Toward a Defeasible Semantics for Symbolic Classifiers

Ruvarashe Madzime[1]

[1]University of the Western Cape and CAIR, Cape Town, South Africa

Abstract

When a machine learning model produces rules like "feature a predicts label 1, except when b is also present," it looks like defeasible reasoning. But looking like it is not the same as being it. This thesis asks: under what conditions does a learned rule-based classifier actually behave like a defeasible theory?

We answer this for a class of classifiers called Exception Closed Conjunctive Rule Sets (ECCRS). When a learned rule set satisfies a single structural condition, called strict global exception closure, its predictions are shown to agree exactly with three well-known defeasible reasoning frameworks: Rational Closure, Lexicographic Closure, and System W. This gives a formal justification for reading the classifier's output as defeasible knowledge, not just as a list of patterns. We also show that strict global exception closure is the only condition that needs to be checked directly, since the other required conditions either follow from it or can be enforced by a simple pruning step that does not change any predictions.

The results open a path toward what we call defeasible machine learning: designing learners that do not just produce readable rules, but produce rules with a principled defeasible meaning built in.

Keywords

defeasible reasoning, nonmonotonic reasoning, symbolic machine learning, interpretable machine learning, rational closure, lexicographic closure, System W

1. Introduction and Motivation

This MSc thesis investigates how symbolic machine learning can be linked to defeasible reasoning. A symbolic learner may produce rules such as

$$a \Rightarrow 1, \qquad a \wedge b \Rightarrow 0, \qquad a \wedge b \wedge c \Rightarrow 1. \tag{1}$$

A person reading these rules does not normally treat them as a flat list of independent predictions. Instead, the rules are read as a structured pattern of defaults and exceptions: feature a normally supports label 1, except when b is present, unless c is also present. The rules look like defeasible knowledge in the sense of Kraus, Lehmann, and Magidor [1].

This resemblance is not accidental. Many real-world classification problems are genuinely defeasible in structure, and a classifier that formalises that structure is more trustworthy than one that merely approximates it. Consider tumour diagnosis: a pathologist examining cellular features reasons that uniform cell size normally indicates a benign tumour, *unless* prominent nucleoli are also present, in which case the conclusion is overridden. Credit assessment follows the same pattern: a clean repayment history normally supports approval, *unless* a recent default is on record. Toxicity classification works similarly: a particular cap shape normally indicates an edible mushroom, *unless* the odour profile is foul. In each of these domains, the exception does not sit alongside the default as an independent rule. It overrides it. A classifier that produces rules with this shape but does not formally behave as a defeasible theory gives no guarantee that the override will be applied correctly, and no principled basis for trusting its explanations.

This is one reason why symbolic models are often seen as useful in interpretable machine learning [2, 3]. They do not just produce outputs; they appear to produce reasons. Rule-based learners such as RIPPER [4], certifiably optimal rule lists [5], and work in inductive logic programming [6] all benefit

Doctoral Consortium of the 23rd International Conference on Principles of Knowledge Representation and Reasoning (KR 2026 DC), July 20-23, 2026, Lisbon, Portugal

madzimeruvarashe@gmail.com (R. Madzime)

0009-0001-3469-9353 (R. Madzime)

from the fact that learned symbolic structures are easier for humans to read and discuss than opaque models [2, 3]. But readability alone does not answer the semantic question. A rule can be easy to read without being semantically well-behaved as defeasible knowledge [1]. The appearance of defeasible structure and the reality of it are very different things, and conflating the two understates how much formal work remains to be done.

This thesis starts from that gap. On one side is symbolic machine learning, where the main goal is to learn compact and useful rule-based models from data [4, 5, 6]. On the other side is defeasible reasoning, where the main goal is to decide how conclusions should be drawn when there are defaults, exceptions, and changes in specificity [1, 7, 8, 9]. The broader goal of this thesis is to investigate whether these two areas can be brought together at the level of formal semantics, so that a learned classifier can be shown not merely to resemble a defeasible theory, but to provably be one.

The main answer developed in this thesis is a new class of classifiers called Exception Closed Conjunctive Rule Sets (ECCRS). Informally, an ECCRS is a conjunctive rule set whose compatible opposite-label conflicts are organised by specificity, so that they behave like genuine defaults and exceptions rather than arbitrary competitors. This thesis builds directly on earlier empirical work [10] which introduced a predecessor defeasible classifier and showed that this kind of rule structure can match or outperform leading interpretable baselines on precisely these kinds of domains. ECCRS formalises and extends that framework by identifying the precise structural condition under which learned rules admit a formal defeasible semantics. The earlier work showed the idea is empirically viable. This thesis shows it can be made theoretically exact.

The larger goal goes beyond ECCRS alone. If one can identify the structural conditions under which a learned rule set admits a principled defeasible reading, those conditions can guide the design of future learners. In that sense, this thesis is a step toward defeasible machine learning, where learned models are judged not only by predictive quality and human readability, but also by whether their structure supports a well-founded defeasible semantics [1, 9, 3].

2. Research Problem and Thesis Questions

The core research question of this thesis is whether a learned symbolic classifier can be understood as a defeasible theory, rather than merely as a predictive rule set.

The thesis answers this by identifying a structural class of rule sets for which the connection can be made exact. It shows that when a learned conjunctive rule set has the exception-closed structure captured by ECCRS, its decision operator, called Most Specific Wins (MSW), agrees with established KLM-style entailment mechanisms. The formal problem is therefore not simply whether a rule learner can produce readable rules, but whether the structure of those rules is strong enough to support a faithful defeasible semantics.

The thesis pursues this through four linked questions. The first asks what structural property a learned conjunctive rule set must satisfy in order for its classifier behaviour to admit a principled translation into a KLM-style defeasible knowledge base. The second asks when the classifier decision $\mathrm{MSW}_{\mathcal{R}}(F) = \ell$ agrees with defeasible entailment in the translated theory, that is, with conclusions of the form $\mathcal{K}(\mathcal{R}) \mid\!\approx F \mid\!\sim \ell$, where $\mid\!\approx$ ranges over three well-known KLM-style entailment relations: Rational Closure ($\mid\!\approx_{\mathrm{RC}}$) [7], Lexicographic Closure ($\mid\!\approx_{\mathrm{LC}}$) [8, 11], and System W ($\mid\!\approx_{\mathrm{W}}$) [9]. The third asks what such agreement reveals about the structure of defaults, exceptions, and counter-exceptions in the learned rule set. The fourth asks whether, when agreement fails outside this structural class, the form of that failure can be used as a diagnostic tool and whether those insights can help guide the design of future learners with defeasibly meaningful behaviour.

Beyond the formal contribution, knowing the structural condition has a direct practical consequence. If a learner can be designed to enforce strict global exception closure during training, then the resulting classifier comes with a built-in guarantee: every prediction it makes is a defeasible entailment, not just a pattern match. In domains like clinical diagnosis or credit assessment, where the override structure of a rule has direct consequences for the decision it justifies, this turns interpretability from a claim

about readability into a claim about meaning. The condition also serves as a diagnostic: if a learned rule set fails to satisfy it, the specific failure points show exactly where and why the rules fall short of principled defeasible behaviour, guiding either repair or redesign.

2.1. How the classifier works

Let $A = \{a_1, \ldots, a_n\}$ be propositional atoms for the feature language. A rule is a pair (B, ℓ) where B is a consistent conjunction of signed literals over A and $\ell \in \{0, 1\}$ is the label. A complete feature pattern F is a complete assignment over A, identified with the conjunction of its true signed literals.

For a learned rule set $\mathcal{R}$, the applicable rules at F are $\mathsf{A}_{\mathcal{R}}(F) := \{(B, \ell) \in \mathcal{R} : F \models B\}$, and the most specific applicable rules are

$$\mathsf{M}_{\mathcal{R}}(F) := \{(B, \ell) \in \mathsf{A}_{\mathcal{R}}(F) : \nexists (C, \ell') \in \mathsf{A}_{\mathcal{R}}(F) \text{ with } B \subset C\}.$$

The classifier predicts by looking only at these most specific rules. If they all agree on a label ℓ, that label is returned; if no rules apply, the classifier abstains. This is the Most Specific Wins operator:

$$\mathrm{MSW}_{\mathcal{R}}(F) = \begin{cases} \ell & \text{if } \mathsf{M}_{\mathcal{R}}(F) \neq \emptyset \text{ and all rules in } \mathsf{M}_{\mathcal{R}}(F) \text{ have label } \ell, \\ \bot & \text{if } \mathsf{M}_{\mathcal{R}}(F) = \emptyset. \end{cases}$$

The three-rule chain in Equation (1) illustrates this directly. If $F = a \wedge b \wedge c$, the unique most specific rule gives $\mathrm{MSW}(F) = 1$. If $F' = a \wedge b \wedge \neg c$, the most specific rule gives $\mathrm{MSW}(F') = 0$. The override is resolved by specificity alone. Under the structural conditions introduced in the next section, the translated defeasible theory yields exactly the same labels for the same reason.

3. What Has Been Established So Far

The current thesis work has established a concrete correspondence for ECCRS. The central structural condition is what we call strict global exception closure.

Definition 1 (Strict global exception closure)**.** *A rule set $\mathcal{R}$ satisfies strict global exception closure if for every pair $(B, \ell), (C, 1-\ell) \in \mathcal{R}$ of compatible opposite-label rules, either $B \subset C$ or $C \subset B$.*

This requires every compatible cross-label conflict to lie on a single specificity chain, so opposite-label rules may disagree only in a genuine default-and-exception pattern, not as incomparable competitors. A rule set satisfying this condition is what the thesis calls an ECCRS. It is precisely the structural property that separates a classifier whose rules merely resemble defeasible knowledge from one whose rules provably are defeasible knowledge.

Two additional conditions are useful in stating the initial result. Sanity requires that there are no pairs $(B, 0), (B, 1) \in \mathcal{R}$ with the same body. No total override requires that every rule remains operationally live, meaning that for each $(B, \ell) \in \mathcal{R}$ there is some feature pattern $F \models B$ such that $(B, \ell) \in \mathsf{M}_{\mathcal{R}}(F)$.

Under these conditions, the first established result is that the exception structure of the learned rule set, measured by exception depth, matches the BaseRank structure of the translated defeasible theory exactly. The second is the main semantic correspondence: the MSW prediction agrees with defeasible entailment under $\mid\!\!\sim_{\mathrm{RC}}$, $\mid\!\!\sim_{\mathrm{LC}}$, and $\mid\!\!\sim_{\mathrm{W}}$ simultaneously. The third is that the assumption set reduces to one: strict global exception closure alone is sufficient, since sanity follows from it and no total override is enforced by a prediction-preserving pruning procedure.

Taken together, these results answer the main question: there is a structurally identified class of learned rule-based models for which a defeasible reading is formally justified. The classifier does not just look like a defeasible theory; under this condition, it provably is one.

4. Current Correspondence Results at a High Level

4.1. Translation and rank

Each rule body becomes the antecedent of a defeasible conditional and the label becomes the consequent. Using a label atom l, the translated knowledge base is $\mathcal{K}(\mathcal{R}) := \{ B \mathrel{|\!\sim} l : (B, 1) \in \mathcal{R} \} \cup \{ B \mathrel{|\!\sim} \neg l : (B, 0) \in \mathcal{R} \}$, with materialisation $\overrightarrow{\mathcal{K}(\mathcal{R})} := \{ B \to l : (B, 1) \in \mathcal{R} \} \cup \{ B \to \neg l : (B, 0) \in \mathcal{R} \}$.

Ranks are built by successive removal: let $E_0 := \overrightarrow{\mathcal{K}(\mathcal{R})}$ and $E_{i+1} := \{\alpha \to \beta \in E_i : E_i \models \neg\alpha\}$. The BaseRank is $\mathrm{br}_{\mathcal{K}(\mathcal{R})}(\alpha) := \min\{i : E_i \not\models \neg\alpha\}$. On the classifier side, exception depth plays the same role: for $(B, \ell) \in \mathcal{R}$, let $\mathrm{Opp}(B) := \{ C : (C, 1 - \ell) \in \mathcal{R},\ C \subset B \}$ and define depth recursively as $\mathrm{depth}(B) = 0$ if $\mathrm{Opp}(B) = \emptyset$, and $1 + \max\{\mathrm{depth}(C) : C \in \mathrm{Opp}(B)\}$ otherwise. For a pattern F, $\mathrm{depth}(F)$ is the maximum depth of any applicable rule body.

Theorem 1 (Depth coincides with semantic rank)**.** *Assume strict global exception closure, sanity, and no total override. Then for each rule body B and each covered complete feature pattern F,*

$$\mathrm{depth}(B) = \mathrm{br}_{\mathcal{K}(\mathcal{R})}(B) \qquad \textit{and} \qquad \mathrm{depth}(F) = \mathrm{br}_{\mathcal{K}(\mathcal{R})}(F).$$

This is the structural backbone of everything that follows. The exception hierarchy visible in the learned rule set is not merely analogous to the BaseRank hierarchy of the translated theory; the two are identical. This shared structure is exactly what makes the correspondence possible.

4.2. Alignment across RC, LC, and System W

Rational Closure is the most conservative of the three: it concludes $\alpha \mathrel{|\!\sim} \beta$ when the world where α holds but β fails is strictly more exceptional, by BaseRank, than the world where both hold [7]. Lexicographic Closure refines this by comparing, for each rank level, how many defaults of that rank each world violates, breaking ties from the most exceptional rank downward; this allows it to recover default inheritance that Rational Closure can block [8, 11]. System W takes a different approach: rather than counting violations at each rank, it compares worlds by the sets of conditionals they falsify within each layer of the ordered tolerance partition of the knowledge base, so a world that falsifies a strict subset of defaults at the highest layer of disagreement is always preferred [9]. The fact that MSW agrees with all three simultaneously is therefore a strong result.

Theorem 2 (Current alignment result)**.** *Let $\mathcal{R}$ be an ECCRS satisfying strict global exception closure, sanity, and no total override, and let $\mathcal{K}(\mathcal{R})$ be its translated defeasible knowledge base. For every covered complete feature pattern F:*
If $\mathrm{MSW}_{\mathcal{R}}(F) = 1$, then $\mathcal{K}(\mathcal{R}) \mathrel{|\!\approx}_{\mathrm{RC}} F \mathrel{|\!\sim} l$, $\mathcal{K}(\mathcal{R}) \mathrel{|\!\approx}_{\mathrm{LC}} F \mathrel{|\!\sim} l$, $\mathcal{K}(\mathcal{R}) \mathrel{|\!\approx}_{\mathrm{W}} F \mathrel{|\!\sim} l$, and $\neg l$ is not entailed by any of the three semantics.
If $\mathrm{MSW}_{\mathcal{R}}(F) = 0$, then $\mathcal{K}(\mathcal{R}) \mathrel{|\!\approx}_{\mathrm{RC}} F \mathrel{|\!\sim} \neg l$, $\mathcal{K}(\mathcal{R}) \mathrel{|\!\approx}_{\mathrm{LC}} F \mathrel{|\!\sim} \neg l$, $\mathcal{K}(\mathcal{R}) \mathrel{|\!\approx}_{\mathrm{W}} F \mathrel{|\!\sim} \neg l$, and l is not entailed by any of the three semantics.

The label that MSW selects is exactly the label that $\mathrel{|\!\approx}_{\mathrm{RC}}$, $\mathrel{|\!\approx}_{\mathrm{LC}}$, and $\mathrel{|\!\approx}_{\mathrm{W}}$ would each derive independently. The classifier and the three defeasible reasoning frameworks are all saying the same thing on every covered pattern. This is not an approximation; it is a provable consequence of the structural condition that defines ECCRS.

4.3. Reducing the assumptions

The three conditions can be reduced to one. If sanity fails at body B, MSW returns $\perp$ and the translated theory forces B to rank ∞, so both sides agree on having nothing to say. Sanity can therefore be dropped as an independent assumption: wherever it fails, both sides remain consistent with each other regardless of what other conditions hold.

No total override is handled by a pruning procedure. Call a rule (B, ℓ) *covered* in S if every completion of B already satisfies some strictly more specific opposite-label rule in S, that is, if $B \models \bigvee\{C :$

$(C, 1-\ell) \in S,\ B \subset C\}$. Such a rule can never be the most specific applicable rule on any pattern, so it never determines an MSW prediction. Starting from $R_0 := \mathcal{R}$, define

$$R_{i+1} := R_i \setminus \{(B, \ell) \in R_i : (B, \ell) \text{ is covered in } R_i\},$$

and let $\mathcal{R}' := \bigcap_{i \geq 0} R_i$ be the fixed point. Since covered rules never fire, removing them preserves every MSW prediction. The procedure terminates because $\mathcal{R}$ is finite, and the fixed point satisfies no total override by construction. Strict global exception closure is therefore the only condition that needs to be imposed directly.

Theorem 3 (Generalised alignment under closure alone)**.** *If $\mathcal{R}$ satisfies strict global exception closure, then there exists a pruning fixed point $\mathcal{R}'$ such that:*

(a) $\mathcal{R}'$ preserves every MSW prediction of $\mathcal{R}$,

(b) $\mathcal{R}'$ satisfies sanity, guaranteed by strict global exception closure,

(c) the correspondence theorem holds for $\mathcal{R}'$.

There is a single structural condition that, when satisfied, guarantees the full triple alignment. Every other assumption either comes for free or can be enforced without changing classifier behaviour. Strict global exception closure is the one thing that genuinely matters.

5. Learning ECCRS and Experiments

5.1. How the learning procedure works

The theoretical results above hold for any rule set satisfying strict global exception closure. We now describe a practical method for learning such rule sets from data.

The procedure starts from a standard decision tree trained on the data. A decision tree makes predictions by testing one feature at a time along each branch until a leaf is reached, and the path from the root to any node can be read as a conjunction of feature conditions. At each node, the training instances that reach it have a majority label, and this majority label can change as the tree grows deeper. When that happens, the deeper node naturally plays the role of an exception: feature a normally supports label 1, except when b is also present.

Rules are extracted not only from the leaves but from every node where the majority label differs from the label of its parent. Each such node produces a rule whose body is the conjunction of feature tests along the path from the root, and whose label is the majority label at that node. Leaf nodes always produce a rule. After extraction, duplicate bodies are removed and totally overridden rules are pruned.

The key point is that extracting rules from all levels of the tree, not just the leaves, is what produces the default-and-exception structure. If only leaf rules were used, no rule could ever strictly extend another, because any two leaves on different branches contain contradicting feature values. By also extracting from internal nodes, the rule set includes both the general pattern and its more specific override.

Strict global exception closure holds by construction for any rule set extracted this way. Two rules on the same root-to-leaf path are related by specificity, since the deeper node adds more feature tests. Two rules on different branches have incompatible bodies, because at the branching point one includes $a_j = 0$ and the other includes $a_j = 1$, so the closure condition does not apply to them. No two compatible, opposite-label, incomparable bodies can therefore exist.

5.2. Experimental results

To confirm that ECCRS rule sets are learnable and competitive in practice, we evaluated the learning procedure on three datasets from the UCI Machine Learning Repository [12] under 10-fold stratified

Table 1
10-fold CV mean (± std). ECCRS rule and exception pair counts are averaged across folds. Neither CART, RIPPER, nor OneR produces any exception pairs.

Dataset	System	Accuracy	F1
Cancer	ECCRS	$.946 \pm .015$	$.941 \pm .016$
	CART	$.957 \pm .016$	$.953 \pm .017$
	RIPPER	$.950 \pm .019$	$.946 \pm .021$
	OneR	$.896 \pm .036$	$.887 \pm .037$
Mushroom	ECCRS	$.999 \pm .001$	$.999 \pm .001$
	CART	$1.00 \pm .000$	$1.00 \pm .000$
	RIPPER	$1.00 \pm .000$	$1.00 \pm .000$
	OneR	$.887 \pm .007$	$.886 \pm .007$
Bank	ECCRS	$.897 \pm .002$	$.620 \pm .011$
	CART	$.867 \pm .004$	$.679 \pm .011$
	RIPPER	$.889 \pm .002$	$.565 \pm .013$
	OneR	$.893 \pm .001$	$.615 \pm .009$

cross-validation. These datasets were chosen because their classification tasks are naturally defeasible in structure: the Breast Cancer Wisconsin dataset involves cytological features where individual markers support a diagnosis by default but are overridden in combination; the Mushroom dataset involves morphological features where shape or colour normally predicts edibility unless overridden by odour or gill characteristics; and the Bank Marketing dataset involves customer attributes where general patterns hold by default but are qualified by specific financial conditions.

The baselines are three standard symbolic classifiers: CART [13] with unrestricted depth (the canonical decision tree learner and the direct source of the trees ECCRS extracts from), RIPPER [4] (a rule induction system that produces ordered rule lists), and OneR [14] (a single-rule baseline). Decision trees are trained using scikit-learn [15]. Table 1 reports mean accuracy and macro-averaged F1 score across folds. Because Bank Marketing is heavily imbalanced (88.3% majority class), F1 provides a more informative picture than accuracy alone on that dataset.

On Cancer, ECCRS achieves 0.946 accuracy with an average of 13 rules and 8 exception pairs, within two percentage points of the best baseline (CART at 0.957). On Mushroom, ECCRS reaches 0.999 with 14 rules and 8 exception pairs, matching the top baselines. On Bank, ECCRS achieves 0.897 accuracy with 38 rules and 6 exception pairs, compared to unrestricted CART at 0.867 with over 6000 rules. The exception-closed hierarchy appears to act as a form of regularisation, preventing the overfitting visible in the unrestricted tree. The F1 scores reveal an additional pattern on Bank: all systems struggle with the minority class, but ECCRS (0.620) outperforms both RIPPER (0.565) and OneR (0.615), while CART achieves the highest F1 (0.679) by trading precision for recall across its 6000+ rules.

Neither CART nor RIPPER produces any exception pairs, and this is structural rather than accidental. CART extracts rules only from leaves, so no rule can ever strictly extend another. RIPPER checks rules sequentially and stops at the first match, producing a priority ordering rather than nested specificity chains. ECCRS is the only system whose rules carry formal defeasible structure.

6. Remaining Challenges and Next Steps

The established results describe the well-behaved case. The next phase of the thesis studies what happens when a learned rule set falls outside the ECCRS structural class, and how such failures should be analysed and handled.

When strict global exception closure fails, alignment can break in two distinct ways. The first is incomparable conflict. Consider $\mathcal{R}_1 = \{(a, 1), (b, 0)\}$ and the pattern $F = a \wedge b$. Both rules are maximally specific at F, they disagree on the label, and MSW returns $\bot$. In the translated theory

$\mathcal{K}(\mathcal{R}_1) = \{a \mathrel{|\!\sim} l,\ b \mathrel{|\!\sim} \neg l\}$, the conjunction $a \wedge b$ is not governed by a single default-and-exception chain, so the correspondence by specificity is lost. The second is total override, illustrated by $\mathcal{R}_2 = \{(a, 1), (a \wedge b, 0), (a \wedge \neg b, 0)\}$, where $(a, 1)$ is never the most specific applicable rule on any completion of a, so it is operationally dead on the classifier side while still present in the translated theory. Understanding these failure modes is what allows the structural condition to serve as a diagnostic, identifying precisely where and why a learned rule set falls short of defeasible behaviour.

The correspondence theorems hold for complete feature patterns, where every feature value is observed. When features are missing, the ECCRS classifier falls back to the most specific rule whose antecedent is fully satisfied by the observed values. This is a natural defeasible response: without evidence that an exception applies, reasoning proceeds from the more general default. However, this behaviour does not in general match any of the three formalisms, since RC, LC, and System W each handle incomplete information differently. Determining which formalism, if any, matches this fallback behaviour under partial observation is a clear direction for future work and could be tested empirically on data with missing values.

7. Related Work

This thesis sits at the intersection of nonmonotonic reasoning and symbolic machine learning.

On the reasoning side, the KLM framework [1] is the foundation. Rational Closure and Lexicographic Closure are two well-known ways of making that account computational [7, 8, 11]. System W extends this by defining a preference relation based on how defaults are verified and falsified, avoiding limitations of weaker ranked approaches such as System Z [9]. What is new here is not a new framework, but a formal bridge: a structural condition on learned rule sets under which their predictions are guaranteed to coincide with all three frameworks.

On the machine learning side, the case for interpretable models is well established [2, 3]. In high-stakes settings, a model whose reasoning can be followed step by step is preferable to one explained after the fact using post-hoc tools such as LIME or SHAP [16, 17]. Rule-based learners sit at the centre of this argument: RIPPER [4], certifiably optimal rule lists [5], and inductive logic programming [6] all produce compact, readable rule sets. A predecessor to ECCRS, the Defeasible Horn Classifier with Exceptions [10], showed empirically that a defeasible rule structure learned via Answer Set Programming can match or outperform leading interpretable baselines on domains with natural default-and-exception structure, but without formal semantic guarantees. What none of these approaches provide is a formal guarantee that the learned rules behave like defeasible knowledge in any principled sense.

Recent work on symbolic explainability [18] provides formal foundations for model interpretability using symbolic languages. ECCRS contributes to this direction by grounding rule-based classifiers in defeasible semantics, so that interpretability is not just a property of the rule format but a consequence of the classifier's formal behaviour.

This is the gap the thesis fills. It turns interpretability from a qualitative claim into a semantic one: saying that a model is interpretable because it produces rules is not the same as saying those rules provably behave as a defeasible theory. The first is an observation about form, the second a guarantee about meaning. A model in the ECCRS class does not just look like it is reasoning with defaults and exceptions; it provably is.

Acknowledgments

I thank my supervisors, Professor Louise Leenen (University of the Western Cape and CAIR) and Professor Thomas Meyer (University of Cape Town and CAIR), for their support and guidance throughout this work. This work is based on the research supported in part by the National Research Foundation of South Africa (REFERENCE NO: SAI240823262612). This work is also supported in part by funding from the Mastercard Foundation Scholars Program.

Declaration on Generative AI

During the preparation of this work, the author used Claude (Anthropic) in order to: Paraphrase and reword, Improve writing style. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] S. Kraus, D. Lehmann, M. Magidor, Nonmonotonic reasoning, preferential models and cumulative logics, Artificial Intelligence 44 (1990) 167–207. doi:10.1016/0004-3702(90)90101-5.

[2] A. A. Freitas, Comprehensible classification models: A position paper, ACM SIGKDD Explorations Newsletter 15 (2014) 1–10. doi:10.1145/2594473.2594475.

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

[4] W. W. Cohen, Fast effective rule induction, in: Proceedings of the 12th International Conference on Machine Learning (ICML), Morgan Kaufmann, 1995, pp. 115–123.

[5] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, C. Rudin, Learning certifiably optimal rule lists for categorical data, Journal of Machine Learning Research 18 (2018) 1–78.

[6] S. Muggleton, L. De Raedt, Inductive logic programming: Theory and methods, Journal of Logic Programming 19–20 (1994) 629–679. doi:10.1016/0743-1066(94)90035-3.

[7] D. Lehmann, M. Magidor, What does a conditional knowledge base entail?, Artificial Intelligence 55 (1992) 1–60. doi:10.1016/0004-3702(92)90041-U.

[8] D. Lehmann, Another perspective on default reasoning, Annals of Mathematics and Artificial Intelligence 15 (1995) 61–82. doi:10.1007/BF01535841.

[9] C. Komo, C. Beierle, Nonmonotonic reasoning from conditional knowledge bases with system W, Annals of Mathematics and Artificial Intelligence 90 (2022) 107–144. doi:10.1007/s10472-021-09777-9.

[10] R. S. Madzime, L. Leenen, T. Meyer, An override-aware classifier for transparent AI, in: Proceedings of the Southern African Conference for Artificial Intelligence Research (SACAIR 2025), 2025. Online Proceedings, Vol. II. ISBN: 978-1-0370-5280-4.

[11] A. Kaliski, An Overview of KLM-Style Defeasible Entailment, Master's thesis, University of Cape Town, 2020. URL: http://hdl.handle.net/11427/32743.

[12] D. Dua, C. Graff, UCI machine learning repository, 2019. URL: https://archive.ics.uci.edu/ml.

[13] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and Regression Trees, Wadsworth, 1984.

[14] R. C. Holte, Very simple classification rules perform well on most commonly used datasets, Machine Learning 11 (1993) 63–91. doi:10.1023/A:1022631118932.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, Édouard Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[16] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.

[17] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems 30, 2017, pp. 4765–4774.

[18] M. Arenas, D. Báez, P. Barceló, J. Pérez, B. Subercaseaux, Foundations of symbolic languages for model interpretability, in: Advances in Neural Information Processing Systems 34 (NeurIPS 2021), 2021, pp. 11690–11701.