Formal Results on Case-Base Consistency: A COMPAS Case Study

Anonymous Author(s)

Abstract

Case-based reasoning is a central theme of AI and law research, providing formal models to assess decision-making consistency. We contribute to this research by proving formal results on case base consistency, and relating them to a consistency analysis of the COMPAS risk assessment dataset. COMPAS, a widely used recidivism prediction tool in the U.S. criminal justice system, has been at the center of debates on fairness and interpretability. We present four theorems related to case-base consistency, analyzing how statistical modeling techniques, feature modifications, and data binning affect consistency outcomes. Our findings demonstrate that generalized linear models necessarily yield consistent decisions, while modifications such as input feature addition/removal and data binning can either increase or decrease consistency, respectively.

CCS Concepts

• **Theory of computation** → *Automated reasoning*; • **Information systems** → *Expert systems*; • **Computing methodologies** → *Knowledge representation and reasoning*.

Keywords

Case-based reasoning, COMPAS, Consistency, Generalized linear models, Data binning, Dimension hierarchy

ACM Reference Format:

Anonymous Author(s). 2018. Formal Results on Case-Base Consistency: A COMPAS Case Study. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

An important aspect of the literature on AI & law has been the development and study of formal models of case-based reasoning [1, 28, 16, 15]. In recent years, these models have increasingly seen applications to the development of interpretable AI [23, 24], post-hoc explanation of AI decisions [27, 26, 33], and normative computational reasoning [6, 7].

In the present work, we contribute to the study of the model of case-based reasoning presented by van Woerkom et al. [34], which is an expanded version of the result model developed of precedential constraint developed by Horty [15]. A key concept in these models is that of case base consistency. Recently, van Woerkom et al. [32] used the model of [34] to analyze the consistency of the COMPAS risk score dataset.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

The COMPAS program (Correctional Offender Management Profiling for Alternative Sanctions) is a risk and need assessment tool, first developed in 1998 by Equivant (then Northepointe), which is widely used by criminal justice agencies in the United States [12, 31]. Over the years it has become the subject of intense debate in the literature on responsible AI, particularly regarding the use of black box AI models for high stake decisions [29]. In this context, a black box AI model is a system of which it is not clear how it functions internally. This can either be because the inner workings of the model are unintelligible, or because the model is proprietary.

One of the early initiators of the debate surrounding COMPAS was a publication by Angwin et al. [2], writing for the ProPublica news outlet, in which it was alleged that COMPAS is racially biased in its decision-making, based on an analysis of a dataset composed of COMPAS risk score assessments. The allegations in [2] have since been refuted [8, 13, 31], but the dataset used in [2] has remained highly relevant to this day.

The main contributions of this work are to prove four theorems about case base consistency, in the sense defined by [34]. We demonstrate each of these theorems through applications to the COMPAS dataset. Our results can be summarized as follows. Firstly, we prove that a large class of popular statistical modeling techniques, called generalized linear models [20], produce decisions that are necessarily consistent (Theorem 4.6). Secondly, we relate the influence of the removal and addition of features to a dataset to changes in case base consistency (Theorem 5.6). Lastly, we prove two theorems about the effects on consistency of data binning, an important technique in data-science. In particular, we show that model outputs increases consistency (Theorem 6.7), and that binning model inputs decreases consistency (Theorem 6.9).

We relate each of these formal results to aspects of the consistency of the COMPAS risk score assignments. In this regard, our work is closely related to the work of van Woerkom et al. [32]. An important difference to note in this regard is that we will use a different version of the COMPAS dataset. One of the problems of COMPAS data published by ProPublica in Angwin et al. [2] is that it is missing input features of the COMPAS program. In attempt to rectify this, Rudin et al. [31] supplemented ProPublica's dataset with probation data purchased from the Broward Clerk's office. We will use this expanded version of the COMPAS data, whereas [32] used ProPublica's version.

Lastly, we note that due to space constraints, we will only provide a formal proof of the first of Theorem 4.6 regarding consistency of the decisions of generalized linear models. The rest of the proofs can be found in an online appendix, together with all the source code that can be used to reproduce our results.[1].

The remainder of this paper is structured as follows. In Section 2, we provide an overview of the COMPAS risk assessment dataset, including its structure and the features it contains. Section 3 recalls the formal model of a fortiori case-based reasoning and its associated notion of consistency. We then turn to our first main

[1]Our source code is available at https://github.com/icailauthor/compascbr.

result, regarding the consistency of linear models, in Section 4, and analyze the consistency of the COMPAS risk scores in light of this result. Section 5 considers the effects on consistency of the addition and removal of data features, and Section 6 examines the impact of dimension binning. Finally, Section 7 concludes the paper and outlines directions for future research.

2 The COMPAS Risk Assessment Dataset

The primary outputs of the COMPAS program are its need and risk scale assessments, computed based on answers to questionnaires; see [3] for examples. The COMPAS need scales measure constructs like financial problems, substance abuse, and depression, while the risk scales predict factors like recidivism, violence, and failure to appear [12]. Additionally, COMPAS provides a "level of supervision" recommendation, ranging from 1 (minimum) to 4 (high).

The inputs to these scales are answers to questionnaires about prior offenses, education, work experience, etc. Some data are self-reported. These answers inform the need scales, which in turn inform the risk scales. The COMPAS risk scales, particularly the "General Recidivism Risk Score" (GRRS) and the "Violent Recidivism Risk Score" (VRRS), have been debated. ProPublica published a dataset in [2] containing COMPAS risk scores assigned between 2013 and 2014 in Broward County, Florida. This dataset includes scores, information on which these scores were presumably computed, and information about whether a person recidivated or committed a violent act after being scored [18].

The dataset has been criticized for missing features needed to compute the COMPAS risk scores [19]. Rudin et al. [31] supplemented ProPublica's dataset with probation data from the Broward Clerk's office to fill in some missing features. However, some information is still missing [31, Table 1]. The dataset published in [31] contains data about 9 features used to compute the GRRSand 13 for the VRRS [31, Table 1 & Tables A4–A8], [10, Tables 1–5].

The inputs to the COMPAS failure to appear risk score are less clear. This score has received less attention in the literature, possibly because the datasets do not contain "true label" information on this scale. The COMPAS Practitioner's Guide [12, Section 4.1] mentions risk scales for general recidivism, violent recidivism, and pretrial misconduct, referring to the GRRS, VRRS, and the "Pretrial Release Risk Scale" (PRRS) respectively. Although it is natural to assume that the failure to appear risk scores in the Broward County COMPAS data correspond to the PRRS, Equivant has clarified that they do not [11]. For this work, we assume the failure to appear risk scores have similar inputs to the PRRS, specifically the history of criminal involvement subscale and marital status information, totaling 8 features.

Lastly, the COMPAS risk scores can be presented in various ways. Initially, COMPAS produces *raw* scores, which can take any value (e.g., −1.54) [12]. For interpretability, these raw scores are converted to *decile* scores by comparing them to those of a *norm group*—a representative sample of the target population of the agency using COMPAS. For instance, a raw score of −1.54 might be converted to a decile score of 6, indicating it is higher than the lowest 50% but lower than the highest 40% of scores in the norm group. These decile scores may then be further represented as "Low" (scores 1–4), "Medium" (5–7), and "High" (8–10). (According to [22], "a surprising number of agencies prefer the traditional labels of low, medium, and high risk.")

We will use the extended dataset published by Rudin et al. [31] to relate formal results about case-based reasoning consistency, to analysis of the COMPAS risk scores. Before turning to these results, we first recall the model of case-based reasoning and its associated notion of consistency.

3 A Model of a Fortiori Constraint

For the sake of a self-contained presentation we now recall the relevant formal definitions of the model of case-based reasoning that we use, which stem from [15, 34, 32]; the reader is referred there for additional details. We illustrate these definitions through examples pertaining to COMPAS, in order to build intuition for our proofs and data analyses in the following sections.

Definition 3.1. A *linear order* on a set d is a relation $\trianglelefteq$ which is:

- Reflexive: $v \trianglelefteq v$ for $v \in d$.
- Transitive: $v \trianglelefteq w$ and $w \trianglelefteq x$ imply $v \trianglelefteq x$ for $v, w, x \in d$.
- Antisymmetric: $v \trianglelefteq w$ and $w \trianglelefteq v$ imply $v = w$ for $v, w \in d$.
- Total: $v \trianglelefteq w$ or $w \trianglelefteq v$ for $v, w \in d$.

As usual, we write $v \triangleleft w$ when $v \trianglelefteq w$ and $v \neq w$.

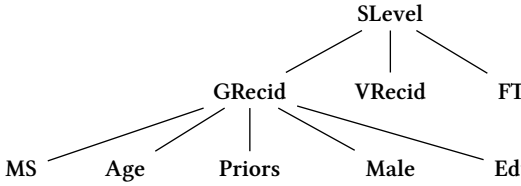
Definition 3.2. A *dimension* is a set d with a linear order $\trianglelefteq$ on d.

We may refer to just the set d as the dimension, leaving the reference to its linear order implicit. Furthermore, in the context of a set of dimensions D, we will, for the sake of brevity, refer to all the orders of the dimensions by just $\trianglelefteq$, because confusion as to which dimension order is being referred to is unlikely to arise. For example, given two dimensions d and e, we will refer to both of the dimensions orders of d and e by $\trianglelefteq$, rather than introducing separate notations such as $\trianglelefteq_d$ and $\trianglelefteq_e$.

Definition 3.3. A *dimension hierarchy* (D, H) is a finite set of dimensions D with a relation H on D such that the transitive closure of H is irreflexive. A dimension is *base-level* if it is H-minimal, and *abstract* otherwise. We denote the pre-image of a dimension d under a hierarchical structure H by $\mathsf{H}(d) = \{e \in D \mid \mathsf{H}(e, d)\}$.

A link $\mathsf{H}(d, e)$ between dimensions d, e in a dimension hierarchy (D, H) indicates that there is a positive correlation between the values of d and e, relative to their orders.

Example 3.4. Below is an example of a dimension hierarchy in the context of the COMPAS risk scores:



(1)

The hierarchical structure H is indicated by the lines, where the higher dimensions indicate an increasing level of abstraction. The associated sets, orders, and meanings of the base-level dimensions,

displayed in the bottom row of (1), is as follows:

$$\begin{aligned}
\textbf{MS} &= (\{0,1\}, \geq), \text{is married or not,}\\
\textbf{Age} &= (\{18,19,20,\dots\}, \geq), \text{the age of the defendant,}\\
\textbf{Priors} &= (\{0,1,2,\dots\}, \leq), \text{the number of prior offenses,}\\
\textbf{Male} &= (\{0,1\}, \leq), \text{is male or not,}\\
\textbf{Educ} &= (\{0,1\}, \geq), \text{completed high school or not.}
\end{aligned}$$

The order $\geq$ of the **MS**, **Age**, and **Educ** dimensions indicates that higher values for these dimensions should generally lead to lower values for the overlying **GRecid** dimensions. Conversely, the $\leq$ order of the other dimensions, such as the number of priors **Priors**, indicates that higher values generally lead to higher values for the **GRecid** dimension.

Above the base-level dimensions is a row of more abstract dimensions. Note that **VRecid** and **FTA** are technically base-level dimensions in this example—in a more realistic version of this hierarchy these would also be dependent on less abstract dimensions. We use the decile scores, ranging from 1 to 10, for this example:

$$\begin{aligned}
\textbf{GRecid} &= (\{1,2,\dots,10\}, \leq), \text{the general recidivism risk score,}\\
\textbf{VRecid} &= (\{1,2,\dots,10\}, \leq), \text{the violent recidivism risk score,}\\
\textbf{FTA} &= (\{1,2,\dots,10\}, \leq), \text{the failure to appear risk score.}
\end{aligned}$$

At the top of the hierarchy is the **SLevel** dimension:

$$\textbf{SLevel} = (\{1,2,3,4\}, \leq), \text{the recommended supervision level.}$$

Remark 3.5. It is somewhat uncommon to assume that the dimension orders satisfy the totality property of Definition 3.1. For example, the related works [15, 27, 24] do not make this assumption. At present, we opt to assume totality because dimension orders that do not satisfy it are rare—especially in the conext of AI. For example, the statistical methods used by [27, 33] to determine dimension orders can only produce orders satisfying totality. Furthermore, some of the results we will state and prove depend on the totality property. We do believe that with appropriate modifications our results can be rephrased to apply to dimensions that do not make the totality assumption, but this would be the topic of future research.

Definition 3.6. A *fact situation* X for a set of dimensions D is a partial choice function on D. We denote the domain of a fact situation X by $\mathrm{dom}(X)$. The set of all fact situations for D is denoted by $\mathcal{X}(D)$, and a *case base* is a finite subset $C \subseteq \mathcal{X}(D)$.

Definition 3.7. Given a case base C and a value $v \in d$, a fact situation X is *lower bounded* in d by C to v, denoted by $C \vDash v \trianglelefteq X(d)$, if and only if either

(1) v is the least element of d, or
(2) $v \trianglelefteq X(d)$, or
(3) d is abstract, and there is $Y \in C$ satisfying $v \trianglelefteq Y(d)$ such that $C \vDash Y(e) \trianglelefteq X(e)$ holds for all $e \in \mathsf{H}(d) \cap \mathrm{dom}(Y)$.

The *upper bound* $C \vDash X(d) \trianglelefteq v$ is defined similarly.

Table 1: Three fact situations X, Y, Z for the example dimension hierarchy for the recidivism risk domain depicted in (1). A dash indicates that the fact situation is undefined on that particular dimension.

	MS	Age	Priors	Male	Educ	GRecid	VRecid	FTA	SLevel
X	1	25	3	0	1	7	4	7	2
Y	0	30	2	0	0	5	8	—	3
Z	1	20	3	1	—	—	9	5	—

Example 3.8. An example case base for the dimension hierarchy of Example 3.4 is listed in Table 1.

$$\{X,Y\} \vDash 5 \trianglelefteq Z(\textbf{GRecid}) \tag{2}$$
$$\text{if } \{X,Y\} \vDash X(e) \trianglelefteq Z(e) \text{ for all } e \in \mathsf{H}(\textbf{GRecid}) \cap \mathrm{dom}(X) \tag{3}$$
$$\text{if } 25 \trianglelefteq Z(\textbf{Age}) \text{ and } 3 \trianglelefteq Z(\textbf{Priors}) \tag{4}$$
$$\text{if } 25 \geq 20 \text{ and } 3 \leq 3 \tag{5}$$

Step (3) corresponds to disjunct (3) of Definition 3.7, and may be applied because **GRecid** is abstract and $5 \trianglelefteq X(\textbf{GRecid}) = 7$. Step (4) follows from disjuncts (3) and (4), as X selects the least elements of **MS**, **Male**, and **Educ**. Step (5) simply fills in the definition of Z, and is a true statement, so that we indeed have a lower-bound constraint $\{X,Y\} \vDash 5 \trianglelefteq Z(\textbf{GRecid})$. Using this, we can in turn derive:

$$\{X,Y\} \vDash 3 \trianglelefteq Z(\textbf{SLevel})$$
$$\text{if } \{X,Y\} \vDash Y(e) \trianglelefteq Z(e) \text{ for all } e \in \mathsf{H}(\textbf{SLevel}) \cap \mathrm{dom}(Y)$$
$$\text{if } \{X,Y\} \vDash 5 \trianglelefteq Z(\textbf{GRecid}) \text{ and } \{X,Y\} \vDash 8 \trianglelefteq Z(\textbf{VRecid})$$

We have already verified that $\{X,Y\} \vDash 5 \trianglelefteq Z(\textbf{GRecid})$ holds, and $\{X,Y\} \vDash 8 \trianglelefteq Z(\textbf{VRecid})$ holds because $8 \trianglelefteq Z(\textbf{VRecid}) = 9$, and so we indeed have a lower-bound constraint $\{X,Y\} \vDash 3 \trianglelefteq Z(\textbf{SLevel})$.

Definition 3.9. Given a dimension d, a fact situation X, and a case base C, we say X is *d-inconsistent with respect to C* if there are values $v, w \in d$ with $v \lhd w$, such that both $C \vDash X(d) \trianglelefteq v$ and $C \vDash w \trianglelefteq X(d)$; otherwise X is *d-consistent*. The *d-consistency percentage* of C, denoted $\mathrm{Cons}_d(C)$, is the relative frequency of d-consistent cases in C:

$$\mathrm{Cons}_d(C) = \frac{|\{X \in C \mid X \text{ is } d\text{-consistent}\}|}{|C|} \tag{6}$$

Remark 3.10. Note that the usage of sets in this model is somewhat informal. For instance, in Example 3.4 we treat **GRecid** and **VRecid** as different dimensions, even though they should strictly speaking be considered identical as sets. We follow the tradition in the literature and refer to dimensions as sets in spite of these concerns (as e.g. Horty did in [15]), but strictly speaking it may be more accurate to speak of *multisets*, which are sets that can can contain multiple copies of an element. This will be particularly relevant for case bases, as the cardinality of a case base appears in the denominator of (6) in Definition 3.9, so we reiterate that case bases can contain multiple instances of a case.

Example 3.11. Reconsidering Table 1, it can be checked that X and Y are **GRecid**- and **SLevel**-consistent with respect to $\{X,Y\}$, so

$$\mathrm{Cons}_{\textbf{GRecid}}(\{X,Y\}) = \mathrm{Cons}_{\textbf{SLevel}}(\{X,Y\}) = 1.$$

Now, suppose we assigned a **GRecid** score of 3 to the fact situation Z, so $Z(\textbf{GRecid}) = 3$. We saw in Example 3.8 that $\{X, Y\} \models 5 \trianglelefteq Z(\textbf{GRecid})$, but now we also have $\{X, Y\} \models Z(\textbf{GRecid}) \trianglelefteq 3$ by Definition 3.7 as $Z(\textbf{GRecid}) = 3$. Therefore, since $3 \lhd 5 \in \textbf{GRecid}$, we see that Z would become **GRecid**-inconsistent with respect to the case base $\{X, Y\}$ as a result of the assignment $Z(\textbf{GRecid}) = 3$.

The rest of this work revolves around Definition 3.9 of case base consistency. To start, we derive a formal result about what can be expected of the consistency of decisions made by a program such as COMPAS, and then compare it to its actual consistency in the data that is available. This analysis will be similar to that of [32], though in the present work we will use the expanded version of the COMPAS dataset published by [31], rather than the original version published by [2], which has less of the input features that COMPAS uses to compute its scores. We use the implementation published by [32] for computing these consistency scores.

4 Consistency of Generalized Linear Models

It is unknown exactly how the COMPAS program works due to its proprietary nature, but its developers have previously indicated that the scores it produces are (at least partially) based on regression models [5, 4, 12, 17]. For example, [5] states that the "Recidivism Risk Scale is a regression model that has been used in COMPAS since 2000," and that "the COMPAS risk and classification models use logistic regression [...] in [...] prediction and classification procedures." Furthermore, the COMPAS Practitioner's Guide [12] states that "linear equations are used to calculate the [general recidivism and violent recidivism] risk scales," and that the violent recidivism risk score, the VRRS, is computed as the following weighted sum:[2]

$$\begin{aligned}\text{VRRS} = {} & (-w_1 \cdot \text{age}) + (-w_2 \cdot \text{age at first arrest}) \\ & + (w_3 \cdot \text{history of violence}) + (w_4 \cdot \text{vocation education}) \\ & + (w_5 \cdot \text{history of noncompliance})\end{aligned}$$

Despite these claims by the developers of COMPAS it is argued in [31], on the basis of a data analysis of the COMPAS dataset, that the scores assigned by COMPAS depend *nonlinearly* on age. Equivant disputed these claims in [17], and reiterated that the COMPAS risk scales make use of logistic regression. Rudin et al. responded in [30] that if COMPAS does operate on the basis of a logistic regression model, the age variable might first undergo a nonlinear transformation before being fed as input to the model.

As we can see, the question of whether the risk scores produced by COMPAS are the output of a relatively simple linear model has been the topic of debate. Indeed, Rudin [29] has argued that it is always better to use interpretable models, such as linear regression or logistic regression, for high stakes decision-making, rather than complex black box machine learning algorithms. In the case of COMPAS, it might be that it is only a black box because of its proprietary nature, and not because it makes use of uninterpretable machine learning algorithms such as neural networks. Ideally, we would be able to verify whether COMPAS is a linear model, without compromising Equivant's intellectual property protections.

In this section we will show that the model of a fortiori case-based reasoning which we reviewed in Section 3 can be used to falsify the claim that a given set of outputs were produced by a linear model. More specifically, we show that for a large class of linear models, called *generalized linear models* (GLMs) [20], a case base of model decisions is necessarily consistent in the sense of Definition 3.9.

We start by recalling the basic definition of a GLM, and give a concrete example in the form of a logistic regression model trained on the COMPAS dataset. This is a representative example, since it has been claimed that COMPAS is itself a form of logistic regression. We then prove a theorem stating that a case base of GLM decisions is necessarily fully consistent. Given this result, we would expect that the consistency of the COMPAS risk score assignments in the dataset made available by [2, 31] is high. We will show that quite the opposite is the case, and we discuss some possible causes of these low consistency percentages.

4.1 Regression Analysis

Regression analysis is a statistical modeling technique used to predict the expected value of a random variable $\boldsymbol{y}$ as a function of a set of observed values $\boldsymbol{x} = (x_1, \ldots, x_n)$. The simplest form of regression is *linear* regression, in which the expected value $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{x})$ of $\boldsymbol{y}$ given $\boldsymbol{x}$ corresponds to the linear combination of $\boldsymbol{x}$ with a vector of coefficients $\beta_0, \beta_1, \ldots, \beta_n$, so $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{x}) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$. This linear combination of the β_i and $\boldsymbol{x}_i$ is called the *linear predictor*. There are many variations on this idea, such as logistic regression, Poisson regression, gamma regression, and so forth.

Nelder and Wedderburn [20] showed that many forms of regression fit in a common class which they called *generalized linear models* (GLMs). A GLM assumes that the random variable $\boldsymbol{y}$ is distributed according to a member of the exponential family of probability distributions, and that the expected value of $\boldsymbol{y}$, conditioned on an observed set of values $\boldsymbol{x}$, is related to the linear predictor by a monotone *link function* g, i.e. that $g(\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{x})) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$ [9]. Note that linear regression is obtained as a GLM by using an identity link function $g(x) = x$. The inverse of the link function is often assumed to exist and is called the *mean function*, denoted by m, as it maps the linear predictor to the mean of the random variable.

The β_i coefficients of the GLM are often estimated from data using techniques such as maximum likelihood estimation [9, Chapter 4]. Once this is done, the GLM can be used for prediction. More specifically, given a new observation $(x_1, \ldots, x_n)$, the GLM can estimate a value $\hat{\boldsymbol{y}}$ of the target random variable $\boldsymbol{y}$ by simply applying the mean function to the linear predictor: $\hat{\boldsymbol{y}} = m(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)$. Note that, for the purpose of prediction, the choice of distribution for $\boldsymbol{y}$ is no longer relevant. Since we are primarily interested in their application to predictive modeling, we will not further consider this aspect of GLMs in this work, and assume that an n-ary GLM $(m, \boldsymbol{\beta})$ is parameterized by two components: a vector of $n+1$ real coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_n)$, and an n-ary monotone mean function m.

Example 4.1. Logistic regression is obtained as a GLM by choosing the sigmoid σ as the mean function:

$$m(x) = \sigma(x) = (1 + \exp(-x))^{-1} \tag{7}$$

We will now demonstrate how logistic regression can be used to produce a risk score such as the COMPAS general recidivism risk scale. We do so by fitting the $\boldsymbol{\beta}$ coefficients to a selection of variables from

[2] Though Equivant has later stated in [17] that this description should not be taken as a complete technical description of the violent recidivism risk scale.

Table 2: Example general recidivism risk assessments, according to the GLM $(\sigma, -0.04, -0.2, -0.03, 0.11, 0.55)$, where σ is defined in Eq. (7), and the parameters are estimated based on the data published by [31]. The example recidivism risk scores are computed according to Eq. (8).

MS	Age	Priors	Male	$\text{GRecid}_{\text{rs}}$
1	25	1	1	0.41
0	21	3	1	0.54
1	67	1	0	0.10
0	18	10	1	0.73
0	46	1	1	0.30

the COMPAS dataset published by [31], where the target variable is a binary indicator whether the person in question recidivated or not (see Section 2 and [31] for a more detailed description of the data). This is, presumably, representative of how the actual COMPAS risk scales were developed. For example [5] states that "The Recidivism Risk Scale is a regression model [...] that was trained to predict new offenses in a probation sample." This is very similar to the data contained in the COMPAS dataset published by [31], as it does not only contain COMPAS risk scores, but also contains labels stating whether the person in question recidivated or not.

We select five features from the dataset, corresponding to some of the dimensions in the hierarchy we discussed in Example 3.4: **MS**, **Age**, **Priors**, and **Male**. Here, the number of priors is given in the data as the number of offenses committed in the 30 days leading up to the COMPAS assessment. The target variable $\boldsymbol{y}$ is the binary label indicating whether the person committed a new crime within two years after the assessment. We used the default Python *Sci-kit learn* implementation to estimate the coefficients [25].

The resulting logistic regression model can be specified as the GLM $(\sigma, -0.04, -0.2, -0.03, 0.11, 0.55)$, which means that its prediction for values of the features **MS**, **Age**, **Priors**, **Male** is given by

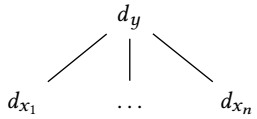
$$(1+\exp(-0.04-0.2\mathbf{MS}-0.03\mathbf{Age}+0.11\mathbf{Priors}+0.55\mathbf{Male}))^{-1} \quad (8)$$

Some example rows of the COMPAS dataset, together with their predicted (raw) general recidivism risk score (according to our example GLM), are displayed in Table 2. The $\mathbf{GRecid}_{\text{rs}}$ values are calculated according to Eq. (8).

4.2 The Consistency of GLM Decisions

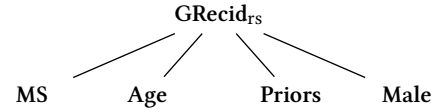
In this section we will prove our first main result regarding case base consistency, which states that the predictions made by a GLM are always fully consistent in the sense of Definition 3.9. To do this, we first show that any GLM can naturally be associated with a dimension hierarchy. Using this associated hierarchy, a dataset of GLM outputs can be translated to a case base C of which the consistency can be calculated with respect to its target variable. Theorem 4.6 below states that the consistency of such a dataset is necessarily equal to 1.

Consider an n-ary GLM $(m, \boldsymbol{\beta})$. We may, without loss of generality, assume that the coefficients β_i are all nonzero. This is because if any coefficient β_i were zero, the corresponding term $\beta_i x_i$ would not contribute to the linear predictor, and so the model's behavior would be the same as if that term were omitted. We define a set of dimensions $D = \{d_{x_i} \mid 1 \le i \le n\} \cup \{d_y\}$. Each d_{x_i} is the set of real numbers $\mathbb{R}$, and is ordered by $\le$ if $\text{sign}(\beta_i) = 1$, and by $\ge$ if $\text{sign}(\beta_i) = -1$. Similarly, d_y is the set of real numbers $\mathbb{R}$, and is ordered by $\le$ if m is order-preserving, and by $\ge$ if m is order-reversing. We order D by the structure $\mathsf{H} = \{(d_{x_i}, d_y) \mid 1 \le i \le n\}$:



Definition 4.2. Let $(m, \boldsymbol{\beta})$ be an n-ary GLM, and (D, H) its associated dimension hierarchy. An $(m, \boldsymbol{\beta})$-*decision* is a fact situation $X \in \mathcal{X}(D)$ with $\text{dom}(X) = D$, and $X(d_y) = m(\beta_0 + \sum_{i=1}^{n} \beta_i X(d_{x_i}))$. An $(m, \boldsymbol{\beta})$-*case base* is a case base $C \subseteq \mathcal{X}(D)$ of $(m, \boldsymbol{\beta})$-decisions.

Example 4.3. Consider the GLM $(\sigma, -0.04, -0.2, -0.03, 0.11, 0.55)$ of Example 4.1, and note that its associated hierarchy is given by:



(9)

The dimension orders are: $\ge$ for **MS** and **Age**, because their coefficients in the GLM are negative, and $\le$ for **Priors** and **Male**, because their coefficients are positive. The order for $\mathbf{GRecid}_{\text{rs}}$ is $\le$ because σ is (strictly) order-preserving. Relative to this hierarchy, the rows of Table 2 constitute a $(\sigma, -0.04, -0.2, -0.03, 0.11, 0.55)$-case base.

To prove our theorem, we will need two lemmas—the first of which states that GLM decisions naturally satisfy the a fortiori principle underlying the model of case-based reasoning.

LEMMA 4.4. *For an n-ary GLM $(m, \boldsymbol{\beta})$ and $(m, \boldsymbol{\beta})$-decisions X and Y: If $Y(d_{x_i}) \trianglelefteq X(d_{x_i})$ for all $1 \le i \le n$, then $Y(d_y) \trianglelefteq X(d_y)$. Similarly, if $X(d_{x_i}) \trianglelefteq Y(d_{x_i})$ for all $1 \le i \le n$, then $X(d_y) \trianglelefteq Y(d_y)$.*

PROOF. The first implication can be derived as follows:

$$Y(d_{x_i}) \trianglelefteq X(d_{x_i}) \text{ for } 1 \le i \le n \quad (10)$$

$$\text{implies} \quad \beta_i Y(d_{x_i}) \le \beta_i X(d_{x_i}) \text{ for } 1 \le i \le n \quad (11)$$

$$\text{implies} \quad \beta_0 + \textstyle\sum_{i=1}^{n} \beta_i Y(d_{x_i}) \le \beta_0 + \sum_{i=1}^{n} \beta_i X(d_{x_i}) \quad (12)$$

$$\text{implies} \quad m\left(\beta_0 + \textstyle\sum_{i=1}^{n} \beta_i Y(d_{x_i})\right) \trianglelefteq m\left(\beta_0 + \sum_{i=1}^{n} \beta_i X(d_{x_i})\right) \quad (13)$$

$$\text{implies} \quad Y(d_y) \trianglelefteq X(d_y). \quad (14)$$

Step (11) follows by definition of the dimension order of d_{x_i}: If $\beta_i > 0$ then $\trianglelefteq = \le$ so $Y(d_{x_i}) \le X(d_{x_i})$ and $\beta_i Y(d_{x_i}) \le \beta_i X(d_{x_i})$; while if $\beta_i < 0$ then $\trianglelefteq = \ge$, so $Y(d_{x_i}) \ge X(d_{x_i})$ and $\beta_i Y(d_{x_i}) \le \beta_i X(d_{x_i})$. Step (13) follows by monotonicity of m and the definition of the dimension order of d_y, and step (14) follows by Definition 4.2. The proof of the second implication is analogous so we omit it. □

The second lemma uses the first to show that when a GLM case base induces constraint on one of its decisions, then this is necessarily the results of disjunct (2) of Definition 3.7.

LEMMA 4.5. *For an n-ary GLM $(m, \boldsymbol{\beta})$, an $(m, \boldsymbol{\beta})$-case base C, some fact situation $X \in C$, and a value $v \in d_y$: If $C \vDash v \trianglelefteq X(d_y)$ then $v \trianglelefteq X(d_y)$. Similarly, if $C \vDash X(d_y) \trianglelefteq v$ then $X(d_y) \trianglelefteq v$.*

Proof. Assume $C \vDash v \trianglelefteq X(d_y)$, we proceed by a case distinction on the disjuncts (1)–(3) in Definition 3.7. We can rule out condition (1) because d_y does not have a minimal element. If condition (2) holds we are done immediately. Lastly, if condition (3) holds, then for some $Y \in C$ with $v \trianglelefteq Y(d_y)$ we have that $Y(d_{x_i}) \trianglelefteq X(d_{x_i})$ for all $1 \leq i \leq n$. Hence, we get $Y(d_y) \trianglelefteq X(d_y)$ from Lemma 4.4, and so $v \trianglelefteq X(d_y)$ by transitivity of $\trianglelefteq$. The proof of the second implication follows the same pattern, so we omit it. □

It is now straightforward to derive our first main result from Lemma 4.5: A case base of GLM decisions is always fully consistent.

Theorem 4.6. *If $(m, \boldsymbol{\beta})$ is a GLM, (D, H) its associated hierarchy, and $C \subseteq \mathcal{X}(D)$ an $(m, \boldsymbol{\beta})$-case base, then* $\mathrm{Cons}_{d_y}(C) = 1$.

Proof. Assume, for sake of contradiction, that $\mathrm{Cons}_{d_y}(C) < 1$; then there is a case $X \in C$ which is d_y-inconsistent. This means that both $C \vDash X(d_y) \trianglelefteq v$ and $C \vDash w \trianglelefteq X(d_y)$ for some values $v, w \in d_y$ with $v \triangleleft w$. By Lemma 4.5 this implies $X(d_y) \trianglelefteq v$ and $w \trianglelefteq X(d_y)$, so $X(d_y) \triangleleft X(d_y)$, meaning $X(d_y) \neq X(d_y)$—a contradiction. □

It is important to note that Theorem 4.6 relies on an accurate construction of the dimension hierarchy associated with the GLM. In practice, when the GLM is a black box, we cannot inspect the signs of its coefficients, or whether its mean function is order-preserving or -reversing. However, estimating the signs of the coefficients is a much easier task than estimating the precise values of the coefficients. Likewise, it is easier to determine whether the function connecting the linear predictor to the output is order-preserving or -reversing, than it is to precisely estimate the function itself.

4.3 The Consistency of the COMPAS Risk Scores

The consistency of the COMPAS risk scores was analyzed by van Woerkom et al. [32, Table 3] using the dataset published by Angwin et al. [2], and they found very low percentages. This is surprising, especially considering Theorem 4.6 and the claims that COMPAS uses linear models. Van Woerkom et al. suggested that the low scores are due to missing inputs in the dataset. This issue is well-known and is one of the main reasons Rudin et al. [31] created and published an extended version of the dataset. Therefore, it is logical to repeat the consistency analysis using this extended dataset.

We have used the implementation by [32] to compute the consistency scores, using the scale inputs as described in Section 2. The dimension orders we used are in agreement with the ones listed in [32, Table 2]; a complete overview can be found in our source code. The results of our analysis are, somewhat surprisingly, the same as those of [32]: The consistency of the raw **GRecid**, **VRecid**, and **FTA** risk scores are all 0%—the opposite of what one would expect based on Theorem 4.6.

There are multiple possible explanations for this. One possibility is that our estimations of the dimension orders are incorrect. However, we consider this unlikely, as the effects of the features involved are quite self-evident. For example, the majority of the features in the dataset correspond to criminal history, and it is clear that higher values of these features should generally lead to higher risk scores. Another possibility is that COMPAS depends on the features in a nonlinear manner, as hypothesized in [31]. However, we follow the assessment in [32] that the most likely cause is the absence of certain dimensions from the dataset. However, it remains surprising that the addition of some of these data by [31] did not improve the consistency scores.

The results of our consistency analysis thus suggest that the low consistency percentages of the COMPAS risk scores may be due to missing dimensions in the dataset. This raises a general question: What is the effect of adding or removing dimensions on the consistency of a case base? In the next section, we formally analyze the impact of modifications to a dimension hierarchy on case base consistency.

5 Consistency with Respect to Subhierarchies

Our definition of consistency operates on the basis of a set D of dimensions, which are assumed to exist and sufficiently describe the domain under consideration. Generally, this set is assumed to be static. This is an idealized assumption which does not always apply in practice. Of course, it is well known that in the domain of law, new cases can bring to light new factors or dimensions, thus expanding the set D. Likewise, we have seen in the case of COMPAS that some of the relevant dimensions may be unknown, and thus omitted from the set. Conversely, it may happen that the set D erroneously contains dimensions which are in fact not relevant to the domain. For example, it might be that there are features in the COMPAS risk score dataset which were in fact not used by COMPAS.

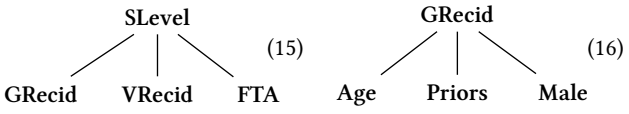
In this section and the next, we consider what the effects are of modifications to the dimension hierarchy on case base consistency. Specifically, in this section, we consider the modification of adding or removing dimensions. We will show that the omission of relevant dimensions decreases consistency, while the addition of (possibly irrelevant) dimensions increases consistency.

5.1 Subhierarchies

To start, we need a definition that captures the addition or removal of dimensions, for which we introduce the notion of subhierarchy.

Definition 5.1. Given a hierarchy (D, H), a *subhierarchy* (E, I) of (D, H), denoted $(E, \mathsf{I}) \subseteq (D, \mathsf{H})$, is a hierarchy satisfying $E \subseteq D$ and $\mathsf{I} \subseteq \mathsf{H}$. A subhierarchy $(E, \mathsf{I}) \subseteq (D, \mathsf{H})$ *preserves base-level dimensions* if any I-minimal element in E is H-minimal in D.

Example 5.2. The hierarchy in (1) has many subhierarchies, one of which is its top layer, depicted as (15) on the left below:



Note that the subhierarchy (15) does not preserve base-level factors, because **GRecid** is base-level in (15) but abstract in the full hierarchy (1). In contrast, its subhierarchy depicted as (16) on the right above does preserve base-level factors.

Given a subhierarchy $(E, \mathsf{I}) \subseteq (D, \mathsf{H})$, any case base $C \subseteq \mathcal{X}(D)$ can be made into a case base $\bar{C} \subseteq \mathcal{X}(E)$, simply by restricting the fact situations in C to the dimensions in E. This can be understood as an instance of the well-known concept of function restriction.

Definition 5.3. Given a function $f : X \to Y$ and a subset $Z \subseteq X$ the *restriction of* f *to* Z, denoted $f\restriction Z$, is the function $f\restriction Z : Z \to Y$ given by $(f\restriction Z)(z) = f(z)$ for all $z \in Z$.

Given a subhierarchy $(E, \mathsf{I}) \subseteq (D, \mathsf{H})$, any fact situation X of D can be made a fact situation $\bar{X} = X\restriction E$ of E. As such, a case base $C \subseteq \mathcal{X}(D)$ can be made a case base $\bar{C} = \{\bar{X} \mid X \in C\} \subseteq \mathcal{X}(E)$.

Example 5.4. The restrictions of the case base listed in Table 1 to the subhierarchies of Example 5.2 are given by removing columns from the table. For example, $X\restriction\{$**Age**, **Priors**, **Male**, **GRecid**$\}$ assigns **Age** to 25, **Priors** to 3, **Male** to 0, and **GRecid** to 7.

Remark 5.5. Note that the restriction operation is not injective: Given fact situations $X, Y \in D$ and a subhierarchy $(E, \mathsf{I}) \subseteq (D, \mathsf{H})$. we may have that $X \neq Y$ while $\bar{X} = \bar{Y}$. This means that if we do not allow multiple occurrances of a case in a case base it might be that $|\bar{C}| < |C|$, but since we *do* allow multiple occurances (cf. Remark 3.10) we have $|\bar{C}| = |C|$.

We now have the required terminology to state the second main result of our work: Restricting a case base to a subhierarchy decreases its consistency.

Theorem 5.6. *If* $(E, \mathsf{I}) \subseteq (D, \mathsf{H})$ *preserves base-level dimensions, then* $\mathrm{Cons}_e(\bar{C}) \leq \mathrm{Cons}_e(C)$ *for any case base* $C \subseteq \mathcal{X}(D)$ *and* $e \in E$.

Theorem 5.6 tells us that when dimensions are missing from the hierarchy, meaning that we are working with a restricted case base, then the consistency of that case base will be lower than it should be. In this regard, it formalizes the intuition that missing features from the COMPAS dataset can explain the low consistency scores we observed in Section 4.3. Conversely, this theorem tells us that the more dimensions that we do take into account, the higher the consistency will tend to be. Intuitively, the reason for this is that the more dimensions there are, the harder it is to satisfy Disjunct (3) of Definition 3.9. This means that there is less constraint, and as such there is less inconsistency, i.e., higher consistency.

This is a significant effect to take into consideration when discussing the consistency of the COMPAS scores, as this means that when we include features as dimensions, this can seemingly increase the consistency percentages of the scores—even when the feature in question was not an input to the program. Consider, as a concrete example, the "race" feature in the COMPAS dataset. This feature is not an input to the risk scores [8], but since it is part of the dataset, we could consider it as a dimension in our computation of the consistency of the risk scores. Theorem 5.6 shows that its inclusion would necessarily lead to an increase in the consistency of the risk scores, which might be mistaken for an indication that COMPAS does make use of this feature.

As discussed in Section 2, the COMPAS risk score dataset published by [31] contains data on a number of features related to inputs of the COMPAS risk scores. In total it contains 26 features. Each of the risk scores use only a subset of these, and we have computed the consistency scores only with respect to that subset [31, Table 1]. However, we can also use more of the features for this calculations, and Theorem 5.6 tells us that this increases the consistency of the scores. To demonstrate this, we successively added features as inputs to the model, and recomputed the resulting consistency scores. The results are shown in Figure 1. In accordance with Theorem 5.6, we see that the scores are monotonically increasing in the number of input dimensions.

6 Effects of Binning on Consistency

In the previous section we considered the effect of expanding or limiting a dimension hierarchy. In this section we consider a second type of modification to the hierarchy, namely that of data *binning*.

Data binning is a preprocessing technique used in machine learning and data science to convert continuous numeric variables into categorical ones by subdividing a range of values into smaller, consecutive, non-overlapping intervals called bins [22, Chapter 4]. A common application of data binning is the histogram, which visualizes the distribution of a dataset by replacing individual data points with their corresponding bins, thus smoothing the data and making general trends easier to see.

In this section, we prove two theorems regarding the effect of binning on case base consistency. Theorem 6.7 shows that binning output dimensions generally increases consistency, while Theorem 6.9 shows that binning input dimensions generally decreases consistency.

6.1 Dimension binning

To begin, we need a formal definition of data binning within our framework, for which we propose the following.

Definition 6.1. Let d be a dimension; a *binning* (bin, e) of d is a dimension e together with a surjective order-preserving function $\mathrm{bin} : d \to e$. The elements of e may be referred to as *bins*.

For the sake of convenience we will refer to just the function $\mathrm{bin} : d \to e$ as the 'binning' of d. The requirement that bin is order-preserving, which means that $v \trianglelefteq w$ implies $\mathrm{bin}(v) \trianglelefteq \mathrm{bin}(w)$, ensures that the order of the bins reflects the order of the original dimension d. The surjectivity requirement, which states that any bin $v \in e$ has a nonempty pre-image, ensures that all bins correspond to some region of the original dimension.

Example 6.2. An example of dimension binning in the context of COMPAS is given by the various presentations of the risk scores which we discussed in Section 2. Let $\mathbf{GRecid}_{\mathrm{rs}}$ denote the dimension corresponding to the raw version of the recidivism risk score, $\mathbf{GRecid}_{\mathrm{ds}}$ the decile version, and $\mathbf{GRecid}_{\mathrm{txt}}$ the textual version (i.e., with the possible values low, medium, and high). The conversion of the raw scores to decile scores corresponds to a binning $\mathrm{bin}_{\mathrm{ds}} : \mathbf{GRecid}_{\mathrm{rs}} \to \mathbf{GRecid}_{\mathrm{ds}}$. The way this mapping works depends on the specific norm group that is used for the conversion. Assuming the conversion specified in [12, Table 2.3] can compute the pre-images of the bins; for example, $\mathrm{bin}_{\mathrm{ds}}^{-1}(4) = (-0.7, -0.4]$, and $\mathrm{bin}_{\mathrm{ds}}^{-1}(5) = (-0.4, -0.2]$, etc. In turn, we have a binning converting deciles scores to text $\mathrm{bin}_{\mathrm{txt}} : \mathbf{GRecid}_{\mathrm{ds}} \to \mathbf{GRecid}_{\mathrm{txt}}$:

$$\mathrm{bin}_{\mathrm{txt}}^{-1}(\mathrm{low}) = [1, 4], \quad \mathrm{bin}_{\mathrm{txt}}^{-1}(\mathrm{med}) = [5, 7], \quad \mathrm{bin}_{\mathrm{txt}}^{-1}(\mathrm{high}) = [8, 10].$$

Example 6.3. In general, for any dimension d there is a trivial *identity* binning $\mathrm{id}_d : d \to d$ defined by $\mathrm{id}_d(v) = v$ for all $v \in d$. This corresponds to putting every value of d in its own unique bin.

We want to investigate the effect that binning one or more dimensions has on case base consistency. To this end, we will use the following definition. In this context, we regard a dimension d

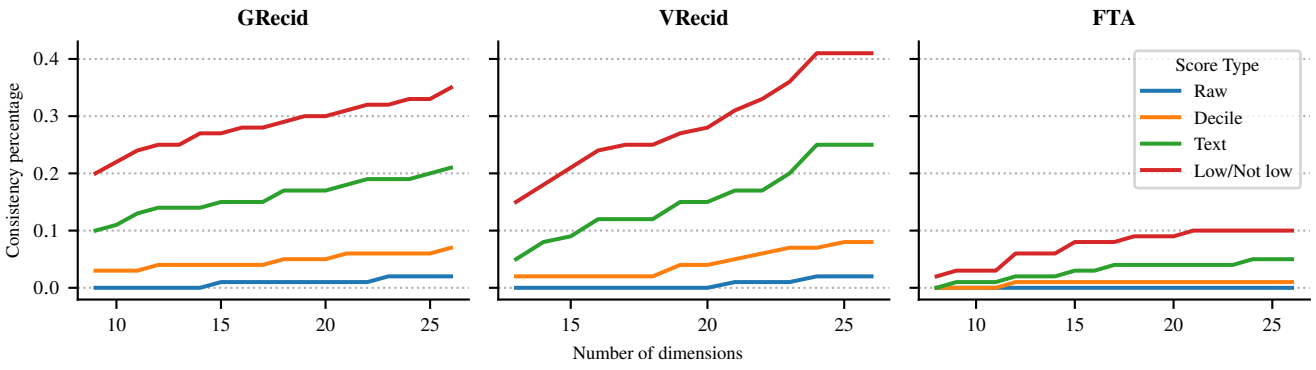


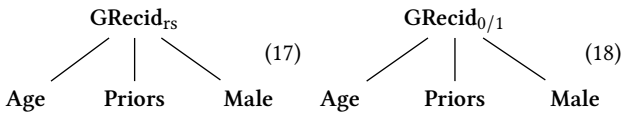
Figure 1: The consistency scores of various versions of the expanded COMPAS dataset published by [31], as a function of number of input dimensions and method of output binning. The "Low/Not Low" score type is a binary label indicating whether the textual representation is "Low" or not, and was considered as part of [2]. The lines are non-decreasing in the number of input dimensions as per per Theorem 5.6. Similarly, they are non decreasing with respect to the order "Raw < Decile < Text < Low/Not low" on the score types as per Theorem 6.7.

about which a decision is made as an *output* dimension, and the dimensions that influence said decision, i.e. those in $\mathsf{H}(d)$, as *input* dimensions.

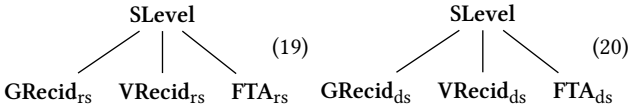
Definition 6.4. Given a set of dimensions D, a D-*binning* is an assignment of a binning $\mathrm{bin}_d : d \longrightarrow \underline{d}$ to every dimension $d \in D$. Furthermore, given $e \in D$, a D-binning is an *input binning of* e if $\mathrm{bin}_e = \mathrm{id}_e$, and an *output binning of* e if $\mathrm{bin}_d = \mathrm{id}_d$ whenever $d \neq e$.

Given a hierarchy (D, H) with a D-binning, the binned version of a dimension $d \in D$ is denoted by $\underline{d}$. Likewise, we will write $\underline{D} = \{\underline{d} \mid d \in D\}$ for the set of all binned dimensions. This set can be given the same hierarchical structure as the original hierarchy by defining $\underline{\mathsf{H}} = \{(\underline{e}, \underline{d}) \mid (e, d) \in \mathsf{H}\}$. There is a canonical way of transforming a fact situation $X \in \mathcal{X}(D)$ into a fact situation $\underline{X} \in \mathcal{X}(\underline{D})$ for the binned version of the hierarchy, by defining $\underline{X}(\underline{d}) = \mathrm{bin}_d(X(d))$. This operation extends in the obvious way to a case base $C \subseteq \mathcal{X}(D)$: We define $\underline{C} = \{\underline{X} \mid X \in C\} \subseteq \mathcal{X}(\underline{D})$.

Example 6.5. A logistic regression model, such as the one we considered in Example 4.1, is not a classifier in the machine learning sense because it outputs probabilities rather than binary labels. Still, it often used as such, simply by tresholding its output to 0 and 1. This operation corresponds to a binning $\mathrm{bin} : \mathbf{GRecid}_{\mathrm{rs}} \longrightarrow \mathbf{GRecid}_{0/1}$ defined by $\mathrm{bin}(x) = 0$ if $x < 0.5$ and $\mathrm{bin}(x) = 1$ if $x \geq 0.5$. This is an example of an output $\mathbf{GRecid}_{\mathrm{rs}}$ binning. Given the hierarchy (D, H) on the left below in (17), its corresponding $(\underline{D}, \underline{\mathsf{H}})$ is drawn on the right below as (18).



Example 6.6. An example of an input binning is given by the various presentations of the COMPAS risk scores, in their connection to the recommended supervision level score. Together the binnings $\mathbf{GRecid}_{\mathrm{rs}} \longrightarrow \mathbf{GRecid}_{\mathrm{ds}}$, $\mathbf{VRecid}_{\mathrm{rs}} \longrightarrow \mathbf{VRecid}_{\mathrm{ds}}$ and $\mathbf{FTA}_{\mathrm{rs}} \longrightarrow \mathbf{FTA}_{\mathrm{ds}}$ give an input **SLevel** binning, with associated hierarchies:



6.2 Output Binning Increases Consistency

We now have the required terminolog to state the third main result of this work: Output binnings increase case base consistency.

THEOREM 6.7. *Given a hierarchy (D, H), a case base $C \subseteq \mathcal{X}(D)$, and an output D-binning of $d \in D$, we have* $\mathrm{Cons}_d(C) \leq \mathrm{Cons}_{\underline{d}}(\underline{C})$.

Theorem 6.7 formalizes the intuition that the more fine-grained a dimension is, the easier it is for an inconsistency to arise. Again, this is important to realize when applying the notion of case base consistency in an AI context. For example, this theorem states that the consistency of the COMPAS risk scores seems to increase when we consider one of the binned versions of the scores, as discussed in Example 6.2. To demonstrate this effect, we have computed the COMPAS consistency risk scores for these various methods of binning, together with a final binary "Low/Not low" binning, which is 0 if the textual representation is "low" and 1 otherwise. These correspond to the presentation of the scores used by ProPublica in [2]. The results can be found in Figure 1, and indeed we see that the consistency scores monotonically increase as we consider increasingly coarse representations of the scores.

Example 6.8. As a final example we can reconsider Example 6.5, in which a logistic regression model is made a classifier through use of a binary output binning. We can now combine Theorems 4.6 and 6.7 to see that a case base of outputs of such a classifier is also necessarily fully consistent.

6.3 Input Binning Decreases Consistency

As the fourth and final formal result of this work, we show that, with respect to input binning, the effect opposite to that of Theorem 6.9 occurs: Input binnings decrease case base consistency.

Theorem 6.9. *Given a hierarchy* (D, H)*, a case base* $C \subseteq \mathcal{X}(D)$*, and an input* D*-binning of* $d \in D$*, we have* $\mathrm{Cons}_d(\underline{C}) \leq \mathrm{Cons}_d(C)$.

Intuitively, the reason for this result is that as we bin input dimensions, it becomes easier to satisfy disjunct (3) of Definition 3.7 of constraint. As such, more constraint is induced, and so there are more opportunities for inconsistencies to arise.

In [32] the consistency scores were computed of the COMPAS recommended level of supervision scores, with respect to both of the hierarchies depicted in (19) and (20). The former, which uses the raw scores, yielded a consistency percentage of 84%; while the latter, which uses the decile scores, yielded a score of 100%. As noted in [32] this should not be possible, and indeed this observation directly contradicts Theorem 6.9.

We note that a manual of the version of COMPAS used in the state of New York explicitly states how the recommended supervision levels are computed [21, Appendix C]. This description exactly matches the scores found in the Broward County, which explains why the consistency is higher when the decile versions of the scores were used. This also shows that the **FTA** score is not an input to the recommended supervision level.

6.4 Reconstructing the Norm Groups

It was hypothesized by van Woerkom et al. [32] that the reason for the discrepancy in the recommended supervision level scores was due to the use of multiple norm groups for the conversion of the raw scores to the decile scores, and Equivant has confirmed this to us upon enquiry [11]. In fact, this possibility was already observed by Dieterich et al. [8, Appendix B]. More recently, it was also remarked by Engel et al. [10, Appendix].

Decile scores computed with different norm groups should not be compared, as they are on different scales. However, many studies, including the original ProPublica publication [2], do exactly that. A more accurate approach uses raw scores, as in [31], but users prefer decile or textual scores [4]. The study in [10] renorms decile scores based on all raw scores, which lowers the average decile risk score and bases the study on hypothetical rather than actual COMPAS output.

To compare decile risk scores accurately, they should be split according to the norm groups used. We show that it is possible to reconstruct these norm groups from the Broward County data.

We estimate the number of norm groups by constructing a graph of COMPAS risk assessments, drawing edges between nodes with mutually inconsistent raw-to-decile score conversions. Applying a graph coloring algorithm [14], we identified three norm groups; see Table 3 for the cut-points. The **GRecid** and **FTA** graphs have fully connected subgraphs of size 3, indicating at least three norm groups for these scores. The **VRecid** scores were converted using a single norm group, but the cut-points listed in [32, Table 7] are incorrect—see Table 3 for the correct list.

We note that, because there is some overlap in the cut-points of the groups, defendants can be "shuffled around." For example, the size of the second largest group of the **GRecid** score could be increased to at least 6,698. Therefore, those wishing to analyze the recidivism scores and the outcome labels could analyze this group in isolation, without having to re-norm the deciles.

Some of the reconstructed groups consist of over over 99.5% males. Since the graph labeling algorithm we used was not in any way instructed to group defendants by sex (in fact, it did not even have access to this information), we consider it likely that gendered norm groups were used for the raw-to-decile conversion in the dataset. We have marked these groups in Table 3. Moreover, one of these groups (corresponding to the second row of Table 3) closely aligns with the cut-points given in [12, Table 2.3] for the GRRS. Due to this, we consider it likely that our reconstructed cut-points are good approximations of the true cut-points.

7 Conclusion

In this work, we analyzed the consistency of the COMPAS risk scores using a formal model of a fortiori case-based reasoning. We showed that for a large class of linear models, called generalized linear models (GLMs), a case base of model decisions is necessarily fully consistent. This suggests that if COMPAS were based on such models, its risk scores should exhibit high consistency. However, our analysis of the COMPAS dataset revealed very low consistency percentages, indicating that the scores may depend on features not present in the dataset or that the model used by COMPAS is more complex than a GLM.

We also investigated the effects of modifying the dimension hierarchy on case base consistency. We showed that omitting relevant dimensions decreases consistency, while adding dimensions increases it. Additionally, we demonstrated that binning output dimensions increases consistency, whereas binning input dimensions decreases it. This highlights the importance of considering the full set of relevant features and the appropriate level of granularity when analyzing the consistency of AI decisions.

Our findings suggest that the low consistency of the COMPAS risk scores may be due to missing input features in the dataset. Future work could focus on obtaining a more complete set of features and further investigating the nature of the model used by COMPAS. Additionally, our formal results on the effects of dimension hierarchy modifications and binning provide a foundation for further research on the robustness and reliability of case-based reasoning models in AI applications.

Lastly, Rudin et al. [31] noted that many defendants with numerous prior offenses received low risk scores in the COMPAS dataset. van Woerkom et al. [33] showed that such outlier cases, termed landmarks, can significantly increase inconsistency in a dataset. We suspect that these cases might overlap with those identified by Rudin et al. [31, Table 6], potentially explaining the low consistency percentages observed for the COMPAS scores. This suggests that inconsistency measures could be useful for detecting outliers in similar datasets.

References

[1] Vincent Aleven. 1997. *Teaching Case-Based Argumentation through a Model and Examples*. Ph.D. Dissertation. 267 pp. https://dl.acm.org/doi/10.5555/926270.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica*. Retrieved Feb. 26, 2024 from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Table 3: The reconstructed cut-points of norm groups used for the raw-to-decile score conversions in the COMPAS dataset. We indicate the highest score for each given decile, in accordance to [12, Table 2.3]. The dashes indicate that there were no defendants grouped in that particular decile of the hypothesized norm group, and so no information is available on its cut-point.

Score	Order	Male	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
GRecid	5,230	×	−1.66	−1.31	−1.03	−0.82	−0.59	−0.37	−0.14	0.12	0.43	2.36
	4,440	✓	−1.39	−0.92	−0.60	−0.39	−0.19	0.01	0.19	0.39	0.67	—
	2,840	✓	—	—	−0.92	−0.60	−0.39	−0.19	0.01	0.19	—	—
VRecid	12,510	×	−2.95	−2.56	−2.24	−1.98	−1.74	−1.50	−1.26	−1.00	−0.63	0.93
FTA	6,631	×	16	19	21	23	25	27	29	31	35	50
	4,950	×	—	16	19	21	22	24	26	28	31	35
	929	✓	—	—	—	—	23	25	27	29	—	—

[3] Berkman Klein Center for Internet and Society. 2025. The risk assessment tools database. https://criminaljustice.tooltrack.org/tool/16627.

[4] Tim Brennan and William Dieterich. 2018. Correctional Offender Management Profiles for Alternative Sanctions (COMPAS). In *Handbook of Recidivism Risk/Needs Assessment Tools*. John Wiley & Sons, Ltd, 49–75. ISBN: 978-1-119-18425-6. DOI: 10.1002/9781119184256.ch3.

[5] Tim Brennan, William Dieterich, and Beate Ehret. 2009. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, 36, 1, (Jan. 1, 2009), 21–40. DOI: 10.1177/0093854808326545.

[6] Ilaria Canavotto and John Horty. 2022. Piecemeal Knowledge Acquisition for Computational Normative Reasoning. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '22). Association for Computing Machinery, New York, NY, USA, (July 27, 2022), 171–180. ISBN: 978-1-4503-9247-1. DOI: 10.1145/3514094.3534182.

[7] Ilaria Canavotto and John Horty. 2023. Reasoning with hierarchies of open-textured predicates. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law* (ICAIL '23). ACM Press, (Sept. 7, 2023), 52–61. ISBN: 9798400701979. DOI: 10.1145/3594536.3595148.

[8] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Research report. Northpointe Inc. Research Department.

[9] Annette J. Dobson. 2001. *An Introduction to Generalized Linear Models*. (2nd ed.). Chapman and Hall/CRC, New York, (Nov. 28, 2001). 240 pp. ISBN: 978-1-4200-5768-3. DOI: 10.1201/9781420057683.

[10] Christoph Engel, Lorenz Linhardt, and Marcel Schubert. 2024. Code is law: How COMPAS affects the way the judiciary handles the risk of recidivism. *Artificial Intelligence and Law*, (Feb. 9, 2024). DOI: 10.1007/s10506-024-09389-8.

[11] Equivant. 2025. E-mail correspondence. (2025).

[12] Equivant. 2019. Practitioner's guide to COMPAS core. (Apr. 4, 2019). https://equivant-supervision.com/resources/white-papers-research-studies/.

[13] Anthony W. Flores, Kristin Bechtel, and Christopher T. Lowenkamp. 2016. False positives, false negatives, and false analyses: a rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.". *The Administrative Office of the U.S. Courts*. Federal Probation 80, 2, 38–46.

[14] Aric Hagberg, Pieter J. Swart, and Daniel A. Schult. 2008. Exploring Network Structure, Dynamics, and Function Using NetworkX. LA-UR-08-05495; LA-UR-08-5495. Los Alamos National Laboratory (LANL), Los Alamos, NM (United States), (Jan. 1, 2008). Retrieved Jan. 29, 2025 from https://www.osti.gov/biblio/960616.

[15] John Horty. 2019. Reasoning with dimensions and magnitudes. *Artificial Intelligence and Law*, 27, 3, (Sept. 1, 2019), 309–345. DOI: 10.1007/s10506-019-09245-0.

[16] John Horty. 2004. The result model of precedent. *Legal Theory*, 10, 1, (Mar. 2004), 19–31. DOI: 10.1017/S1352325204000151.

[17] Eugenie Jackson and Christina Mendoza. 2020. Setting the record straight: What the COMPAS core risk and need assessment is and is not. *Harvard Data Science Review*, 2, 1, (Jan. 31, 2020). DOI: 10.1162/99608f92.1b3dadaa.

[18] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. ProPublica. Retrieved May 16, 2024 from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[19] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2021. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Vol. 1. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021.

[20] J. A. Nelder and R. W. M. Wedderburn. 1972. Generalized Linear Models. *The Journal of the Royal Statistical Society*. A 135, 3, (May 1, 1972), 370–384. DOI: 10.2307/2344614.

[21] New York State Division of Criminal Justice Services, Office of Probation and Correctional Alternatives. 2015. Practitioner Guidance for Probation and Community Corrections Agencies. (2015). https://apps.criminaljustice.ny.gov/opca/pdfs/2015-5-NYCOMPAS-Guidance-August-4-2015.pdf.

[22] Robert Nisbet, Gary Miner, and Ken Yale. 2018. *Handbook of Statistical Analysis and Data Mining Applications*. (Second ed.). Academic Press, (Jan. 1, 2018). ISBN: 978-0-12-416632-5. DOI: 10.1016/C2012-0-06451-4.

[23] Daphne Odekerken and Floris Bex. 2020. Towards transparent human-in-the-loop classification of fraudulent web shops. In *Legal Knowledge and Information Systems. JURIX 2020: The Thirty-third Annual Conference*. Serena Villata, Jakub Harašta, and Petr Křemen, (Eds.) IOS Press, 239–242. DOI: 10.3233/FAIA200873.

[24] Daphne Odekerken, Floris Bex, and Henry Prakken. 2024. Precedent-based reasoning with incomplete information for human-in-the-loop decision support. *Artificial Intelligence and Law*, (Dec. 12, 2024). DOI: 10.1007/s10506-024-09421-x.

[25] Fabian Pedregosa et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 85, 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html.

[26] Joeri G.T. Peters, Floris Bex, and Henry Prakken. 2023. Model- and data-agnostic justifications with a fortiori case-based argumentation. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law* (ICAIL '23). Association for Computing Machinery, New York, NY, USA, 207–216. ISBN: 9798400701979. DOI: 10.1145/3594536.3595164.

[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. 1, 2022), 159–194. DOI: 10.3233/AAC-210009.

[28] Henry Prakken and Giovanni Sartor. 1998. Modelling reasoning with precedents in a formal dialogue game. *Artificial Intelligence and Law*, 6, 2, (June 1, 1998), 231–287. DOI: 10.1023/A:1008278309945.

[29] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 5, (May 2019), 206–215. DOI: 10.1038/s42256-019-0048-x.

[30] Cynthia Rudin, Caroline Wang, and Beau Coker. 2020. Broader Issues Surrounding Model Transparency in Criminal Justice Risk Scoring. *Harvard Data Science Review*, 2, 1, (Jan. 31, 2020). DOI: 10.1162/99608f92.038c43fe.

[31] Cynthia Rudin, Caroline Wang, and Beau Coker. 2020. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2, 1, (Jan. 31, 2020). DOI: 10.1162/99608f92.6ed64b30.

[32] Wijnand van Woerkom, Davide Grossi, Henry Prakken, and Bart Verheij. 2024. A Case-Based-Reasoning Analysis of the COMPAS Dataset. In *Legal Knowledge and Information Systems. JURIX 2024: The Thirty-seventh Annual Conference*. IOS Press, (Dec. 2024).

[33] Wijnand van Woerkom, Davide Grossi, Henry Prakken, and Bart Verheij. 2024. *A Fortiori* case-based reasoning: From theory to data. *Journal of Artificial Intelligence Research*, 81, (Oct. 2024), 401–441. DOI: 10.1613/jair.1.15178.

[34] Wijnand van Woerkom, Davide Grossi, Henry Prakken, and Bart Verheij. 2023. Hierarchical *a fortiori* reasoning with dimensions. In *Legal Knowledge and Information Systems. JURIX 2023: The Thirty-sixth Annual Conference*. Giovanni Sileno, Jerry Spanakis, and Gijs van Dijck, (Eds.) IOS Press, (Dec. 2023), 43–52. DOI: 10.3233/FAIA230944.

8 Proofs

Here we provide the proofs of Theorems 5.6, 6.7, and 6.9.

The notation that we use for constraint, $C \vDash v \trianglelefteq X(d)$, does not explicitly refer to the dimension hierarchy (D, H) with respect to which the constraint is defined, instead leaving this to be inferred from the context. We do this to avoid clutter, but since the proofs in this appendix relate constraint statements between different hierarchies we caution that some confusion may arise.

As a notational convenience, we will write $C, Y \vDash v \trianglelefteq X(d)$ as a shorthand for disjunct (3) of Definition 3.7. More specifically, the notation $C, Y \vDash v \trianglelefteq X(d)$ means that Y is a fact situation $Y \in C$ such that d is an abstract dimension, Y satisfies $v \trianglelefteq Y(d)$, and $C \vDash Y(e) \trianglelefteq X(e)$ holds for all $e \in \mathsf{H}(d) \cap \mathrm{dom}(Y)$.

Lastly, we note that most of the proofs in this appendix use structural induction on the hierarchical structure of Definition 3.3. This technique is a generalization of natural number induction, and is applicable here because we assume a dimension hierarchy is finite and does not contain loops. It works as follows; in order to prove that every dimension d in a dimension hierarchy (D, H) satisfies a property P, meaning $P(d)$ holds, it suffices to prove the following two statements:

(1) For any base-level dimension $b \in D$, $P(b)$ holds.
(2) For any abstract dimension $a \in D$, if $P(e)$ holds for every $e \in \mathsf{H}(a)$ then $P(a)$ holds.

These statements are the base case and the induction case of the induction proof, respectively. Given an abstract dimension a, the assumption that "$P(e)$ holds for every $e \in \mathsf{H}(a)$" is called the *induction hypothesis*.

8.1 Proof of Theorem 5.6

Lemma 8.1. *Consider a subhierarchy $(E, \mathsf{I}) \subseteq (D, \mathsf{H})$ that preserves base-level dimensions, a value v in a dimension $e \in E$, a fact situation $X \in \mathcal{X}(D)$, and a case base $C \subseteq \mathcal{X}(D)$; if $C \vDash v \trianglelefteq X(e)$ then $\bar{C} \vDash v \trianglelefteq \bar{X}(e)$. Similarly, $C \vDash X(e) \trianglelefteq v$ implies $\bar{C} \vDash \bar{X}(e) \trianglelefteq v$.*

Proof. We proceed by structural induction on the position of e in the hierarchy (E, I). We make a case distinction on $C \vDash v \trianglelefteq X(e)$.

(1) If v is the least element of e, then $\bar{C} \vDash v \trianglelefteq \bar{X}(e)$ by definition.
(2) If $v \trianglelefteq X(e) = \bar{X}(e)$ then, again by definition, $\bar{C} \vDash v \trianglelefteq \bar{X}(e)$.
(3) Suppose $e \in A$, $Y \in C$ with $v \trianglelefteq Y(e)$, and $C \vDash Y(d) \trianglelefteq X(d)$ for all $d \in \mathsf{H}(d) \cap \mathrm{dom}(Y)$. The induction hypothesis states that for any $f \in \mathsf{I}(e)$ and $w \in f$: if $C \vDash w \trianglelefteq X(f)$ then $\bar{C} \vDash w \trianglelefteq \bar{X}(f)$. We want to show that $\bar{C}, \bar{Y} \vDash v \trianglelefteq \bar{X}(e)$. Note that since e is an abstract dimension in (D, H), it is still an abstract dimension in (E, I) because it preserves base-level dimensions. Furthermore, since $v \trianglelefteq Y(e)$, also $v \trianglelefteq \bar{Y}(e)$. Let $f \in \mathsf{I}(e) \cap \mathrm{dom}(\bar{Y})$, we are done if $\bar{C} \vDash \bar{Y}(f) \trianglelefteq \bar{X}(f)$. Note that $\mathsf{I}(e) \cap \mathrm{dom}(\bar{Y}) \subseteq \mathsf{H}(e) \cap \mathrm{dom}(Y)$, and so by assumption $C \vDash Y(d) \trianglelefteq X(d)$. It thus follows from the induction hypothesis that $\bar{C} \vDash Y(d) \trianglelefteq \bar{X}(d)$, and $\bar{C} \vDash \bar{Y}(d) \trianglelefteq \bar{X}(d)$ by definition of $\bar{Y}$. □

Theorem (5.6). *If $(E, \mathsf{I}) \subseteq (D, \mathsf{H})$ preserves base-level dimensions, then $\mathrm{Cons}_e(\bar{C}) \leq \mathrm{Cons}_e(C)$ for any $C \subseteq \mathcal{X}(D)$ and $e \in E$.*

Proof. It suffices to show that if a fact situation $X \in \mathcal{X}(D)$ is e-inconsistent with respect to C, then $\bar{X} \in \mathcal{X}(E)$ is e-inconsistent with respect to $\bar{C}$ (recall that we are using multisets, so $|C| = |\bar{C}|$); but this is immediate from Lemma 8.1: If $v \lhd w \in e$ such that $C \vDash X(e) \trianglelefteq v$ and $C \vDash w \trianglelefteq X(e)$, then $\bar{C} \vDash \bar{X}(e) \trianglelefteq v$ and $\bar{C} \vDash w \trianglelefteq \bar{X}(e)$ and so $\bar{X}$ is indeed e-inconsistent. □

8.2 Proof of Theorem 6.9

Lemma 8.2. *Consider a hierarchy (D, H) with a D-binning. For a dimension $d \in D$, a value $v \in d$, a fact situation $X \in \mathcal{X}(D)$, and a case base $C \subseteq \mathcal{X}(D)$: If $C \vDash v \trianglelefteq X(d)$ then $\underline{C} \vDash \mathrm{bin}(v) \trianglelefteq \underline{X}(\underline{d})$; and similarly, if $C \vDash X(d) \trianglelefteq v$ then $\underline{C} \vDash \underline{X}(\underline{d}) \trianglelefteq \mathrm{bin}(v)$.*

Proof. We proceed by structural induction on the position of d in H, and apply a case distinction on $C \vDash v \trianglelefteq X(d)$.

(1) If v is the least element of d, then $\mathrm{bin}(v)$ is the least element of $\underline{d}$ because bin is order-preserving and surjective, so that indeed $\underline{C} \vDash \mathrm{bin}(v) \trianglelefteq \underline{X}(\underline{d})$.
(2) Likewise, if $v \trianglelefteq X(d)$ then because bin is order-preserving we have $\mathrm{bin}(v) \trianglelefteq \mathrm{bin}(X(d)) = \underline{X}(\underline{d})$, so $\underline{C} \vDash \mathrm{bin}(v) \trianglelefteq \underline{X}(\underline{d})$.
(3) If $C, Y \vDash v \trianglelefteq X(d)$, then $d \in A$, $v \trianglelefteq Y(d)$, and for all $e \in \mathsf{H}(d) \cap \mathrm{dom}(Y)$: $C \vDash Y(e) \trianglelefteq X(e)$. The induction hypothesis states that for all $w \in e \in \mathsf{H}(d)$: $C \vDash w \trianglelefteq X(e)$ implies $\underline{C} \vDash \mathrm{bin}(w) \trianglelefteq \underline{X}(\underline{e})$. We are done if $\underline{C}, \underline{Y} \vDash \mathrm{bin}(v) \trianglelefteq \underline{X}(\underline{d})$. Note that $\underline{d}$ is abstract in $(\underline{D}, \underline{\mathsf{H}})$, and that $\mathrm{bin}(v) \trianglelefteq \mathrm{bin}(Y(d)) = \underline{Y}(\underline{d})$ as bin is order-preserving. Therefore, it only remains to show that for all $\underline{e} \in \underline{\mathsf{H}}(\underline{d}) \cap \mathrm{dom}(\underline{Y})$: $\underline{C} \vDash \underline{Y}(\underline{e}) \trianglelefteq \underline{X}(\underline{e})$, so consider such $\underline{e}$. By definition of $\underline{\mathsf{H}}$ and $\underline{Y}$, this means that $e \in \mathsf{H}(d) \cap \mathrm{dom}(Y)$ and so $C \vDash Y(e) \trianglelefteq X(e)$, which means $\underline{C} \vDash \underline{Y}(\underline{e}) \trianglelefteq \underline{X}(\underline{e})$ follows from the induction hypothesis.

The proof of the other implication is analogous, so we omit it. □

Theorem (6.9). *Given a hierarchy (D, H), a case base $C \subseteq \mathcal{X}(D)$, and an input D-binning of $d \in D$, we have $\mathrm{Cons}_d(\underline{C}) \leq \mathrm{Cons}_d(C)$.*

Proof. It suffices to show that if $X \in C$ is d-inconsistent with respect to C, then $\underline{X}$ is d-inconsistent with respect to $\underline{C}$; so consider such $X \in C$. This means there are $v \lhd w \in d$ with $C \vDash X(d) \trianglelefteq v$ and $C \vDash w \trianglelefteq X(d)$. Lemma 8.2 gives us $\underline{C} \vDash \underline{X}(d) \trianglelefteq v$ and $\underline{C} \vDash w \trianglelefteq \underline{X}(d)$, so $\underline{X}$ is indeed d-inconsistent with respect to $\underline{C}$. □

8.3 Proof of Theorem 6.7

Definition 8.3. Given a dimension hierarchy (D, H), a subhierarchy (E, I) of (D, H) is *wide* if $\mathsf{I}(e) = \mathsf{H}(e)$ for all $e \in E$.

Note that a wide subhierarchy must also preserve base-level dimensions. If a subhierarchy is wide then we can obtain a converse of Lemma 8.1, as the following lemma shows.

Lemma 8.4. *Consider a wide subhierarchy $(E, \mathsf{I}) \subseteq (D, \mathsf{H})$, a value v in a dimension $e \in E$, a fact situation $X \in \mathcal{X}(D)$, and a case base $C \subseteq \mathcal{X}(D)$: If $\bar{C} \vDash v \trianglelefteq \bar{X}(e)$ then $C \vDash v \trianglelefteq X(e)$, and similarly, if $\bar{C} \vDash \bar{X}(e) \trianglelefteq v$ then $C \vDash X(e) \trianglelefteq v$.*

Proof. Assume that $\bar{C} \vDash v \trianglelefteq \bar{X}(e)$; we proceed by structural induction on the position of e in I, and apply a case distinction on $\bar{C} \vDash v \trianglelefteq \bar{X}(e)$:

(1) If v is the least element of e then $C \vDash v \trianglelefteq X(e)$.
(2) If $v \trianglelefteq \bar{X}(e)$ then $v \trianglelefteq X(e)$ so $C \vDash v \trianglelefteq X(e)$.

(3) If $\bar{C}, \bar{Y} \vDash v \trianglelefteq \bar{X}(e)$ then e is abstract so we have an induction hypothesis stating that for all $w \in f \in \mathsf{I}(e)$: If $\bar{C} \vDash w \trianglelefteq \bar{X}(f)$ then $C \vDash w \trianglelefteq X(f)$. We are done if we have $C, Y \vDash v \trianglelefteq X(e)$. Since e is abstract in (E, I) it must be abstract in (D, H). Furthermore, since $v \trianglelefteq \bar{Y}(e)$ we have $v \trianglelefteq Y(e)$, and so it remains only to show that for all $f \in \mathsf{H}(e) \cap \mathrm{dom}(Y)$: $C \vDash Y(f) \trianglelefteq X(f)$. Consider such f, then $f \in \mathsf{I}(e) \cap \mathrm{dom}(\bar{Y})$ because (E, I) is wide and so $H(e) = \mathsf{I}(e)$. It thus follows from $\bar{C}, \bar{Y} \vDash v \trianglelefteq \bar{X}(e)$ that $\bar{C} \vDash \bar{Y}(f) \trianglelefteq \bar{X}(f)$, and from the induction hypothesis that $C \vDash Y(f) \trianglelefteq X(f)$ as desired.

The proof of the second implication is analogous, so we omit it. □

A straightforward way of obtaining a (wide) subhierarchy from an existing one is to restrict the hierarchy to a specific dimension and include everything below it.

Definition 8.5. The *restriction* of a hierarchy (D, H) to a dimension $d \in D$, denoted $(D, \mathsf{H}) \restriction d$, is a subhierarchy of (D, H) given by (E, I) where E is given by $E = \{e \in D \mid \mathsf{H}^+(e, d)\}$, with H^+ denoting the transitive closure of H, and $\mathsf{I} = (E \times E) \cap \mathsf{H}$.

Lemma 8.6. *Consider a hierarchy (D, H), an output D-binning of some $d \in D$, a value $v \in d$, fact situations $X, Y \in \mathcal{X}(D)$ and a case base $C \subseteq \mathcal{X}(D)$: If $\underline{C}, \underline{Y} \vDash \mathrm{bin}(v) \trianglelefteq \underline{X}(\underline{d})$ then $C, Y \vDash Y(d) \trianglelefteq X(d)$, and similarly, if $\underline{C}, \underline{Y} \vDash \underline{X}(\underline{d}) \trianglelefteq \mathrm{bin}(v)$ then $C, Y \vDash X(d) \trianglelefteq Y(d)$.*

Proof. Assume that $\underline{C}, \underline{Y} \vDash \mathrm{bin}(v) \trianglelefteq \underline{X}(\underline{d})$. This means $\underline{d}$ is abstract, and therefore so is d. Clearly $Y(d) \trianglelefteq Y(d)$, so it remains to be shown that for all $e \in \mathsf{H}(d) \cap \mathrm{dom}(Y)$: $C \vDash Y(e) \trianglelefteq X(e)$, so consider such an e. We know that $\underline{C} \vDash \underline{Y}(e) \trianglelefteq \underline{X}(e)$ by assumption. Note that $(\underline{D}, \underline{\mathsf{H}}) \restriction e = (D, \mathsf{H}) \restriction e$ because we are considering an output binning of D. It is easy to see that these restrictions are wide and preserve base-level dimensions, and so it follows by applications Lemmas 8.1 and 8.4 to $\underline{C} \vDash \underline{Y}(e) \trianglelefteq \underline{X}(e)$ that $C \vDash Y(e) \trianglelefteq X(e)$, as desired. □

Theorem (6.7). *Given a hierarchy (D, H), a case base $C \subseteq \mathcal{X}(D)$, and an output D-binning of $d \in D$, we have $\mathrm{Cons}_d(C) \leq \mathrm{Cons}_{\underline{d}}(\underline{C})$.*

Proof. Let $X \in C$, it suffices to show that if $\underline{X}$ is $\underline{d}$-inconsistent then X is d-inconsistent. To this end, assume that $\underline{X}$ is $\underline{d}$-inconsistent; then, as bin is surjective, there are $v, w \in d$ with $\mathrm{bin}(v) \triangleleft \mathrm{bin}(w)$ such that $\underline{C} \vDash \underline{X}(\underline{d}) \trianglelefteq \mathrm{bin}(v)$ and $\underline{C} \vDash \mathrm{bin}(w) \trianglelefteq \underline{X}(\underline{d})$. By applying a case distinction to $\underline{C} \vDash \underline{X}(\underline{d}) \trianglelefteq \mathrm{bin}(v)$ we can see that there must exist some $Y \in C$ such that $\underline{Y}(\underline{d}) \trianglelefteq \mathrm{bin}(v)$ and $C \vDash X(d) \trianglelefteq Y(d)$:

(1) This case can be ruled out: If $\mathrm{bin}(v)$ is the greatest element of $\underline{d}$ then $\mathrm{bin}(w) \trianglelefteq \mathrm{bin}(v)$, which contradicts $\mathrm{bin}(v) \triangleleft \mathrm{bin}(w)$.
(2) In this case $\underline{X}(\underline{d}) \trianglelefteq \mathrm{bin}(v)$ and so there does exist such a witness $Y \in C$, namely X itself.
(3) Now $\underline{C}, \underline{Y} \vDash \underline{X}(\underline{d}) \trianglelefteq \mathrm{bin}(v)$, which by definition means $\underline{Y}(\underline{d}) \trianglelefteq \mathrm{bin}(v)$, and $C \vDash X(d) \trianglelefteq Y(d)$ follows by Lemma 8.6.

By similar reasoning we can deduce from $\underline{C} \vDash \mathrm{bin}(w) \trianglelefteq \underline{X}(\underline{d})$ that there is some $Z \in C$ with $\mathrm{bin}(w) \trianglelefteq \underline{Z}(\underline{d})$ and $C \vDash Z(d) \trianglelefteq X(d)$. To conclude that X is d-inconsistent it thus remains only to show that $Y(d) \triangleleft Z(d)$. Note that $\underline{Y}(d) \trianglelefteq \mathrm{bin}(v) \triangleleft \mathrm{bin}(w) \trianglelefteq \underline{Z}(\underline{d})$, so $\underline{Y}(\underline{d}) \triangleleft \underline{Z}(\underline{d})$. This implies that $Y(d) \triangleleft Z(d)$, because otherwise totality of $\trianglelefteq$ would give $Z(d) \trianglelefteq Y(d)$ and $\underline{Z}(\underline{d}) \trianglelefteq \underline{Y}(\underline{d})$ by the fact that bin is order preserving, contradicting $\underline{Y}(\underline{d}) \triangleleft \underline{Z}(\underline{d})$. □

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009