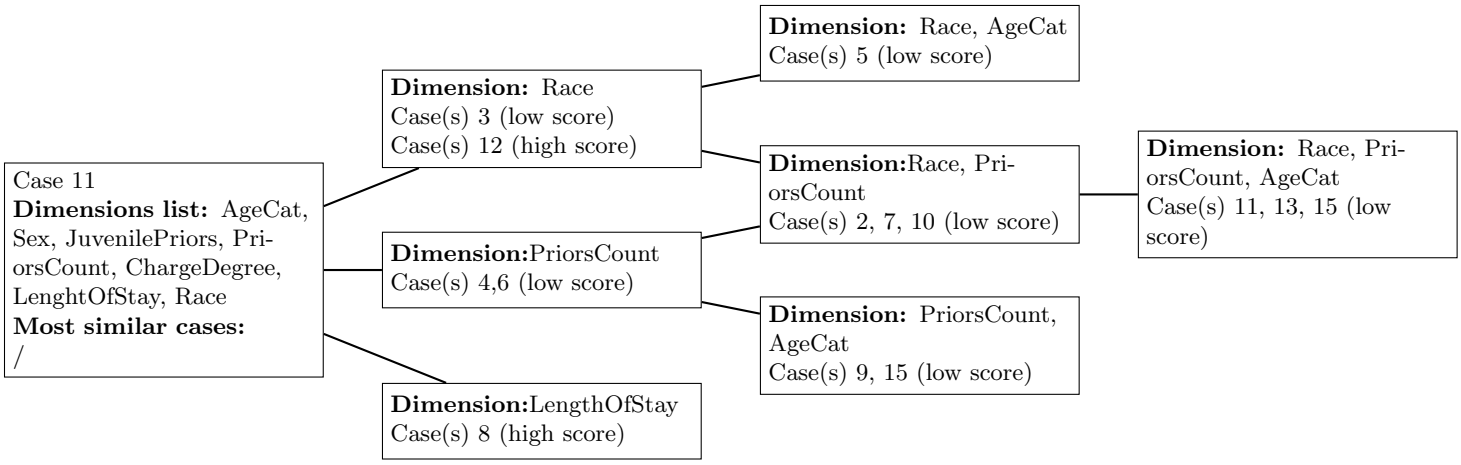

Figure 58. Claim Lattice for case 11.

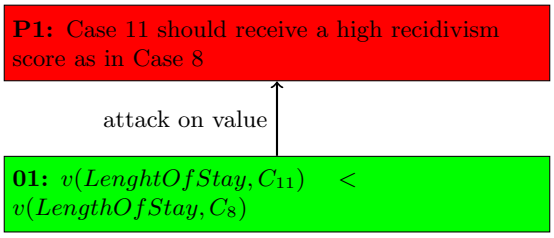

Figure 59. Argument Game for case 11.

ANALYSIS		
AgeCat	25-45	neutral
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	$x \geq 4$	con
ChargeDegree	F	con
LenghtOfStay	x < 50	pro
Race	Caucasian	neutral

Table 46. Dimensional Analysis for case 12.

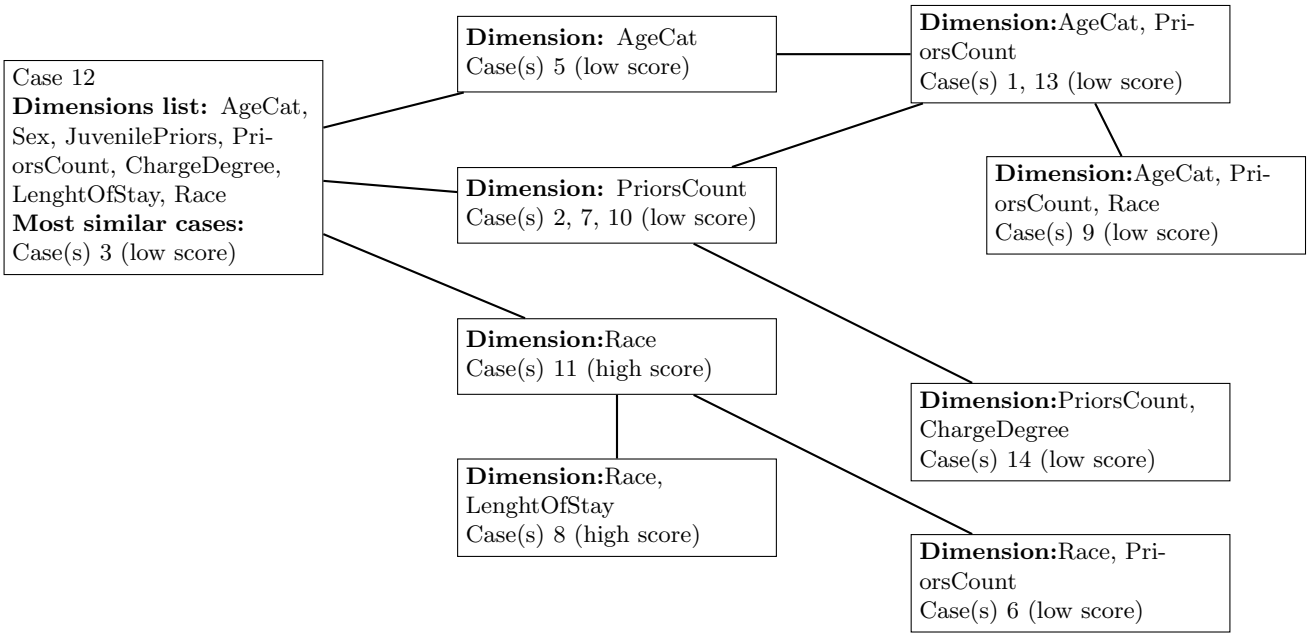

Figure 60. Claim Lattice for case 12.

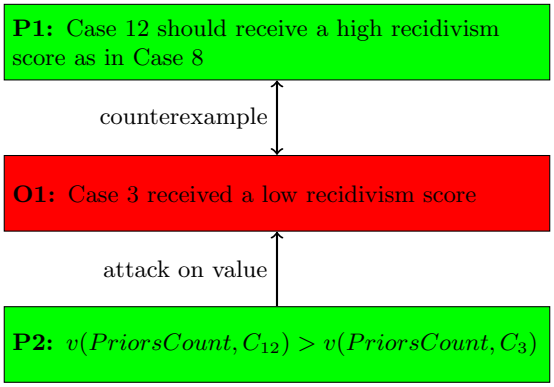

Figure 61. Argument Game for case 12.

	ANALYSIS	
`AgeCat`	x > 45	pro
`Sex`	M	neutral
`JuvenilePriors`	No	pro
`PriorsCount`	0	pro
`ChargeDegree`	F	con
`LenghtOfStay`	x < 50	pro
`Race`	Asian	neutral

Table 47. Dimensional Analysis for case 13.

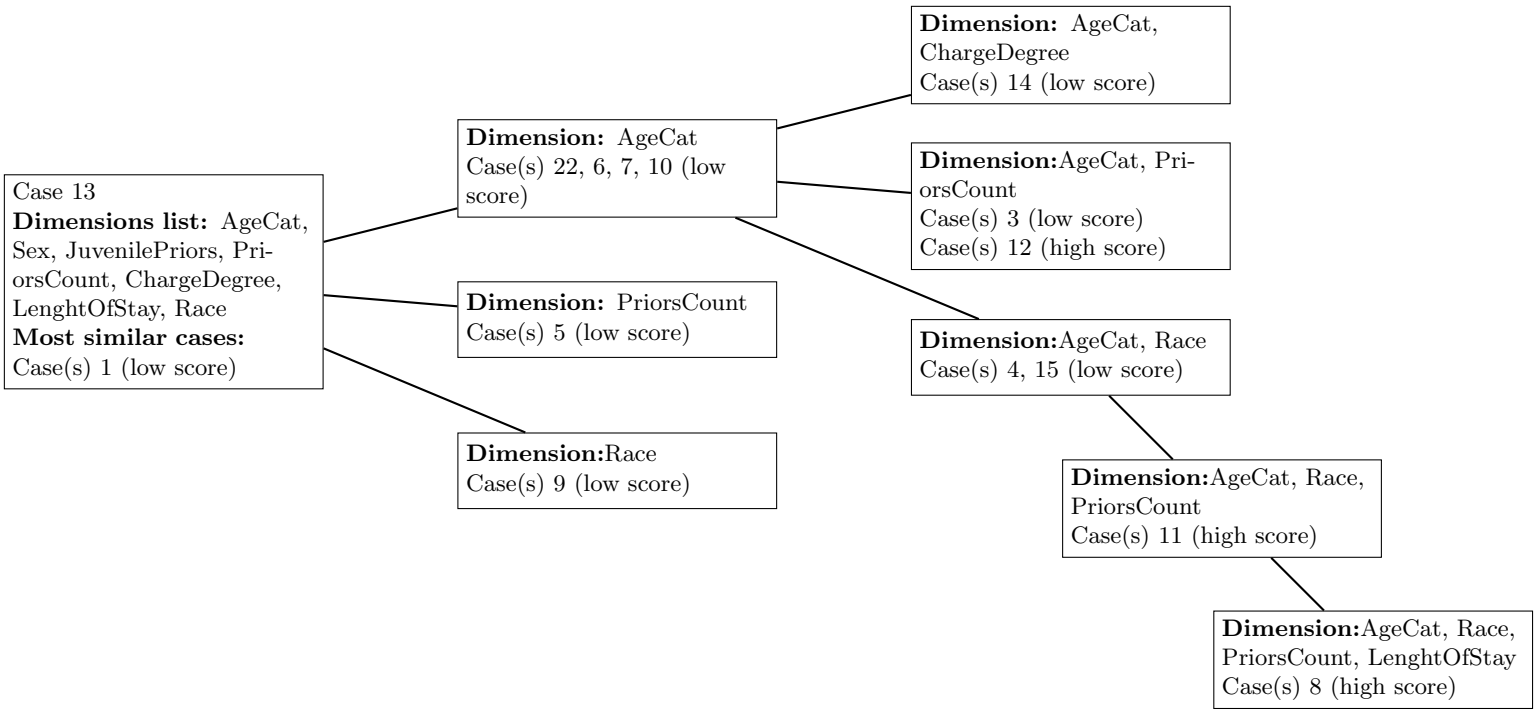

Figure 62. Claim Lattice for case 13.

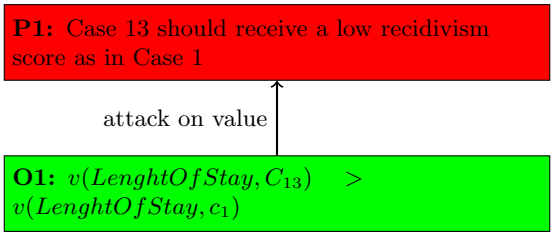

Figure 63. Argument Game for case 13.

ANALYSIS		
AgeCat	25-45	neutral
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	2	neutral
ChargeDegree	F	con
LenghtOfStay	x < 50	pro
Race	other	neutral

Table 48. Dimensional Analysis for case 16.

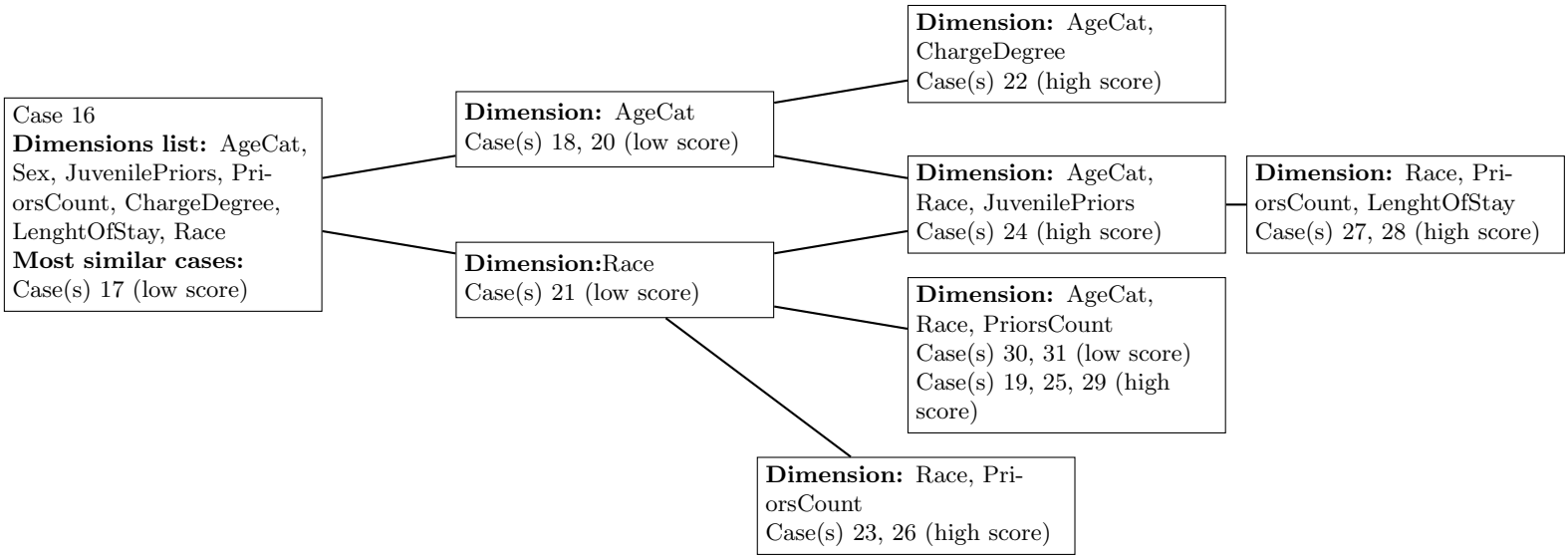

Figure 64. Claim Lattice for case 16.

P1: Case 16 should receive a low recidivism score as in Case 17

Figure 65. Argument Game for case 16.

ANALYSIS		
AgeCat	25-45	neutral
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	3	neutral
ChargeDegree	F	con
LenghtOfStay	x < 50	pro
Race	other	neutral

Table 49. Dimensional Analysis for case 17.

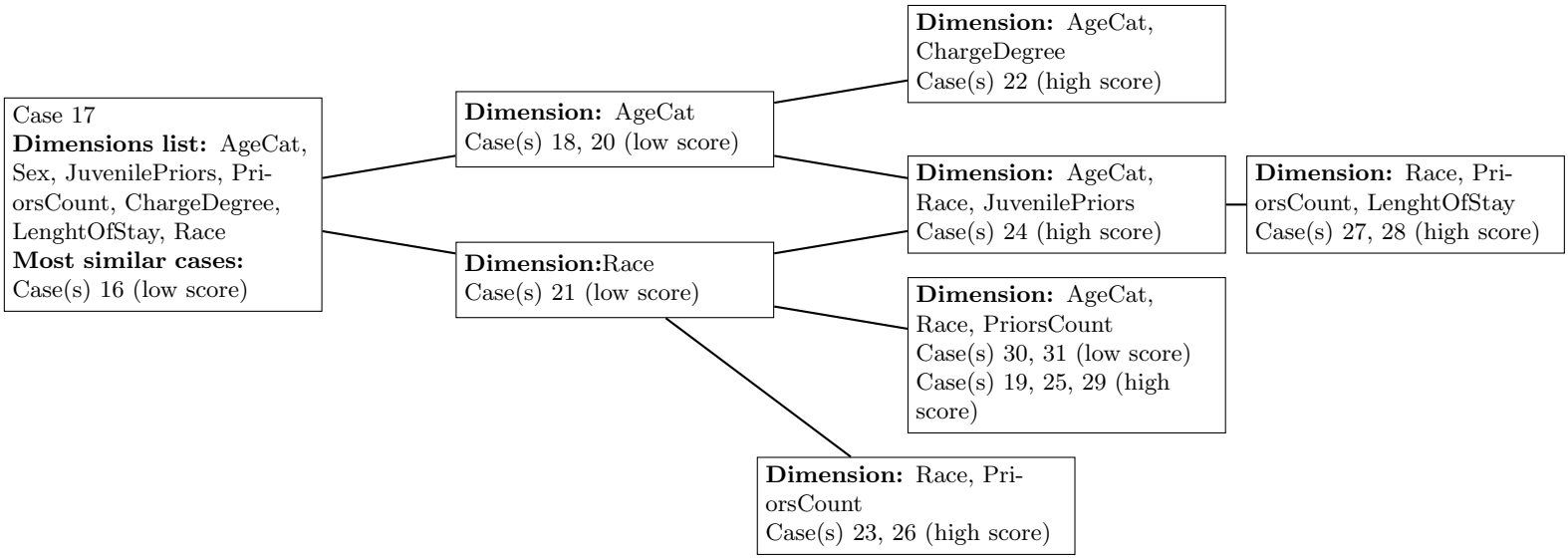

Figure 66. Claim Lattice for case 17.

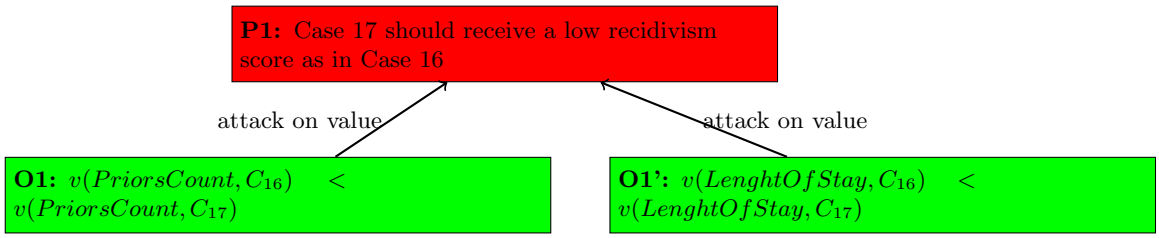

Figure 67. Argument Game for case 17.

ANALYSIS		
AgeCat	> 45	pro
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	13	con
ChargeDegree	F	con
LenghtOfStay	x < 50	pro
Race	African-American	con

Table 50. Dimensional Analysis for case 19.

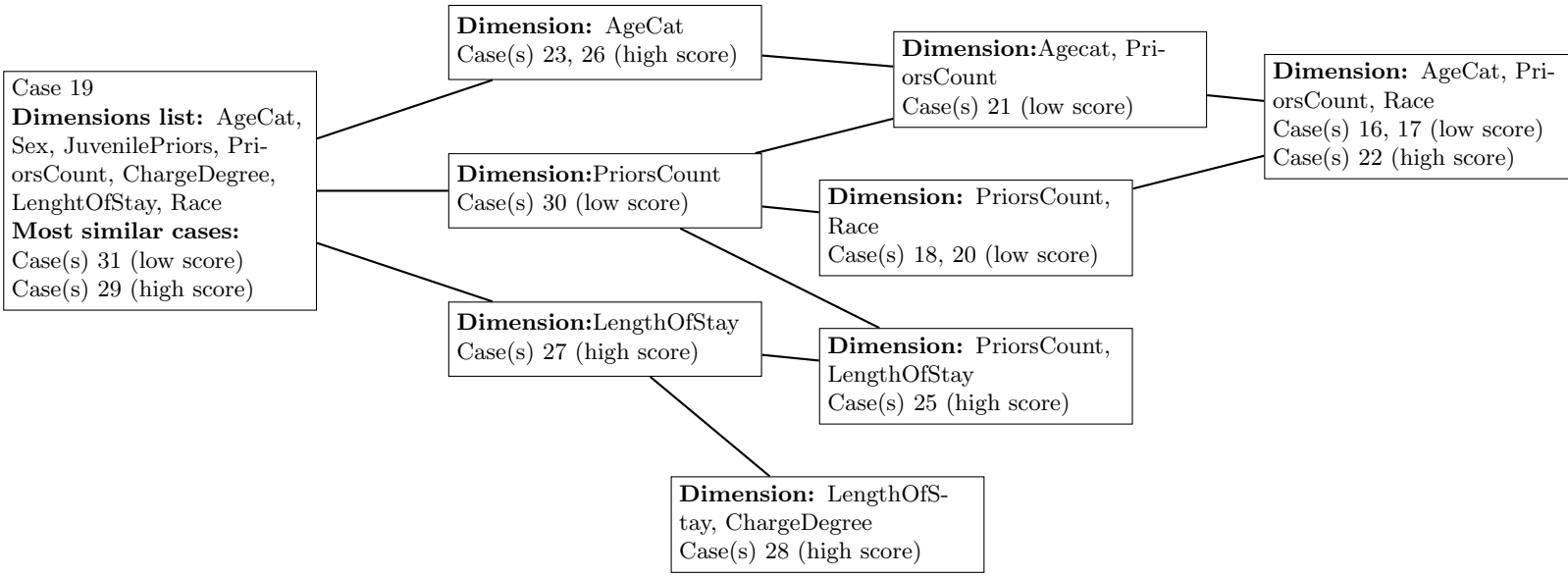

Figure 68. Claim Lattice for case 19.

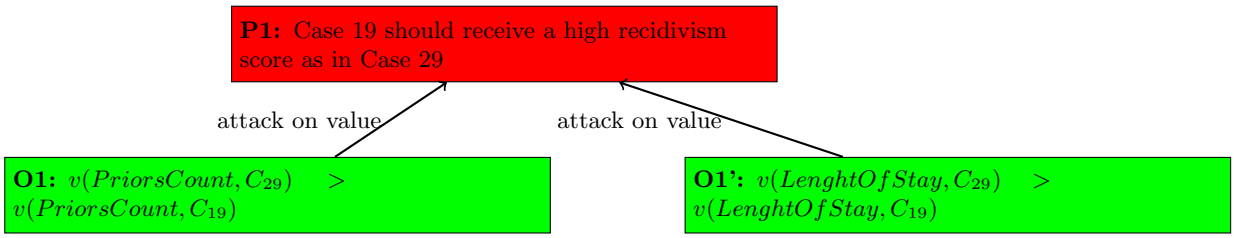

Figure 69. Argument Game for case 19.

ANALYSIS		
AgeCat	x < 25	pro
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	1	neutral
ChargeDegree	F	con
LenghtOfStay	x < 50	pro
Race	Caucasian	neutral

Table 51. Dimensional Analysis for case 22.

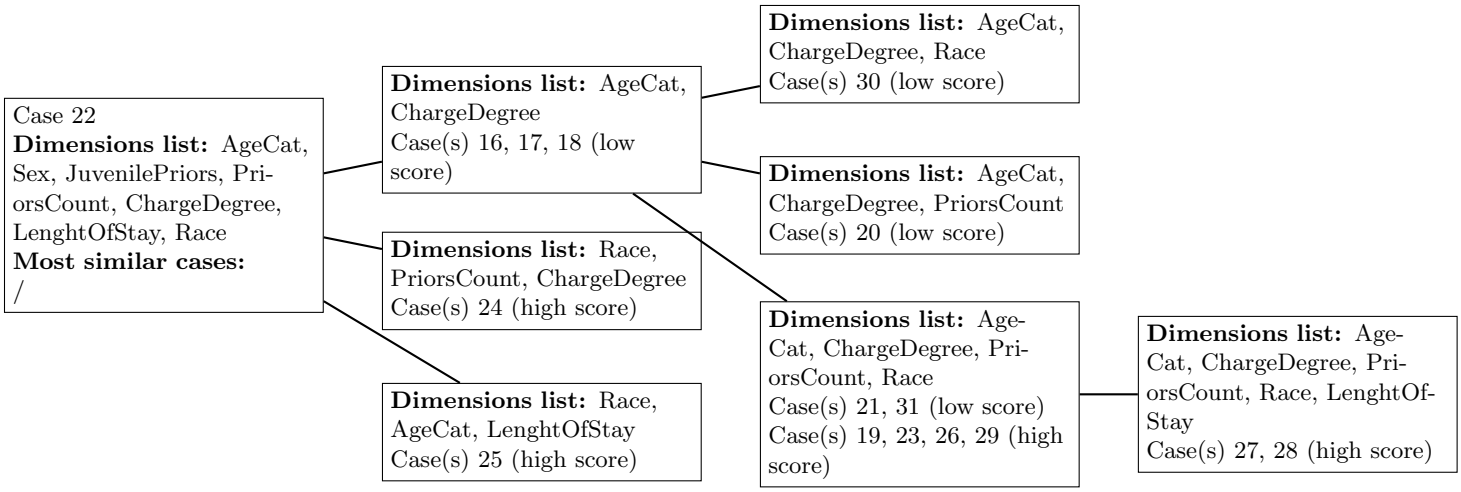

Figure 70. Claim Lattice for case 22.

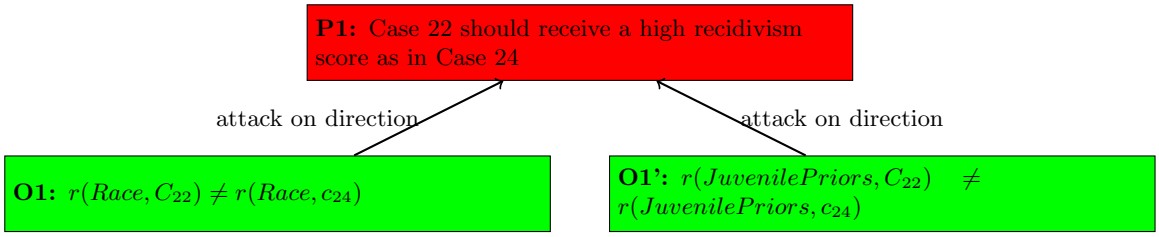

Figure 71. Argument Game for case 22.

	ANALYSIS	
AgeCat	25-45	neutral
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	x > 4	con
ChargeDegree	F	con
LenghtOfStay	x < 50	pro
Race	African-American	con

Table 52. Dimensional Analysis for case 23.

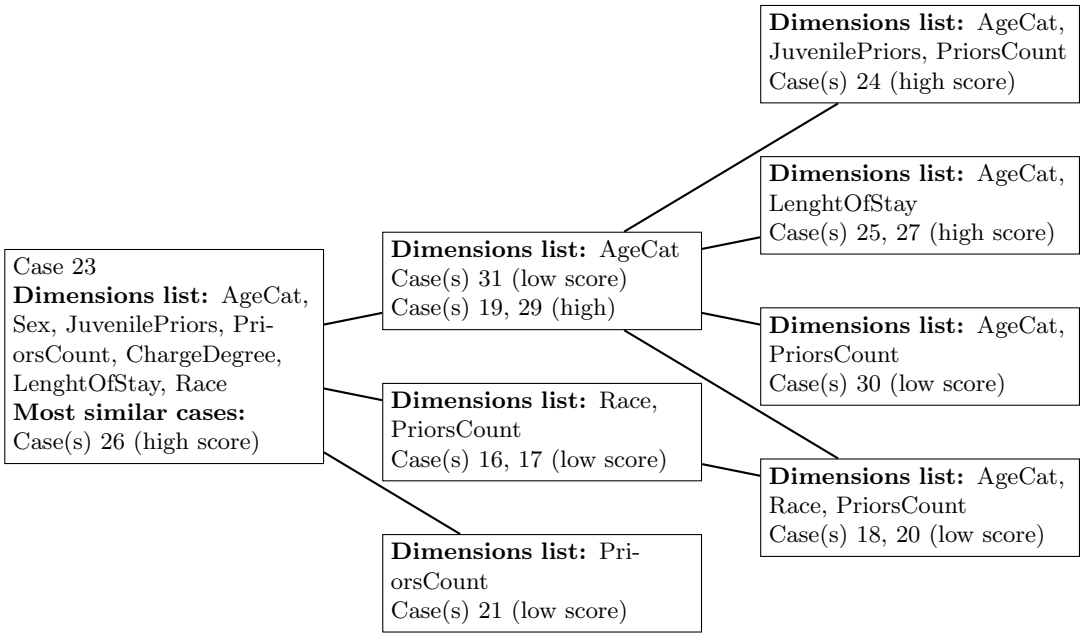

Figure 72. Claim Lattice for case 23.

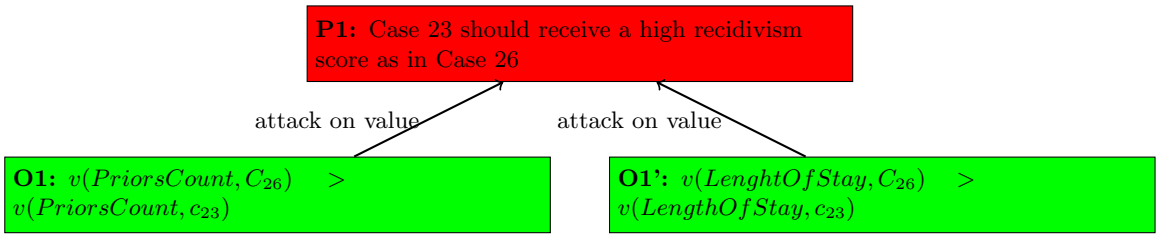

Figure 73. Argument Game for case 23.

ANALYSIS		
AgeCat	x > 45	pro
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	3	neutral
ChargeDegree	F	con
LenghtOfStay	50 < x < 500	neutral
Race	African-American	con

Table 53. Dimensional Analysis for case 25.

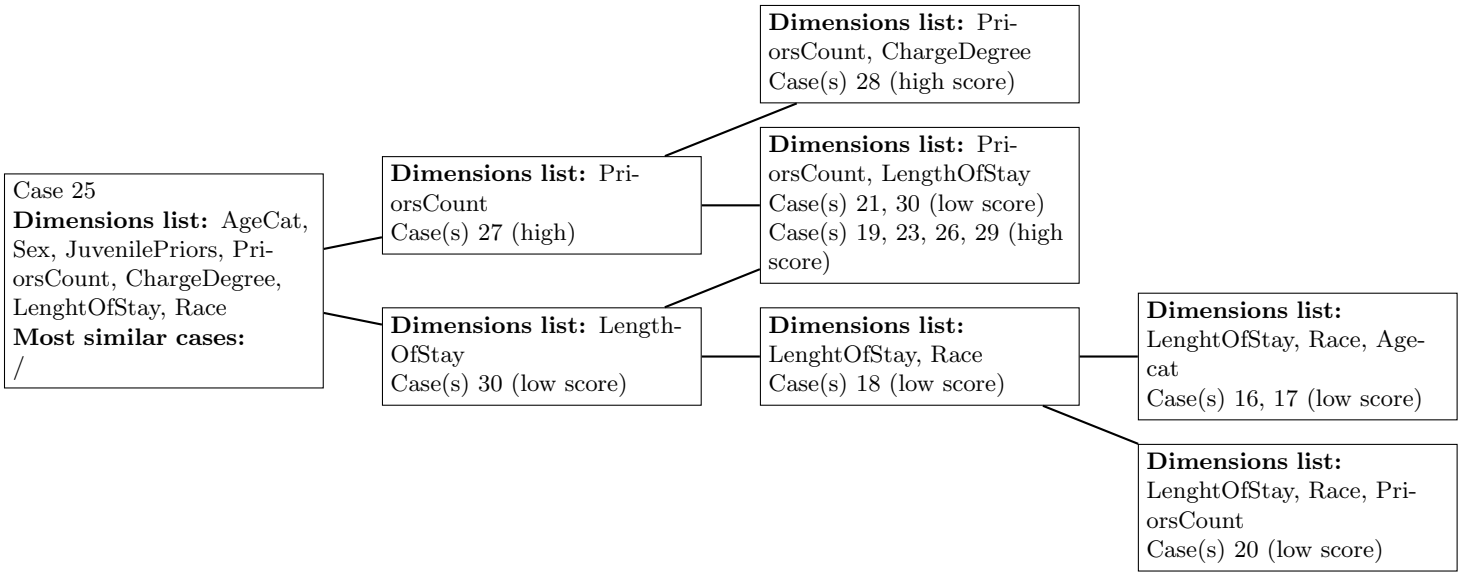

Figure 74. Claim Lattice for case 25.

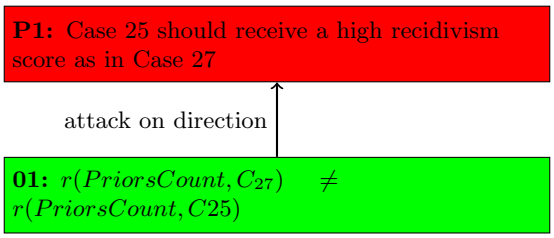

Figure 75. Argument Game for case 25.

	ANALYSIS	
AgeCat	25-45	neutral
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	x > 4	con
ChargeDegree	F	con
LenghtOfStay	x < 50	pro
Race	African-American	con

Table 54. Dimensional Analysis for case 26.

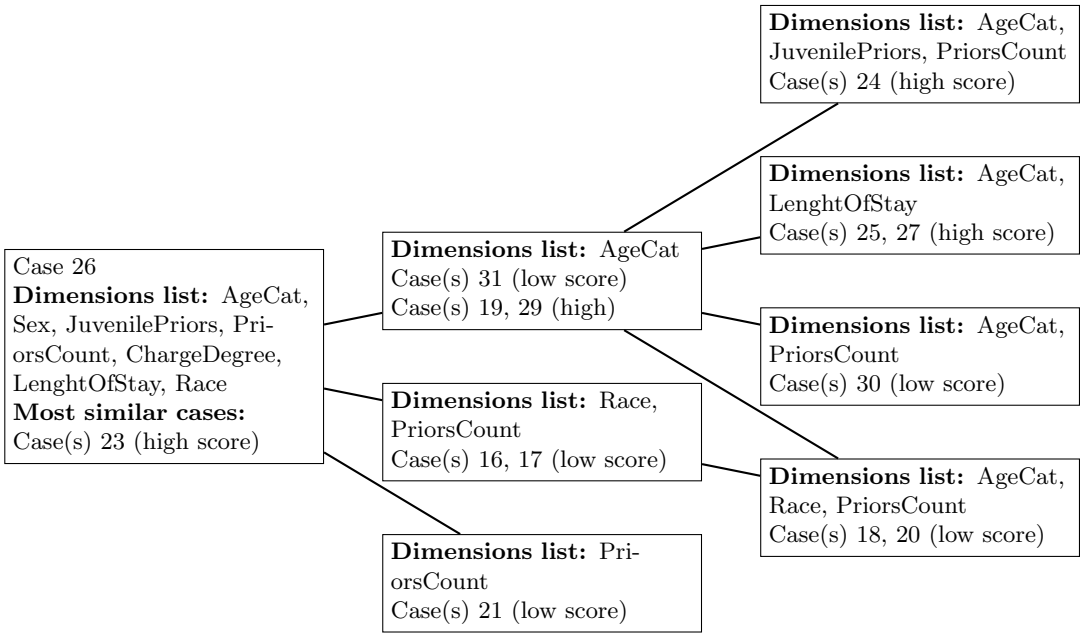

Figure 76. Claim Lattice for case 26.

P1: Case 26 should receive a high recidivism score as in Case 23

Figure 77. Argument Game for case 26.

ANALYSIS		
AgeCat	> 45	pro
Sex	M	neutral
JuvenilePriors	No	pro
PriorsCount	21	con
ChargeDegree	M	neutral
LenghtOfStay	50 <x < 500	neutral
Race	African-American	con

Table 55. Dimensional Analysis for case 29.

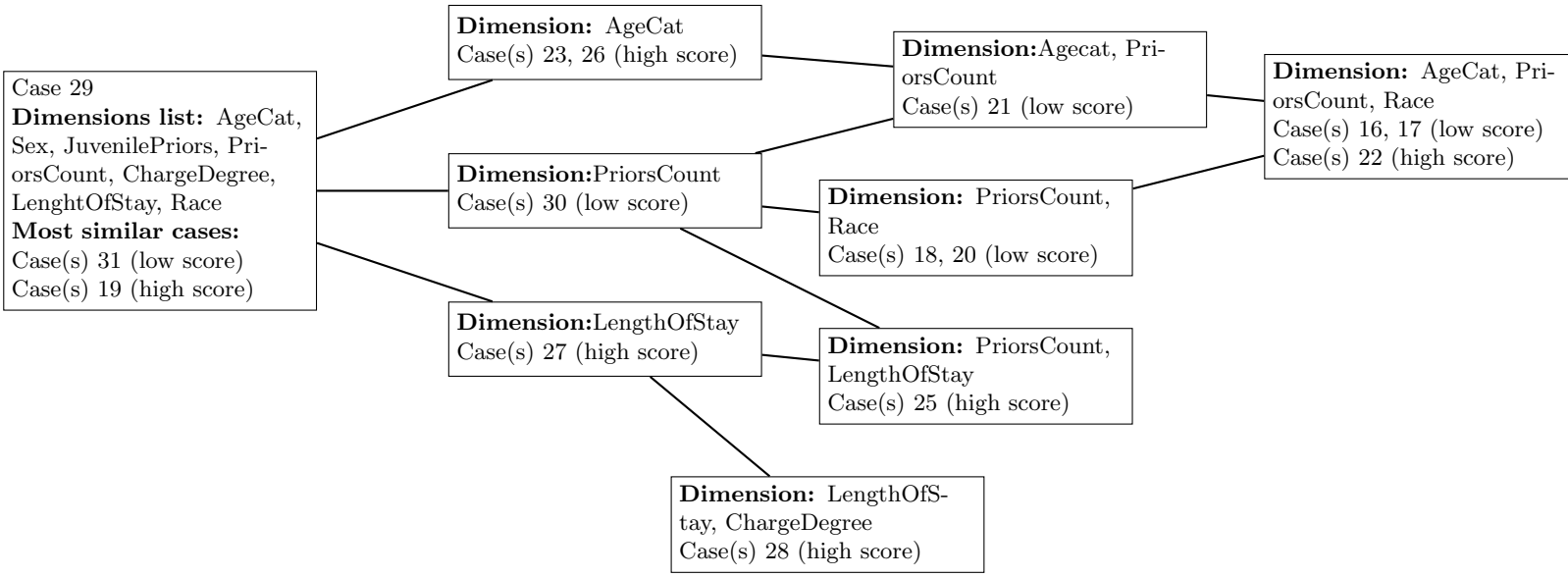

Figure 78. Claim Lattice for case 29.

P1: Case 29 should receive a high recidivism score as in Case 19

Figure 79. Argument Game for case 29.

ANALYSIS		
AgeCat	> 45	pro
Sex	F	neutral
JuvenilePriors	No	pro
PriorsCount	5	con
ChargeDegree	F	con
LenghtOfStay	x < 50	pro
Race	African-American	con

table 56. Dimensional Analysis for case 31.

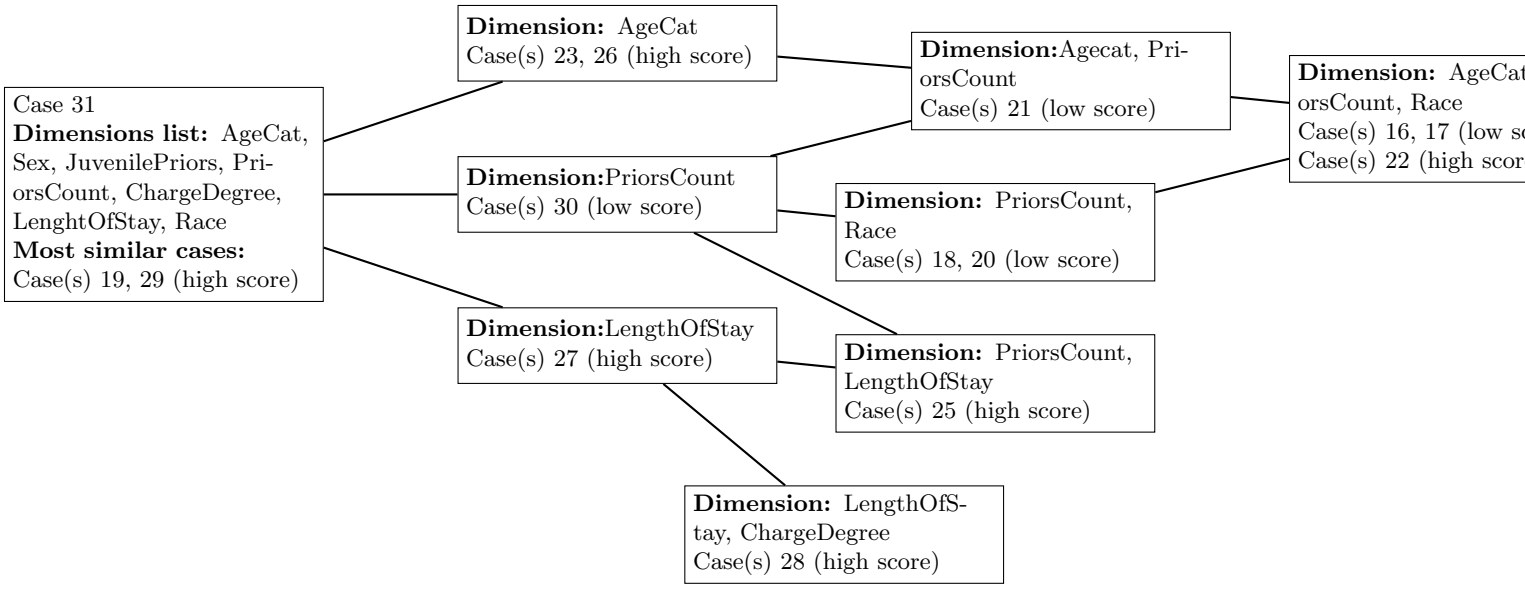

Figure 80. Claim Lattice for case 31.

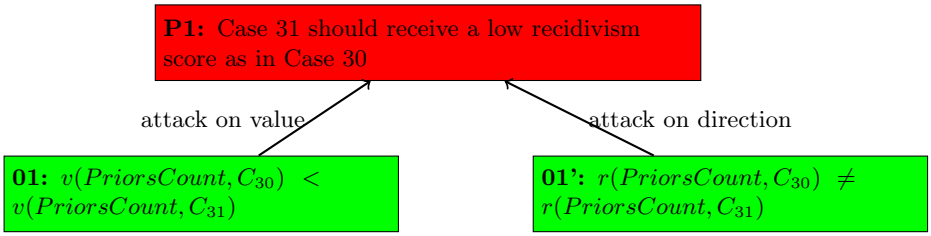

Figure 81. Argument Game for case 31.

References

Angwin, J., Larson, J., Mattu, S., Kirchner, L., ProPublica. (2016, May 23). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks, How we analyzed the COMPAS recidivism algorithm.

Aleven, V. (2003). Using background knowledge in case-based legal reason- ing: a computational model and an intelligent learning environment. *Artificial Intelligence*, 150(1-2), 183-237.

Ashley, K. D. (1991). Reasoning with cases and hypotheticals in HYPO. *International Journal of Man-Machine Studies*, 34(6), 753-796.

Bench-Capon, T. J. (1993). Neural networks and open texture. In *Proceedings of the International Conference on Artificial intelligence and Law* (pp. 292-297). ACM.

Bench-Capon, T. J., & Rissland, E. L. (2001). Back to the Future: Dimensions Revisited. In Verheij, B., Lodder, A.R., Ronald, L. P., & Muntjewerff, A.J., *Legal Knowledge and Information Systems. Jurix 2001: The Fourteenth Annual Conference* (pp. 41-52). IOS Press.

Bench-Capon, T. J. (2017). Arguing with Dimensions in Legal Cases. In *CMNA@ ICAIL* (pp. 2-6).

Feteris, E., & Kloosterhuis, H. (2011). Law and argumentation theory: theoretical approaches to legal justification, in van Klink, B., & Taekema, S., *Law and Method*, Tübingen: Mohr Siebeck.

Friscione, E. (2018). Understanding the machines: argumentation as a tool for interpretability. Utrecht University.

Gordon, T. F., & Walton, D. (2009). Legal reasoning with argumentation schemes. In *Proceedings of the International Conference on Artificial Intelligence and Law* (pp. 137-146). ACM.

Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A survey of methods for explaining black box models. *arXiv preprint arXiv:1802.01933*.

Grabmair, M. (2017). Predicting trade secret case outcomes using argument schemes and learned quantitative value effect tradeoffs. In *Proceedings of the International Conference on Articial Intelligence and Law* (pp. 89-98). ACM.

Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). Accountable algorithms. *U. Pa. L. Rev., 165*, 633.

Lipton, Z. C. (2016). The mythos of model interpretability. In *2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, New York, NY, USA.

Miller, T. (2017). Explanation in artificial intelligence: insights from the social sciences. *arXiv preprint arXiv:1706.07269*.

Modgil, S., & Caminada, M. (2009). Proof theories and algorithms for abstract argumentation frameworks. In *Argumentation in artificial intelligence* (pp. 105-129). Springer, Boston, MA.

Prakken, H., & Sartor, G. (2004). The three faces of defeasibility in the law. *Ratio Juris, 17*(1), 118-139.

Prakken, H. (2017). Commonsense reasoning and argumentation, course manual. Utrecht University.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107(2), 358.

Twining, W., & Miers, D. (1999). *How to Do Things with Rules: a Primer of Interpretation*. Cambridge University Press.

Van Camp, W. (2014). Explaining understanding (or understanding explanation). *European Journal for Philosophy of Science, 4*(1), 95-114.