# Model- and data-agnostic justifications with *A Fortiori* Case-Based Argumentation

JOERI G.T. PETERS, Utrecht University, The Netherlands and Netherlands National Police, The Netherlands

FLORIS J. BEX, Utrecht University, The Netherlands and Tilburg University, The Netherlands

HENRY PRAKKEN, Utrecht University, The Netherlands and University of Groningen, The Netherlands

AF-CBA is an example-based approach to XAI that draws on the case-based argumentation tradition in AI & Law. It means to explain binary classifications made by an opaque machine-learning model by presenting an argument graph to the user, which represents an argument game about the classification of a case on the basis of precedents derived from labelled data used in the training phase of the classifier. We improve the robustness of this method by modifying it to better handle inconsistent labelling and evaluate an alternative setup that does not require access to the labelled data by using earlier predictions instead.

**ACM Reference Format:**

Joeri G.T. Peters, Floris J. Bex, and Henry Prakken. 2023. Model- and data-agnostic justifications with *A Fortiori* Case-Based Argumentation. 1, 1 (June 2023), 18 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

The accuracy of some machine learning (ML) classifiers comes at the cost of their transparency [19]. For certain purposes, transparent alternatives may not be able to achieve sufficient accuracy to be a viable option. In many cases, perceived classifier opacity is due to technical complexity, but this is relative to a person's level of understanding, so even very basic approaches would be considered opaque by some. Proprietary protection can be another cause of opacity and renders even an otherwise highly interpretable approach opaque. Regardless of the underlying reason, highly opaque classifiers are commonly known as 'black boxes' [13, 19].

A central concern with black-box models is their trustworthiness. There may be ethical concerns such as unfair treatment and biases that remain hidden with an opaque model. Transparency is often a legal requirement for (semi-)automatic decision-making processes in practice [6], even if the decisions are not within the legal domain themselves. For this reason, improving transparency is essential to the AI & Law domain. Explainable Artificial Intelligence (XAI) is aimed at increasing the transparency of black-box models [20].

In this paper, we are concerned with example-based XAI, which is one of the lines of research within XAI [4, 22]. XAI methods can be categorised in various ways. One distinction is between methods that generate local explanations (explaining individual instances) and those that generate global explanations (explaining a whole model). Some methods access the trained model itself, whilst others are *model-agnostic*. Approaches that generate explanations after the fact are also known as '*post hoc analyses*' [19]. We use the term '*justifications*' for the subclass of explanations generated by

Authors' addresses: Joeri G.T. Peters, j.g.t.peters@uu.nl, Utrecht University, Utrecht, The Netherlands and Netherlands National Police, Driebergen, The Netherlands; Floris J. Bex, f.j.bex@uu.nl, Utrecht University, Utrecht, The Netherlands and Tilburg University, Tilburg, The Netherlands; Henry Prakken, h.prakken@uu.nl, Utrecht University, Utrecht, The Netherlands and University of Groningen, Groningen, The Netherlands.

model-agnostic XAI methods which do not explain a model's actual behaviour. Instead, they present to a human user the assumptions under which the model's decision can be justified.

When one considers justifying binary class labels, the predictions of the classifier (trained on labelled data in its training phase) can be thought of as analogous to court decisions on the basis of judicial precedents. This is why, in order to explain predictions made by such a classifier, Prakken & Ratsma [27] draw on AI & Law research to propose a top-level model using case-based argumentation (CBA) based on Horty's model of *a fortiori* reasoning [16], hereafter referred to as 'A Fortiori Case-Based Argumentation' (AF-CBA). AF-CBA is inspired by CATO [2] and work by Čyras et al. [9, 10]. In the present work, we improve the usability of AF-CBA in two respects, namely the ability to handle label inconsistency[1] and the ability to explain predictions without having access to the actual training data.

The context of AF-CBA is depicted in Figure 1. The labelled dataset (or $X$ in most ML literature) is a random sample from the overall population of a sufficient size to make it representative of the population, labelled by annotators or decision makers. $X$ is used in the training phase of a classification approach to produce a classifier. A *focus case* is a single, random sample case from the same population. It receives a predicted outcome (some label $s$) from the classifier. Because the classifier is a black box, it cannot provide an explanation for why it came to the decision to predict outcome $s$. AF-CBA justifies its outcome by initiating the labelled set $X$ as a case base and using it to play an argument game between a proponent and opponent of the outcome $s$. A winning strategy for the proponent is presented as a justification of the predicted outcome $s$ in the form of an argument graph.

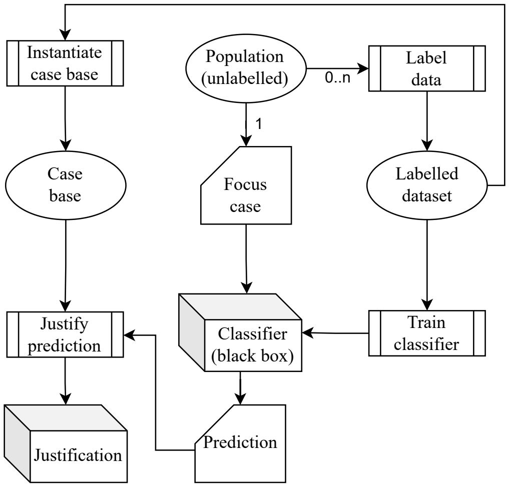

Fig. 1. A schematic depiction of the context in which AF-CBA is applied.

In this paper, we investigate two limitations of AF-CBA and the extent to which we can circumvent them, thereby allowing AF-CBA to be more widely usable. The first limitation is that, as a consequence of its reliance on precedential constraint, label inconsistency is a significant concern for AF-CBA [27]. The second limitation is that AF-CBA presumes access to the labelled dataset used in training the black-box classifier, which is often not the case, i.e. it is not *data-agnostic*. We demonstrate a solution to the first limitation by constraining AF-CBA's selection of best precedents and to the second by implementing a data-agnostic alternative.

[1] This portion of the current paper is a further investigation based on preliminary work [25], using additional datasets and performance metrics.

The rest of this paper is structured as follows. We formally describe AF-CBA in Section 2. We address the problem of inconsistency in Section 3 and that of data-agnosticism in Section 4. Subsequently, we experiment with these modifications in Section 5 and discuss the results and future work in Section 6. Finally, we consider some related work in Section 7.

## 2 AF-CBA

AF-CBA justifies predictions by referring to cases which are highly similar to the one whose class is being predicted (the focus case), relying on Horty's notion of precedential constraint [15, 16] used to model a fortiori reasoning as used with case law. This requires no knowledge of ML to understand. It does, however, require a body of case law. ML classification is a supervised approach, which means there is a training set that was used to train the classifier. AF-CBA requires that this set be accessible to be used as a case base. The underlying a fortiori assumption is the following notion of precedential constraint: the focus case should have the same outcome (label) as a precedent if differences between the two only make the focus case stronger for that same outcome [27].

AF-CBA produces an argument graph through an argument game, which has a fixed set of allowed moves inspired by HYPO [3] and CATO [2]. These moves are modelled as an abstract argumentation framework in the sense of Dung [11]. The argument game is modelled as the game for grounded semantics of abstract argumentation frameworks [26]. A proponent argues why the focus case should receive the same outcome as a *best precedent* (a most similar case) and the opponent argues against this. They both cite precedents from the CB and make moves to set cases apart or to downplay these differences. Deciding for a focus case is *forced* if the precedent has no relevant differences with the focus case [27].

We will now formally introduce these and other concepts involved in AF-CBA, largely identical to those found in [27], with some differences in notation. As a running example, we rely on a feature subset of the Telco Customer Churn dataset [17], which describes the customers of a telecommunications provider and whether they have churned, i.e. switched to an alternative provider. This is valuable information, because the provider might wish to take action when a customer is likely to churn, such as offering a discount. If this occurs automatically, it is a case of automatic decision-making, which implies that the transparency requirements of the General Data Protection Regulation should apply [32].

A *case* is a member of a *case base* (CB) and consists of an *outcome* as well as a *fact situation.* The outcome of a case is simply a binary label, $o$ or $o'$. The variables $s$ and $\bar{s}$ denote the two sides, meaning that $s = o$ if $\bar{s} = o'$ and vice versa. A fact situation consists of dimensions (features), with each dimension a tuple $d = (V, \leq_o, \leq_{o'})$, with value set $V$ and two partial orderings on $V$, $\leq_o$ and $\leq_{o'}$, such that $v \leq_o v'$ iff $v' \leq_{o'} v$ for $v, v' \in V$.

A dimension has a *tendency*, where a positive tendency means a higher value assignment for that dimension is associated with one outcome (e.g. 1 or $\top$) and vice versa for the other. Table 1 illustrates the dimensions in our running example, showing the optional superscript plus or minus notation to reflect a dimension's tendency. Three of these dimensions have a negative tendency and only a higher value assignment for *high cost* is associated with a customer churning.

A value assignment is denoted as a pair $(d, v)$ and the value $x$ of dimension $d$ as $v(d, c) = x$ for case $c \in CB$. The value assignments to all dimensions $d$ of the non-empty set $D$ constitute a fact situation $F$. We presume that two fact situations refer to the same set $D$. A case is defined as $c = (F, outcome(c))$ with $outcome(c) \in \{o, o'\}$, and we can denote the fact situation of case $c$ as $F(c)$.

In Table 2, Alice currently is and Bob used to be a customer of the telecommunications provider, as is seen by the outcome. Their fact situations are provided as well, with value assignments to each of the four dimensions. Charlie (the

Table 1. The dimensions used in the running example.

| Dimension | Name | Description |
|---|---|---|
| $d_1^-$ | Gift | Whether the customer received a gift |
| $d_2^-$ | Present | Whether the customer was present during a recent event |
| $d_3^-$ | Website | The number of times a customer logged into their online profile |
| $d_4^+$ | High cost | Whether the customer has a subscription in the high-cost category |

focus case) is an additional customer and the provider has a classifier in place to predict whether Charlie will churn. Say that the classifier predicts that he will not. Our approach is then to explain that outcome on the basis of Charlie's case's similarity to the other two cases.

Table 2. An example CB labelled with the outcome churn.

| Customer | $d_1^-$ | $d_2^-$ | $d_3^-$ | $d_4^+$ | Outcome |
|---|---|---|---|---|---|
| Alice | 1 | 1 | 3 | 0 | 0 |
| Bob | 1 | 0 | 1 | 1 | 1 |
| Charlie (focus) | 0 | 1 | 3 | 0 | ? |

Given two fact situations and the tendencies of their dimensions, one fact situation may be 'stronger' for a particular outcome than the other. The outcome of a focus case is *forced* if there exists a precedent in the CB with the same outcome for which all differences between the focus case and that precedent merely make the focus case an even stronger case for that very outcome [15].

Definition 1 (Preference relation for fact situations). *Given two fact situations $F$ and $F'$, $F \leq_s F'$ iff $v \leq_s v'$ for all $(d, v) \in F$ and $(d, v') \in F'$.*

Definition 2 (Precedential constraint). *Given case base CB and fact situation $F$, deciding $F$ for $s$ is forced iff CB contains a case $c = (F', s)$ such that $F' \leq_s F$.*

In Table 2, Alice's case is stronger than Charlie's for the outcome not churning when it comes to $d_1^-$ and $d_3^-$. Alice's case therefore does not force the focus case. If Alice were instead the focus case and Charlie the precedent, Alice's case would be stronger than Charlie's for churning in all dimensions and so Charlie would indeed force the outcome. In other words, we would then say that if Charlie did not churn, surely Alice must not churn either.

A fact situation could be forced for both $s$ and $\bar{s}$, which brings us to Horty's definition of CB consistency:

Definition 3 (Case base consistency). *A case base CB is consistent iff it does not contain two cases $c = (F, s)$ and $c' = (F', \bar{s})$ such that $F \leq_s F'$. Otherwise it is inconsistent.*

A best precedent and a focus case not only have the same outcome, but also as few as possible *relevant differences*. Multiple cases can meet these criteria. A lower number of best precedents is preferable, because of computational reasons and because one could say that a higher number of possible citations would make a single explanation somewhat arbitrary.

Definition 4 (Differences between cases). *Let* $c = (F(c), outcome(c))$ *and* $f = (F(f), outcome(f))$ *be two cases. The set* $D(c, f)$ *of differences between* $c$ *and* $f$ *is* $D(c, f) = \{(d, v) \in F(c) \mid v(d, c) \not\leq_{outcome(f)} v(d, f)\}$.

Definition 5 (Best precedent). *Let* $c = (F(c), outcome(c))$ *and* $f = (F(f), outcome(f))$ *be two cases, where* $c \in CB$ *and* $f \notin CB$. $c$ *is a best precedent for* $f$ *iff:*

- $outcome(c) = outcome(f)$ *and*
- *there is no* $c' \in CB$ *such that* $outcome(c') = outcome(c)$ *and* $D(c', f) \subset D(c, f)$.

For instance, a relevant difference between Charlie and Alice in Table 2 is dimension $d_1^-$, where Alice received a gift and Charlie did not, making her case better for staying rather than churning. Alice would be selected as a best precedent, as Bob received the opposite outcome.

The opponent is looking to reply to the proponent's initial citation, either by citing a counterexample or by playing a distinguishing move. The distinguishing moves are $Worse(c, x)$ (the focus case is worse than the precedent $c$ for dimensions $x$), $Compensates(c, x, y)$ (the dimensions $y$ compensate for the dimensions $x$ on which the focus case is not at least as good as the precedent $c$) and $Transformed(c, c')$ (the citation can be transformed by the distinguishing moves into a case for which $D(c, f) = \emptyset$). For the sake of brevity, see [27] for formal motivations of these moves and the need to allow the *Compensates* move to be empty in order to state that the differences with the focus case do not matter. The proponent is then able to reply in turn with these same moves, and so on, until the opponent cannot make any further moves.

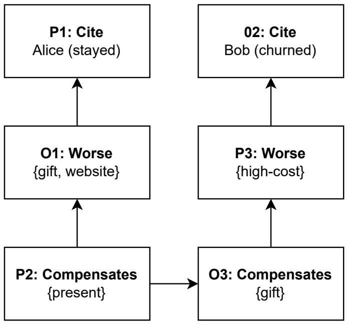

Fig. 2. A fictional example of an explanation (dialogue between proponent and opponent).

Returning to our example, Figure 2 presents the resulting explanation as an argument game, which can be read as follows. P1: Alice stayed and her case is similar to Charlie's. O1: Charlie's scores for $d_1^-$ and $d_3^-$ make him worse for staying than Alice. P2: Charlie's score for $d_2^-$ compensates for O1. O2: Bob churned and his case is similar to Charlie's. P3: Charlie's score for $d_4^+$ makes him worse for churning than Bob. O3: Charlie's score for $d_1^-$ compensates for P3. P2: Charlie's score for $d_2^-$ compensates for O3. After this, the opponent has run out of possible moves to make and the proponent wins. The similarity to Alice's case has held up and acts as an explanation for the prediction that Charlie will stay as well.

When the proponent cites a precedent for which there are no relevant differences with the focus case, the opponent cannot distinguish the focus case from that precedent. If the CB does not contain a case that has no relevant differences with the focus case but with the opposite outcome (which would be a source of inconsistency), the argument game simply ends, since the opponent cannot cite a counterexample. If the CB does contain such a case, the opponent cites it as a counterexample which also has no relevant differences, to which the proponent responds with an empty downplaying move. In either scenario, we refer to this as a trivial winning strategy (Definition 6) and to the corresponding focus case as a trivial case, as opposed to a non-trivial winning strategy and a non-trivial case.

Definition 6 (Trivial winning strategies). *Given a case base CB and a focus case $f$, $f$ has a trivial winning strategy iff there is a best precedent $c \in CB$ for which $D(c, f) = \emptyset$. When $f$ has at least one trivial winning strategy, it is considered to be a trivial case.*

## 3 CB INCONSISTENCY

The labels in labelled training data are typically gathered through annotation of unlabelled data. Although annotators (who label data to this end) produce a labelled dataset specifically for the purpose of training a model, they may not necessarily be fully consistent when doing so [21]. Annotators might disagree or make an occasional mistake, leading to label inconsistency. Labels may also be produced by decision makers as part of some decision process. Take for example judges who decide on court cases, with verdicts being stored as case law, which can contain conflicting opinions and interpretations. Finally, the feature vector may be but a subset of relevant details that influenced a decision, thereby potentially lacking necessary data to discriminate between seemingly similar cases [12]. These sources of noise may make the labelling inconsistent. The exact same feature vector may be labelled with class label 0 in one instance and class label 1 in another.

Some label inconsistency is to be expected in many situations in practice. If a CB corresponds to the training data, as in AF-CBA, it follows that the CB is likewise expected to contain inconsistencies. CB inconsistency is however not limited to the conflict between identical feature vectors with opposing labels. Given Horty's a fortiori assumption [16], a case which is at least as good as another yet receives the opposite outcome is a source of inconsistency. If feature vector $A$ can be said to be more strongly associated with outcome 1 than feature vector $B$, which is indeed labelled as 1, it would be inconsistent with this assumption if $A$ received outcome 0 instead. Thus, CB inconsistency is a broader notion than label inconsistency and is all the more to be expected.

AF-CBA does not strictly require that the CB be consistent, but it is not entirely robust in its handling of CB inconsistency. A source of inconsistency is typically a somewhat exceptional case with a surprising outcome, such as when one case $c_1$ is at least as good in all dimensions for outcome $s$ as another case $c_2$, yet received outcome $\bar{s}$. Citing such a case as a precedent can lead to the focus case being forced for both outcomes ('*inconsistent forcing*'). This results in an explanation that hinges on the acknowledgement of the CB's inconsistency. It can be argued that this is unsatisfactory and not ideal when attempting to raise the transparency of a black-box model. The larger the number of inconsistent forcings ($N_{inc}$), the larger the number of explanations where this problem occurs.

In experiments by Prakken & Ratsma [27], significant portions of CBs had to be ignored (by removing a minimal number of cases when instantiating the CB) in order to make them consistent. This is unfortunate, as it weakens the merit of justifications and makes the whole approach less transparent. We would prefer to use the whole training set as a CB. This problem is exasperated by feature selection techniques, which would otherwise help keep explanations simple. In conclusion, CB consistency forms a problematic constraint.

We present a modification of AF-CBA that takes into account the degree to which the CB consistently supports a precedent. We refer to this measure as the 'authoritativeness' of a precedent. Using authoritativeness prevents inconsistent forcing by modifying the selection of best precedents to cite. We experiment with several possible alternatives of quantifying the authoritativeness and demonstrate that it has a positive effect on AF-CBA without adversely affecting its explanations.

Instead of mitigating the problem through case deletion [27], we explicitly take inconsistencies into account. Informally, one might say that when there is consistency, a precedential case has a strong backing when cited and should indeed immediately force the outcome; if there is inconsistency, it has less backing and thus should not. We therefore introduce the concept of *precedential authoritativeness*, by which we mean that, given any case $c \in CB$, the authoritativeness $\alpha(c)$ numerically expresses (normalised between 0 and 1) the degree to which the rest of the CB supports the citing of $c$ for $outcome(c)$. We subsequently use $\alpha(c)$ as an additional criterion in the selection of best precedents. The intuition behind authoritativeness is that whereas the a fortiori rule applied to a consistent CB can be expressed as the phrase 'cases like this always receive outcome $o$,' our idea of authoritativeness changes this phrase to 'cases like this *usually* receive outcome $o$'—where 'usually' has to be quantified in some manner which expresses the inconsistency of the CB with regards to the focus case. Since $\alpha(c)$ is a number, we can have a total ordering $\leq$ on the authoritativeness of cases.

Table 3 is another instance of our Churn example. Depending on how one chooses to define $\alpha(c)$, $c_1$ and $c_2$ should arguably receive a higher value for $\alpha(c)$ than $c_3$ due to $c_4$ having the opposite outcome.

Table 3. Example of a CB with two identical cases that are consistent with each other and two identical cases which contradict each other.

| Customer | $d_1^-$ | $d_2^-$ | $d_3^-$ | $d_4^+$ | $outcome$ |
|---|---|---|---|---|---|
| $c_1$ | 1 | 1 | 0 | 0 | $s$ |
| $c_2$ | 1 | 1 | 0 | 0 | $s$ |
| $c_3$ | 1 | 1 | 5 | 0 | $s$ |
| $c_4$ | 1 | 1 | 5 | 0 | $\bar{s}$ |

First of all, the definition of best precedent has to be modified to reflect the additional criterion of maximising the authoritativeness:

Definition 7. *(Best authoritative precedent) Let CB be a case base and let* $c = (F(c), outcome(c))$ *and* $f = (F(f), outcome(f))$ *be two cases, where* $c \in CB$ *and* $f \notin CB$. $c$ *is a best precedent for* $f$ *iff:*

- $outcome(c) = outcome(f)$,
- *there is no* $c' \in CB$ *such that* $outcome(c') = outcome(c)$ *while* $D(c', f) \subset D(c, f)$ *and* $\alpha(c') \geq \alpha(c)$.

We require expressions of agreement and disagreement between a precedent and the rest of the CB:

Definition 8. *(Agreement) Let CB be a case base. Given* $c \in CB$, *the agreement* $n_a(c)$ *is defined as:*
$n_a(c) = |\ \{c' \in CB \mid outcome(c') = outcome(c) \text{ and } D(c, c') = \emptyset\}\ |$

Definition 9. *(Disagreement) Let CB be a case base. Given* $c \in CB$, *the disagreement* $n_d(c)$ *is defined as:*
$n_d(c) = |\ \{c' \in CB \mid outcome(c') \neq outcome(c) \text{ and } D(c, c') = \emptyset\}\ |$

We understand $n_a(c)$ as the number of cases which have the same outcome as the precedent case and are at least as good for that outcome as $c$. Similarly, $n_d(c)$ is the number of cases which have the opposite outcome yet are at least as good for $outcome(c)$. The agreement $n_a(c)$ has at least a value of 1 due to $c$ itself being a member of the CB. The disagreement $n_d(c)$ can have a value of 0.

Exactly how the level of agreement relates to authoritativeness is not self-evident, as various expressions may have equal merit. For example, given a case $c \in CB$, we could express the authoritativeness $\alpha(c)$ as the relative number of cases which lend further support to $c$ (1). In Table 3, $c_3$ is supported by (other than itself) $c_1$ and $c_2$, but opposed by $c_4$. So in that situation, $\alpha(c_3) = 3/(3+1) = 0.75$.

$$\alpha(c) = \frac{n_a(c)}{n_a(c) + n_d(c)} \tag{1}$$

However, this overlooks any intuitive understanding of authoritativeness which stems from the absolute number of cases that can act as precedents (2). Intuitively, obscure cases are less authoritative than common ones. In Table 4, $c_1$ is supported by two other cases (again, other than itself), namely $c_2$ and $c_3$, while $c_5$ is supported by $c_1$ through $c_4$. We divide by $\mid CB \mid$ to normalise the expression between 0 and 1. So for example $\alpha(c_1) = 3/(3+0) = 1$ according to (1) but $\alpha(c_1) = 3/7 \approx 0.429$ according to (2).

$$\alpha(c) = \frac{n_a(c)}{\mid CB \mid} \tag{2}$$

Table 4. Example of an inconsistent CB showcasing different levels of support.

| Customer | $d_1^-$ | $d_2^-$ | $d_3^-$ | $d_4^+$ | *outcome* |
|---|---|---|---|---|---|
| $c_1$ | 1 | 1 | 0 | 0 | $s$ |
| $c_2$ | 1 | 1 | 0 | 0 | $s$ |
| $c_3$ | 1 | 1 | 0 | 0 | $s$ |
| $c_4$ | 1 | 1 | 2 | 0 | $s$ |
| $c_5$ | 1 | 1 | 2 | 0 | $s$ |
| $c_6$ | 1 | 1 | 2 | 0 | $\bar{s}$ |
| $c_7$ | 1 | 1 | 15 | 0 | $s$ |

Both *relative authoritativeness* (1) and *absolute authoritativeness* (2) would appear to have some merit. Using a combination of the two seems even more intuitive. One option is to create a *product authoritativeness* (3) by taking the product of (1) and (2), essentially using (1) as a weight factor for (2).

$$\alpha(c) = \frac{n_a(c)}{n_a(c) + n_d(c)} \cdot \frac{n_a(c)}{\mid CB \mid} \tag{3}$$

Alternatively, (1) and (2) can be combined as a weighted harmonic mean (4). This *harmonic authoritativeness* introduces a parameter $\beta$, the relative importance of one expression over the other. The added advantage of this is that (1) could be considered twice as important as (2), for instance. At a value of $\beta = 1$ (the unweighted harmonic mean), the two are equally important.

$$\alpha(c) = (1 + \beta^2) \cdot \frac{\frac{n_a(c)}{n_a(c)+n_d(c)} \cdot \frac{n_a(c)}{|CB|}}{\left(\beta^2 \cdot \frac{n_a(c)}{n_a(c)+n_d(c)}\right) + \frac{n_a(c)}{|CB|}} \tag{4}$$

## 4 DATA-AGNOSTICISM

One point of criticism typically aimed at approaches such as AF-CBA is that it is limited to a very particular set of circumstances, namely that of a black-box classifier in combination with an accessible labelled dataset. We encounter such situations in practice at the Netherlands National Police when data scientists have trained a complex classifier

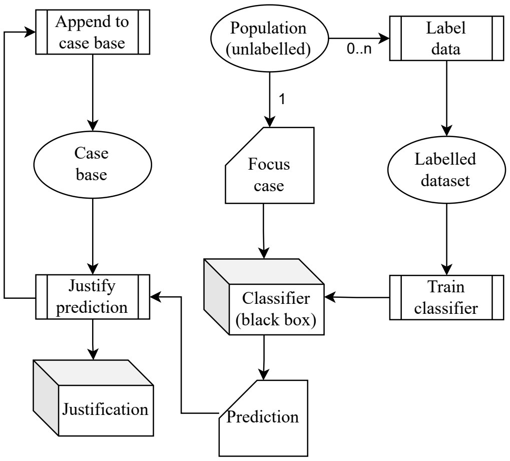

Fig. 3. A schematic depiction of the context in which AF-CBA is applied without requiring access to the labelled dataset ($X$).

(e.g. a neural network) to be used by colleagues, in which case the labelled data is readily available to be used by AF-CBA. However, classifiers can also be considered black boxes because of proprietary protection rather than intrinsic complexity, in which case the labelled data is frequently also kept from clients. Furthermore, classifiers may have been trained a long time ago, making it all the more likely that the data is no longer available. In practice, we have previously encountered situations where the labelled data was inaccessible to non-data scientist employees due to privacy concerns, whilst a classifier trained on this data could still be applied to make predictions. By requiring access to the labelled data used to train a classifier, AF-CBA belongs to a very particular category of XAI which is clearly not always applicable.

To improve the applicability of AF-CBA and place it in both the model-agnostic and 'data-agnostic' category of XAI, we could instead instantiate the CB on the basis of a set of earlier focus cases. The predicted outcome of a new focus case $f$ would then be based on the CB as usual. The caveat with this modification is that the outcomes in the CB are now derived from the classifier instead of annotators or decision makers. Its labels are inferred instead of observed and so no longer represent the ground truth. Nevertheless, whether justifications are based on the classifier's output rather than input labels does not change the fact that they refer to an approximation of what the classifier has learnt. This alternative approach is graphically depicted in Figure 3. The data-agnostic approach presumes that there are at least some precedents for either outcome in the CB on which to base justifications. This is a small limitation and only relevant when the approach is first implemented, and should therefore not be a serious concern.

We would normally construct training set $X$ by labelling a set sampled from the overall population $P$. The classifier thus trained is then able to predict the outcome of an unseen instance from $P$ (a focus case) to some degree of accuracy. We call the set of all focus cases $Q$. AF-CBA would normally base its justifications for any $f \in Q$ on the CB constructed from $X$. If instead the CB were to equal the set $Q$, AF-CBA would still construct justifications based on precedents. The distribution of frequent and infrequent fact situations in $X$ and $Q$ may be very different. However, if we can assume that $X$ is a representative sample of $P$ (as good training data is supposed to be), this concern diminishes as $Q$ grows and becomes a more representative sample of $P$ itself.

With this modification, the outcomes in $Q$ become predicted outcomes rather than training labels. This means that justifications generated from $Q$ refer to the labelling that the classifier has learnt to apply and not, as previously, to the labelling that the classifier was expected to learn. The labelled dataset represents a ground truth from which the classifier has learnt to make predictions and which we use to justify those predictions, but the benefit of this ground truth can be questioned. We have already stated that annotators and decision makers are not infallible, which explains at least some inconsistency in some datasets. Furthermore, we know that the classifier is not infallible. Even a state-of-the-art ML classifier is likely not to achieve perfect performance metrics and thus some misclassification is to be expected. With the standard (i.e. not data-agnostic) approach, those misclassifications are justified on the basis of precedents which, being the ground truth, do not contain that misclassification. With the data-agnostic approach, however, that misclassification is carried over into the CB used to make justifications. In a sense, that brings justifications closer to the black-box classifier whose predictions we wish to justify. We believe this to be an advantage of our data-agnostic modification.

## 5 EXPERIMENTS

The present work contributes to AF-CBA in two ways and so this paper's evaluation consists of two experiments. First, we compare alternative expressions for the authoritativeness of precedents using multiple datasets. From these results, we take the most promising expression and use it in a second experiment to study the applicability of our proposed solution to the problem of data-agnosticism. In the latter case, the evaluation involves the usage of a trained classifier to be treated as a black box. In order to actually obtain a black box, we train a classifier, but its performance is not the true concern of the experiment.

### 5.1 Expressions of authoritativeness

The evaluation approach is schematically depicted in Figure 4, whereby the usability of AF-CBA is evaluated using a series of metrics (descriptive statistics). Whilst $\mu(CB, Q)$ represents the total mean number of best precedents (Definition 10), we use $\mu_n(CB, Q)$ to represent the mean number of best precedents for non-trivial cases only (Definition 11). The number of inconsistent forcings (Definition 12) is represented as $N_{inc}(CB)$ and $N_{del}(CB)$ denotes the number of case deletions required to obtain a consistent subset (Definition 13). In the rest of this paper, we denote these metrics simply as $\mu$, $\mu_n$, $N_{inc}$ and $N_{del}$ if there is no risk of confusion.

Definition 10 (Number of best precedents). *Given a case base $CB$ and a set of focus cases $Q$, $\mu(CB, Q)$ is the mean number of best precedents in $CB$ for each focus case $f \in Q$.*

Definition 11 (Number of best precedents for non-trivial cases). *Given a case base $CB$ and a set of focus cases $Q$, $\mu_n(CB, Q)$ is the mean number of best precedents in $CB$ for each focus case $f \in Q$ where $f$ is not a trivial case.*

Definition 12 (Number of inconsistent forcings). *Given a case base $CB$ and a set of focus cases $Q$ with predicted outcome $s$, the number of inconsistent forcings $N_{inc}(CB, Q)$ is equal to the number of best precedents which force the decision for $f \in Q$ for $\bar{s}$ summed over all focus cases in $Q$.*

Definition 13 (Number of case deletions). *Given a case base $CB$, the number of case deletions $N_{del}(CB)$ is equal to the minimal number of cases $c \in CB$ that must be removed from $CB$ in order to be left with a consistent case base.*

Higher values of $\mu$ and/or $\mu_n$ suggests a larger number of potential winning strategies, making the selected justification more arbitrary. We therefore aim for low values for $\mu$ and $\mu_n$. Inconsistent forcings decrease the usability of AF-CBA's

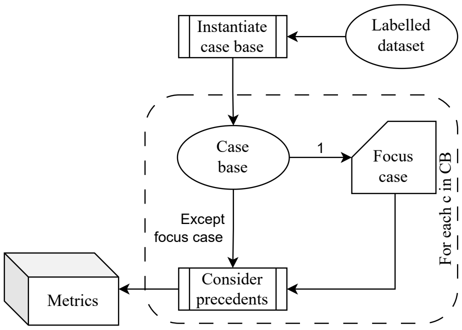

Fig. 4. An informal depiction of the evaluation setup, in which each case in the labelled set is used in turn as a focus case to determine total descriptive statistics.

Table 5. The datasets used in the first experiment.

| | CB size | Inconsistent cases |
|---|---|---|
| Admission | 500 | 3.20% |
| Churn | 7032 | 8.43% |
| SCOTUS | 6000 | 16.83% |
| No-shows | 6000 | 14.90% |
| COMPAS | 6000 | 21.13% |
| Fraud | 6000 | 0% |

justifications, as they tell the user that Horty's precedential constraint does not consistently explain the predicted outcome given the CB. As such, we would prefer $N_{inc}$ to stay as low as possible. The same goes for $N_{del}$, for which a value of 0 shows that we have circumvented the problem of inconsistency without case deletion (as was necessary in [27]).

We use the following datasets in our experiments: Admission [1] (applicants to a master's programme and whether they were accepted), Churn [17] (customers of a telecommunications provider and whether they switched providers), SCOTUS [29] (US Supreme Court cases and their decisions), No-shows [14] (medical patients and whether they showed up to an appointment), COMPAS [24] (convicted criminals and whether they are recidivists) and Fraud [28] (online payment details and whether they were fraudulent). The tendencies of all dimensions are determined using the Pearson correlation coefficient. We ignored unsubstantive features and we took random (fixed seed) samples for the larger datasets. See Table 5 for an overview of the sample sizes and the percentages of cases responsible for inconsistent forcings.[2]

We report the results for the four different metrics in Table 6 for the varying versions of authoritativeness together with the default usage (where authoritativeness is not used) with which we compare them. For the harmonic expression (4), we use $\beta = 1$, which yielded the best results for each dataset.

Whereas a higher value for $N_{del}$ often corresponds to a higher value for $N_{inc}$ in Table 6, No-shows has a particularly high value for $N_{inc}$ for the default approach (suggesting some surprising outliers). The number of inconsistent forcings $N_{inc}$ drops significantly compared to the default AF-CBA. For relative, product and harmonic authoritativeness, $N_{inc}$ is

[2] For a detailed account of the features used and the data preprocessing steps taken, as well as the source code of the experiments, see: https://github.com/JGTP/ICAIL.

Table 6. Results per dataset and per expression of authoritativeness, including the default approach where authoritativeness is not used.

| | Default | Relative $\alpha$ | Absolute $\alpha$ | Product $\alpha$ | Harmonic $\alpha$ |
|---|---|---|---|---|---|
| Admission | $\mu = 105.67$<br>$\mu_n = 6.12$<br>$N_{inc} = 496$<br>$N_{del} = 16$ | $\mu = 112.1$<br>$\mu_n = 22.59$<br>$N_{inc} = 0$<br>$N_{del} = 0$ | $\mu = 105.95$<br>$\mu_n = 7.34$<br>$N_{inc} = 0$<br>$N_{del} = 0$ | $\mu = 106.0$<br>$\mu_n = 7.61$<br>$N_{inc} = 0$<br>$N_{del} = 0$ | $\mu = 105.97$<br>$\mu_n = 7.44$<br>$N_{inc} = 0$<br>$N_{del} = 0$ |
| Churn | $\mu = 33.71$<br>$\mu_n = 26.13$<br>$N_{inc} = 19484$<br>$N_{del} = 593$ | $\mu = 83.8$<br>$\mu_n = 109.86$<br>$N_{inc} = 6$<br>$N_{del} = 3$ | $\mu = 43.21$<br>$\mu_n = 37.24$<br>$N_{inc} = 42$<br>$N_{del} = 20$ | $\mu = 43.68$<br>$\mu_n = 38.28$<br>$N_{inc} = 4$<br>$N_{del} = 2$ | $\mu = 43.32$<br>$\mu_n = 37.56$<br>$N_{inc} = 4$<br>$N_{del} = 2$ |
| SCOTUS | $\mu = 28.96$<br>$\mu_n = 27.28$<br>$N_{inc} = 62600$<br>$N_{del} = 1010$ | $\mu = 345.42$<br>$\mu_n = 411.82$<br>$N_{inc} = 52$<br>$N_{del} = 23$ | $\mu = 44.1$<br>$\mu_n = 40.89$<br>$N_{inc} = 262$<br>$N_{del} = 66$ | $\mu = 43.52$<br>$\mu_n = 40.72$<br>$N_{inc} = 16$<br>$N_{del} = 7$ | $\mu = 44.2$<br>$\mu_n = 41.1$<br>$N_{inc} = 16$<br>$N_{del} = 7$ |
| No-shows | $\mu = 235.98$<br>$\mu_n = 60.29$<br>$N_{inc} = 333584$<br>$N_{del} = 894$ | $\mu = 556.16$<br>$\mu_n = 379.85$<br>$N_{inc} = 0$<br>$N_{del} = 0$ | $\mu = 246.51$<br>$\mu_n = 62.59$<br>$N_{inc} = 86$<br>$N_{del} = 34$ | $\mu = 246.74$<br>$\mu_n = 62.7$<br>$N_{inc} = 0$<br>$N_{del} = 0$ | $\mu = 246.62$<br>$\mu_n = 62.65$<br>$N_{inc} = 0$<br>$N_{del} = 0$ |
| COMPAS | $\mu = 284.43$<br>$\mu_n = 36.68$<br>$N_{inc} = 492176$<br>$N_{del} = 1268$ | $\mu = 480.4$<br>$\mu_n = 306.84$<br>$N_{inc} = 0$<br>$N_{del} = 0$ | $\mu = 286.4$<br>$\mu_n = 40.97$<br>$N_{inc} = 256$<br>$N_{del} = 73$ | $\mu = 287.13$<br>$\mu_n = 42.33$<br>$N_{inc} = 0$<br>$N_{del} = 0$ | $\mu = 286.77$<br>$\mu_n = 41.81$<br>$N_{inc} = 0$<br>$N_{del} = 0$ |
| Fraud | $\mu = 62.5$<br>$\mu_n = 63.03$<br>$N_{inc} = 0$<br>$N_{del} = 0$ | $\mu = 62.5$<br>$\mu_n = 63.03$<br>$N_{inc} = 0$<br>$N_{del} = 0$ | $\mu = 67.96$<br>$\mu_n = 68.61$<br>$N_{inc} = 0$<br>$N_{del} = 0$ | $\mu = 67.96$<br>$\mu_n = 68.61$<br>$N_{inc} = 0$<br>$N_{del} = 0$ | $\mu = 67.96$<br>$\mu_n = 68.61$<br>$N_{inc} = 0$<br>$N_{del} = 0$ |

0 for all but Churn and SCOTUS. Absolute authoritativeness still has a few inconsistent forcings for all but Admission and Fraud. Those results for absolute authoritativeness, however, are still significantly lower than for the default. We observe that $N_{del}$ likewise drops drastically with all expressions, suggesting that remaining occurrences of inconsistent forcing are typically caused by a very small number of cases.

We see that the impact of authoritativeness on $\mu$ and $\mu_n$ is quite limited for most versions of authoritativeness (except relative authoritativeness), suggesting that explanations should not become more arbitrary by using authoritativeness in the identification of best precedents in most scenarios. They rise slightly if inconsistency is not completely reduced. Since Fraud is an entirely consistent dataset, $N_{inc}$ and $N_{del}$ remain zero for all expressions of authoritativeness. We see in Table 6 that in this scenario, $\mu$ and $\mu_n$ are hardly affected. This suggests that authoritativeness could be safely applied without checking for the degree of inconsistency.

From these results we conclude that our modification successfully deals with the problem of CB inconsistency, that our modification is therefore preferable to the default identification of best precedents and that the harmonic expression of authoritativeness (4) with $\beta = 1$ is typically preferable (although it differs little from the product expression of authoritativeness (3)).

### 5.2 data-agnostic approach

We conduct a second experiment to investigate the consequences of the modification described in Section 4 allowing for a data-agnostic version of AF-CBA. In order to evaluate this modification of the approach, we randomly split the complete Churn dataset of size 7032 evenly into $X$ (including the labels) and $Q$ (excluding the labels). Dataset $X$ is used to train a classifier[3] and the cases from $Q$ are incrementally used as focus cases and appended to the CB after receiving a prediction from the classifier. The evaluation metrics can thus be recalculated as the CB grows. This process (including the random split) is repeated three times and the metrics are averaged, to make the results more generalisable. See Figure 5 for a depiction of this setup.

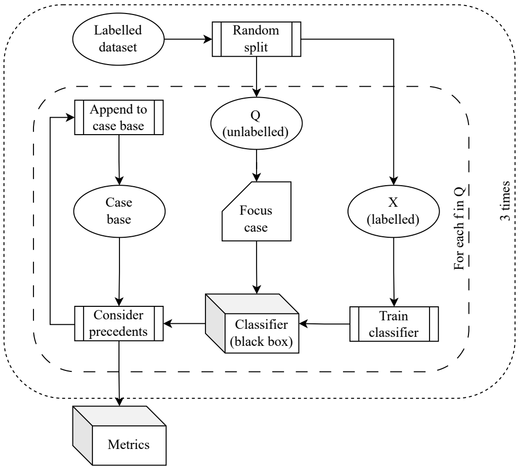

Fig. 5. A schematic depiction of the evaluation setup for the data-agnostic approach.

The data-agnostic approach is evaluated using in part the same metrics as before, to see if these behave in an unexpected way. We expect $\mu$ and $\mu_n$ to grow steadily with $\mid CB \mid$ and eventually level off, showing that the CB is becoming more similar to $X$ (presumably because it is becoming more representative of the population $P$). We expect $N_{inc}$ to grow much more slowly, as inconsistent forcings are often caused by exceptional cases (outliers), which are likely to be misclassified by the classifier and thus receive the opposite (consistent) outcome in the CB. We expect the same for $N_{del}$.

In addition to these previous four metrics, we introduce two metrics that were unaffected by authoritativeness but can vary here, namely the number of trivial and non-trivial winning strategies. This distinction was also used by Prakken & Ratsma [27]. Trivial winning strategies do not play to the full the strength of AF-CBA, as they make for very simple justifications. We would prefer them to be relatively rare. We measure the number of trivial winning strategies by the metric $N_{tws}(CB, Q)$ (Definition 14) and the number of non-trivial winning strategies as $N_n(CB, Q)$ (Definition 15), hereafter denoted as $N_{tws}$ and $N_n$.

Definition 14 (Number of cases with trivial winning strategies). *Given a case base $CB$ and a set of focus cases $Q$, the number of trivial winning strategies $N_{tws}(CB, Q)$ is equal to the number of focus cases $f \in Q$ for which there is a case $c \in CB$ for which $D(f, c) = 0$.*

Definition 15 (Number of cases with non-trivial winning strategies). *Given a case base $CB$ and a set of focus cases $Q$, the number of non-trivial winning strategies $N_n(CB, Q)$ is equal to the number of focus cases $f \in Q$ for which there is no case $c \in CB$ for which $D(f, c) = 0$.*

[3] We follow the same strategy using XGBoost as a popular notebook (reported accuracy: 0.83 [5]) found on the data science website Kaggle for the Churn dataset. With our smaller training set, this reaches an average accuracy of 0.77.

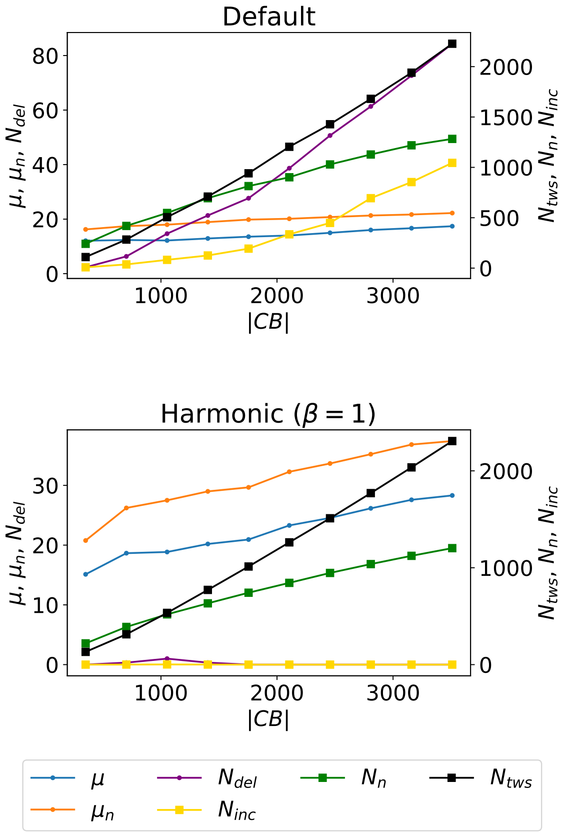

Fig. 6. The various metrics as a function of the size of the CB in the data-agnostic approach for the churn dataset, averaged over three iterations for the default approach (top) and the harmonic ($\beta = 1$) authoritativeness approach (bottom) to best precedent identification.

We want to compare the evaluation metrics for the growing CB using predicted outcome labels to what the metrics would have been if we had used the original labels ($| CB |= 7032/2 = 3516$) instead (default (mean): $\mu = 20.49$, $\mu_n = 21.56$, $N_{inc} = 4906.67$, $N_{del} = 240.33$, $N_{tws} = 2435.33$, $N_n = 1080.67$; harmonic (mean): $\mu = 26.17$, $\mu_n = 29.46$, $N_{inc} = 0$, $N_{del} = 0$, $N_{tws} = 2529$, $N_n = 987$). We want to see if the values using the predicted labels converge on those using the original labels.

The performance is presented in Figure 6. We see that all metrics grow steadily with $| CB |$, albeit not at the same rate. As was to be expected given our earlier results, both $N_{inc}$ and $N_{del}$ are nearly absent in the bottom plot, re-emphasising the usefulness of authoritativeness. Both metrics are lowered in the top plot too compared to the means using the original labels, suggesting that the classifier is misclassifying cases that were originally the cause of inconsistent forcing relations to now be consistent with the CB. In both plots, both $N_{tws}$ and $N_n$ do not deviate much from their values using the original labels, we only see a small tilt towards $N_n$.

The mean number of precedents $\mu$ and $\mu_n$ do not deviate much from their values with the original labels, except that $\mu_n$ is slightly higher for harmonic authoritativeness. This difference is not large enough to cause concern and nothing else suggests that using authoritativeness provides any disadvantage for the data-agnostic approach. Overall, these results suggest that the data-agnostic approach produces no significant adverse effects.

## 6 DISCUSSION AND FUTURE WORK

Our results suggest that $\mu$ and $\mu_n$ can show a small increase when using harmonic authoritativeness (4) for datasets where $N_{inc}$ and $N_{del}$ are not completely reduced to zero, although it does still largely mitigate the problem of inconsistency. This could imply that a comparison between the default approach and harmonic authoritativeness could sometimes be warranted when implementing AF-CBA. However, the reduction in $N_{inc}$ and $N_{del}$ is so strong that it does not seem very likely that the default approach should ever be preferable because of a small rise in $\mu$ and $\mu_n$.

None of our four expressions for authoritativeness (1)-(4) ever reach a value of 0 for any case in the CB. This seems intuitive, since any case should have at least some authoritativeness simply due to its being a precedent. A value of $\alpha(c) = 1$ is only realistic when using our relative expression of authoritativeness (1). This would only be a problem if values due to different expressions of authoritativeness would have to be compared to each other, which is not currently part of our approach. If multiple explanations are ever to be compared as part of some overarching approach, these (and possibly other) characteristics of alternative authoritativeness expressions would have to be taken into account.

Our data-agnostic modification raises a rather difficult question to answer: is it preferable to justify a misclassification using precedents which are (presumably) correctly labelled, or to justify them using precedents which are labelled by the same classifier which made that misclassification and thus may themselves be misclassified? In both cases, AF-CBA provides justifications for the outcome and the debate is therefore related to the debate surrounding explainability through justifications [6]: is it better to attempt to provide an interpretable representation of an abstract mathematical formula or to provide a train of thought that makes a decision meaningful to the user? The number of inconsistent forcings may be smoothed out in the data-agnostic approach, but misclassifications are justified to the user in both circumstances. Furthermore, we have argued for the data-agnostic approach by stating that training data is often not available in practice, thus eliminating the standard approach as a possibility. Most importantly, this decision to use the data-agnostic approach should be transparent and made apparent to the users of AF-CBA at all times.

Our modification of AF-CBA relies on the intuition that one precedent can be more authoritative than another. We have demonstrated its consequences, given that higher or lower metric scores indicate better explanations (as was argued in the original paper [27]). However, it could be argued that these metrics offer a limited insight into the quality of AF-CBA's explanations. As AF-CBA is intended to justify predictions to human users and since the effectiveness of justifications can be surprising in practice (see, for example, work by Branting et al. [7]), testing its performance thoroughly requires a usability study. Any alternative modifications and additional metrics could then be compared to study the effect in a real-world setting. This could also allow us to investigate the relation between the nature of a dataset and the perceived applicability of a fortiori reasoning, as the notion of precedential constraint may be more suitable for those contexts in which internal consistency plays a large role.

There may be qualitative reasons for ranking precedents in addition to quantitative ones like the notion of authoritativeness as used in this paper. One might rank decisions of a supreme court higher than that of a lower court, for instance. How to combine qualitative and quantitative reasons for ranking precedents could be a topic of future work. Furthermore, additional modifications to AF-CBA could include incorporating complex arguments in the explanations (AF-CBA is qualified as a 'top-level' model due to the possibility of providing it with a set of definitions as to why specific downplaying moves can be played) or accounting for dimensions which are highly dependent. Another possibility is an alteration that allows dimensions to have a more complex effect on predictions than the tendencies used in this paper. There exist binary classification tasks for which this would be desirable. For example, a dimension such as blood pressure could be a predictor for illness both at very low and very high values, with a value in the intermediate range being a predictor for the patient not being ill. We intend to include this in our future work.

## 7 RELATED WORK

The problem of (factor-based) reasoning with an inconsistent CB is studied from a different perspective by Canavotto [8]. A generalised notion of precedential constraint allows for a conflict-free deontic logic that allows a court using an inconsistent CB to be either required to decide for one side or the other, or permitted to decide for either side. Rather than requiring the court to preserve the consistency of a consistent CB, courts are required to avoid new inconsistencies when extending an existing CB.

AF-CBA is emphatically not used to classify cases, but Horty's underlying a fortiori reasoning can be used as such. One example is by Odekerken & Bex [23], who use precedential constraint in a transparent classification system of fraudulent web shops. Within their legal case-based reasoning module, the factors of a web shop are compared to labelled precedents to determine whether it should receive the label *mala fide*, *bona fide* or neither. The result, including the precedent for which precedential constraint applies, are presented to a human user for further analysis. If the user disagrees with the received label, they can add factors to the case along with the corrected label. As such, the CB grows more elaborate and class predictions gradually improve.

The original top-level model [27] is rephrased by Van Woerkom et al. [30] in terms of justification and citation relations. The explanation model is thus shown to be equivalent to adding an extension of the forcing relation to the theory of precedential constraint. In a related paper [31], Van Woerkom et al. introduce a notion of 'landmark cases' and use it to characterise applicable datasets.

Consistency plays perhaps a larger role in case-based reasoning specifically than it does in ML at large, due to the notion of inconsistent forcing resulting from precedential constraint. However, there is recent work on notions of consistency within the ML literature. For example, various types of label noise are investigated by Frénay & Verleysen [12]. Furthermore, various types of consistency are considered in work by Li et al. [18]. They argue that optimising a classifier only for accuracy has its downsides, such as neglecting a need for internal consistency across examples. They present a learning framework in which logic rules allow models to be regularised away from inconsistencies by relaxing those rules using t-norms, thereby producing differentiable functions which can be used as loss terms. It is shown how annotation consistency simply results in standard cross-entropy loss when converted with product t-norms and transformed to the negative log space, thereby giving credence to their approach. By combining several constraints in the overall loss function, they show that one does not have to come at the cost of another and that the interplay between loss terms can even be beneficial. One could perhaps view their notion of additional consistencies as the application of domain rules, which raises the question whether domain rules could have a role to play in example-based XAI as well.

## CONCLUSION

AF-CBA is an approach from the example-based XAI tradition [4] inspired by work in AI & Law [2, 9, 10]. AF-CBA uses case-based argumentation to provide *post hoc* justifications for label predictions by an opaque (black box) machine-learning classifier. This paper presents intuitive graphical clarifications of the processes in AF-CBA, and extends AF-CBA in two novel ways:

- We have presented the notion of precedential authoritativeness to mitigate the problem of case base inconsistency.

- We have presented a data-agnostic version of AF-CBA which does not rely on the original training data.

We initially modelled AF-CBA's criteria for best precedents by adding a quantified expression of how authoritative a precedent is in light of the degree to which that case base consistently supports the conclusion of that precedent's outcome (label). We experimented with alternative quantifications of this criterion to study which expression is the most fruitful regarding the handling of inconsistency without adversely affecting the explanations. This was determined to be the harmonic expression of authoritativeness.

We subsequently replaced the case base (previously equated with the labelled data used to train the classifier) with one constructed from earlier predictions. We have shown that this does not have any serious adverse effects on the evaluation metrics used for AF-CBA and that the advantages of using our authoritativeness criterion translate to this data-agnostic approach, thus allowing AF-CBA to be used without access to the original labelled data.

## ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their feedback and suggestions.

## REFERENCES

[1] Mohan Acharya, Asfia Armaan, and Aneeta Antony. 2019. A Comparison of Regression Models for Prediction of Graduate Admissions. In *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*. 1–5.

[2] Vincent Aleven. 1997. *Teaching Case-Based Argumentation through a Model and Examples*. Ph. D. Dissertation. University of Pittsburgh, Pittsburgh.

[3] K. D. Ashley. 1989. Modelling Legal Argument: Reasoning with Cases and Hypotheticals. (1989), 1.

[4] Katie Atkinson, Trevor Bench-Capon, and Danushka Bollegala. 2020. Explanation in AI and Law: Past, Present and Future. *Artificial Intelligence* 289 (Dec. 2020), 103387.

[5] Atindra Bandi. 2019. Telecom Churn Prediction. https://kaggle.com/code/bandiatindra/telecom-churn-prediction.

[6] Adrien Bibal, Michael Lognoul, Alexandre de Streel, and Benoît Frénay. 2021. Legal Requirements on Explainability in Machine Learning. *Artificial Intelligence and Law* 29, 2 (June 2021), 149–169.

[7] L. Karl Branting, Craig Pfeifer, Bradford Brown, Lisa Ferro, John Aberdeen, Brandy Weiss, Mark Pfaff, and Bill Liao. 2021. Scalable and Explainable Legal Prediction. *Artificial Intelligence and Law* 29, 2 (June 2021), 213–238.

[8] Ilaria Canavotto. 2022. Precedential Constraint Derived from Inconsistent Case Bases. *Legal Knowledge and Information Systems* (2022), 23–32.

[9] Kristijonas Čyras, David Birch, Yike Guo, Francesca Toni, Rajvinder Dulay, Sally Turvey, Daniel Greenberg, and Tharindi Hapuarachchi. 2019. Explanations by Arbitrated Argumentative Dispute. *Expert Systems with Applications* 127 (Aug. 2019), 141–156.

[10] Kristijonas Čyras, Ken Satoh, and Francesca Toni. 2016. Explanation for Case-Based Reasoning via Abstract Argumentation. *Computational Models of Argument* (2016), 243–254.

[11] Phan Minh Dung. 1995. On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence* 2, 77 (1995), 321–357.

[12] Benoît Frénay and Michel Verleysen. 2014. Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems* 25, 5 (May 2014), 845–869.

[13] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *Comput. Surveys* 51, 5 (2018), 93:1–93:42.

[14] Joni Hoppen. 2017. Medical Appointment No Shows. https://www.kaggle.com/datasets/joniarroba/noshowappointments.

[15] John Horty. 2011. Rules and Reasons in the Theory of Precedent. *Legal Theory* 17 (2011), 1–34.

[16] John Horty. 2019. Reasoning with Dimensions and Magnitudes. *Artificial Intelligence and Law* 27, 3 (2019), 309–345.

[17] IBM. 2018. Telco Customer Churn. https://www.kaggle.com/blastchar/telco-customer-churn.

[18] Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A Logic-Driven Framework for Consistency of Neural Models. *arXiv:1909.00126 [cs]* (Sept. 2019).

[19] Zachary Lipton. 2016. The Mythos of Model Interpretability. *Commun. ACM* 61 (2016), 96–100.

[20] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267 (2019), 1–38.

[21] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *arXiv:2103.14749 [cs, stat]* (2021).

[22] Conor Nugent and Pádraig Cunningham. 2005. A Case-Based Explanation System for Black-Box Systems. *Artificial Intelligence Review* 24, 2 (Oct. 2005), 163–178.

[23] Daphne Odekerken and Floris Bex. 2020. Towards Transparent Human-in-the-Loop Classification of Fraudulent Web Shops. (2020), 4.

[24] Dan Ofer. 2017. COMPAS Recidivism Racial Bias. https://www.kaggle.com/datasets/danofer/compass.

[25] Joeri G T Peters, Floris J Bex, and Henry Prakken. 2022. Justifications Derived from Inconsistent Case Bases Using Authoritativeness. In *Proceedings of the 1st International Workshop on Argumentation for eXplainable AI (ArgXAI 2022) Co-Located with 9th International Conference on Computational Models of Argument (COMMA 2022)*, Vol. 3209. CEUR WS, Cardiff, Wales, UK, 1–13.

[26] Henry Prakken. 1999. Dialectical Proof Theory for Defeasible Argumentation with Defeasible Priorities (Preliminary Report). In *Formal Models of Agents (Lecture Notes in Computer Science)*, John-Jules Ch. Meyer and Pierre-Yves Schobbens (Eds.). Springer, Berlin, Heidelberg, 202–215.

[27] Henry Prakken and Rosa Ratsma. 2022. A Top-Level Model of Case-Based Argumentation for Explanation: Formalisation and Experiments. *Argument & Computation* 13, 2 (Jan. 2022), 159–194.

[28] Rupak Roy. 2022. Online Payments Fraud Detection Dataset. https://www.kaggle.com/datasets/rupakroy/online-payments-fraud-detection-dataset.

[29] Harlod J Spaeth, Lee Epstein, Andrew D Martin, Jeffrey A Segal, Theodore J Ruger, and Sara C Benesh. 2022. Supreme Court Database (2022 Release 1). http://scdb.wustl.edu/data.php.

[30] Wijnand van Woerkom, Davide Grossi, Henry Prakken, and Bart Verheij. 2022. Justification in Case-Based Reasoning. In *Proceedings of the 1st International Workshop on Argumentation for eXplainable AI (ArgXAI 2022) Co-Located with 9th International Conference on Computational Models of Argument (COMMA 2022)*. CEUR WS, Cardiff, Wales, UK, 1–13.

[31] Wijnand van Woerkom, Davide Grossi, Henry Prakken, and Bart Verheij. 2022. Landmarks in Case-Based Reasoning: From Theory to Data. *HHAI2022: Augmenting Human Intellect* (2022), 212–224.

[32] Paul Voigt and Axel von dem Bussche. 2017. *The EU General Data Protection Regulation (GDPR).* Springer International Publishing, Cham.