

Project Proposal: Explanation Using Defeasible Rulesets

Antony Baker**
BKRANT001@myuct.ac.za
University of Cape Town
Cape Town, Western Cape, South
Africa

Daniel Schlesinger††
SCHDAN066@myuct.ac.za
University of Cape Town
Cape Town, Western Cape, South
Africa

Augustine Mochoeneng‡‡
MCHKOK003@myuct.ac.za
University of Cape Town
Cape Town, Western Cape, South
Africa

Abstract

Defeasible reasoning formalizes the intuition that conclusions hold by default but may be retracted when more specific information becomes available. Exception Closed Conjunctive Rule Sets embody this pattern in a symbolic classifier whose predictions are determined entirely by structural specificity, the most specific rule applicable wins, yielding explanation traces that are faithful and traceable by construction. In this project, we aim to build a workbench that formally certifies these explanations by translating the ECCRS ruleset into a KLM-style knowledge base, a framework grounded in preferential semantics that represents typicality via ranked possible worlds, and verifying whether three prominent defeasible entailment relations, namely Rational Closure, System W, and Lexicographic Closure, each produce predictions consistent with the ECCRS Most Specific Wins classifier. Agreement across these relations would provide formal grounding for ECCRS explanations; disagreement would expose the limits of that correspondence.

1 Introduction

Machine learning models are being increasingly deployed in consequential settings such as medical diagnosis, credit assessment, and criminal risk profiling. In these areas, merely providing a prediction alone is not sufficient. Users, regulators, and affected individuals need to understand not just what a model decided, but why. Post-hoc explanation methods, such as surrogate models, attempt to address this by approximating a black-box model with a simpler, interpretable one. However, as Rudin [17] argues, a surrogate is not the model it describes, and in high-stakes environments this gap is unacceptable. Symbolic rule-based classifiers offer an alternative. By representing decisions as explicit if-then rules over meaningful features, they make every prediction traceable by design, without requiring any additional explanation machinery. However, existing symbolic approaches introduce their own structural problems. Decision lists resolve conflicts through rule ordering, meaning that reordering the same rules can change the prediction [16]. Unordered rule sets avoid this but require external tie-breaking mechanisms [5] (such as scoring, voting, or hand-tuned priorities) that lie outside the rules themselves and therefore outside the explanation. Exception Closed Conjunctive Rule Sets (ECCRS) [14] address both problems by encoding the relationship between defaults and exceptions directly in the structure of the rules. When two opposite-label rules can both apply to the same instance, one must be strictly more specific than the other. Prediction then follows a Most Specific Wins (MSW) principle: the most specific applicable rules determine the label. This makes explanations short, stable under reordering, and

based off a visible chain of defaults and overrides that reflects how a human would justify the decision. An ECCRS ruleset can be translated into a KLM-style conditional knowledge base [10], and the same classification queries can then be answered using standard defeasible entailment methods. In our case this will include Rational Closure, Lexicographic Closure, and System W. An alignment theorem [14] guarantees that Rational Closure agrees exactly with MSW predictions under three structural conditions on the ruleset. Whether this agreement extends to the richer entailment methods, and what any divergence would reveal about the limits of MSW, are open questions. This project builds a translation and verification workbench that answers those questions. Starting from an already-learned ECCRS ruleset, the workbench validates the structural assumptions required for translation, constructs the corresponding KLM knowledge base, and runs all three closure methods against the same classification queries as MSW. For each prediction, it produces explanation traces on both sides, namely the ECCRS default-exception narrative and the formal defeasible entailment, making the verification legible rather than purely symbolic.

2 Related Work

2.1 The KLM Framework

The KLM framework, introduced by Kraus, Lehmann, and Magidor [10], provides the formal foundation for defeasible reasoning used in this project. Rather than proposing a single reasoning system, KLM characterises families of non-monotonic consequence relations by the properties they satisfy. The core idea is that classical logic is too rigid for real-world reasoning because it reinforces that once a conclusion follows from a set of premises, it holds no matter what new information is added. Defeasible reasoning relaxes this by allowing conclusions to be overridden when more specific information is available. For example, rather than saying birds *always* fly, we say birds *typically* fly, and reserve the right to override this when we know we are dealing with a penguin.

The semantic core of the framework is the preferential model, which imposes a plausibility ordering over possible worlds. A defeasible conditional $\alpha \sim \beta$ holds if all the most plausible worlds satisfying α also satisfy β . Lehmann and Magidor [13] refine this with ranked preferential models, assigning each world a rank representing how typical or exceptional it is. Lower ranks correspond to more normal situations and higher ranks to more exceptional ones, providing a clean and principled basis for resolving conflicts between defaults and exceptions.

2.2 Rational Closure

Rational Closure [13] is the most widely used entailment method in the KLM framework and the starting point for this project. It ranks

*
†
‡

defeasible conditionals by how exceptional their antecedents are, and answers queries by temporarily discarding the most general information when it conflicts with more specific knowledge. The ranks are computed by the BaseRank algorithm [8], which iteratively removes conditionals whose antecedents are exceptional at the current level until no further removals are possible. A query $\alpha \sim \beta$ is entailed when knowing α places us at a strictly lower rank than knowing α together with $\neg\beta$. Intuitively, β is what we would *typically* expect given α .

Rational Closure is computationally feasible, requiring only a polynomial number of consistency checks with respect to the size of the knowledge base. It is well understood, but has a known limitation, it does not allow exceptional subclasses to inherit generic properties from their superclass. When a subclass is ranked as exceptional, all of its superclass properties become inaccessible, even those entirely unrelated to the source of its exceptionality. The classic example is the penguin.

Knowing that penguins do not fly causes Rational Closure to rank penguins as exceptional birds [8], which can prevent it from concluding that penguins have wings even though nothing in the knowledge base suggests any conflict between penguins and wings. This over-blocking motivates the stronger entailment methods described below.

2.3 System W

System W is an inductive inference operator that exhibits properties that extend rational closure, satisfy syntax splitting and avoid the drowning problem phenomenon [6, 9]. Komo and Beierle presented System W, which can capture and extend both c-inference and system Z inference [9]. It is noted that lexicographic inference, modified to be an inductive inference operator, can capture System W^+ , an modified version of System W retaining its core inference properties [6]. System W is defined using both the inclusion-maximal ordered partition, $OP(R)$, and the information on the set of falsified conditionals [9]. System W tracks the conditionals at each rank and falsifies them by a strict subset of a strict partial order $<_R^W$. The preference relation of a preferred world ω is preferred if it falsifies fewer conditionals than ω' and ω' falsifies at least the same conditionals as ω .

The Drowning Problem refers to inferences that can be made about non-exceptional subclasses and exceptional subclasses. An inference suffers from the drowning problem if it is not allowed to infer properties of a superclass for a subclass that is exceptional with respect to another property because the respective conditional is "drowned" by others. E.g. penguins are exceptional birds with respect to flying but not with respect to having wings. Thus we expect penguins to have wings. The issue arises in System Z (Rational Closure) inference because it cannot make inferences about exceptional conditionals because they have higher ranks and thus any inference about the lower subclasses cannot be made. Since System W ranks sets of conditionals it can distinguish between conditionals because it needs to falsify the same number of conditionals. The algorithm for system W inference was implemented by Beierle et al. [2] as Partial MaxSAT problems. The model for system W follows that given an input of the belief base Δ and sets of formulas A and B , the output would be the truth value of if $A \sim_{\Delta}^W B$ holds.

2.4 Lexicographic Closure

Lexicographic Closure [12] is a strengthening of Rational Closure that addresses its over-blocking limitation. Rather than discarding entire ranks when a subclass is identified as exceptional, Lexicographic Closure preserves as much knowledge as possible and discards only what is strictly necessary to restore consistency. The guiding principle is that while Rational Closure is maximally cautious about drawing conclusions, Lexicographic Closure is maximally cautious about discarding knowledge [3].

This is achieved through a seriousness ordering over subsets of the knowledge base, $\mathcal{K}$. When exceptional cases force some conditionals to be ignored, Lexicographic Closure selects the least costly subset to discard, preferring to sacrifice general knowledge over specific knowledge, and fewer conditionals over more. Given two subsets A and B of $\mathcal{K}$, A is considered more serious than B if A contains conditionals of higher base rank, or if both share the same highest rank but A contains more of them [8].

This ordering is formalised by associating each subset $A \subseteq \mathcal{K}$ with a tuple

$$v(A) = \langle |\{\alpha \sim \beta \in A : br(\alpha) = \infty\}|, \dots, |\{\alpha \sim \beta \in A : br(\alpha) = 0\}| \rangle$$

where each entry $|\{\alpha \sim \beta \in A : br(\alpha) = k\}|$ counts, from left to right, the number of conditionals in A whose antecedent α has base rank, $(br) k$. The tuple lists these counts from the highest rank ∞ (corresponding to classical knowledge) down to 0 (the most general defeasible conditional). Intuitively, $v(A)$ is a profile of how much high-ranked, and therefore more important, knowledge the set A contains. The seriousness ordering is then the natural lexicographic ordering on these tuples:

$$A <_S B \iff v(A) <_{lex} v(B)$$

where $<_{lex}$ compares tuples position by position from left to right, so A is less serious than B if A contains fewer high-ranked conditionals.

This ordering is a modular partial order and induces a ranked interpretation $\mathcal{R}_{LC}^{\mathcal{K}}$ [8]. A ranked interpretation assign each possible world a rank, with lower ranked worlds considered to be more normal. The lexicographic ranked function of a formula α with respect to a knowledge base $\mathcal{K}$ is then defined as:

$$r_{\mathcal{K}}^{LC}(\alpha) := \min\{\mathcal{R}_{LC}^{\mathcal{K}}(v) : v \in \hat{\alpha}\}$$

where $\hat{\alpha}$ denotes the set of all worlds satisfying α , so $r_{\mathcal{K}}^{LC}(\alpha)$ is the rank of the most normal in which α holds.

Lexicographic Closure strictly subsumes Rational Closure, recovering all conclusions that Rational Closure draws while additionally allowing exceptional subclasses to inherit unrelated properties from their superclass. It is a well-motivated entailment relation that sits naturally between Rational Closure and the stronger methods described in the literature [6].

2.5 Symbolic Rule Based Classifiers

Symbolic rule-based classifiers represent decisions as explicit if-then rules over discrete features, making every prediction traceable by design. Angelino et al. [1] demonstrate with CORELS that certifiably optimal rule lists can match the predictive accuracy of black-box models while remaining fully transparent. However, rule lists

handle exceptions through ordering (a more specific exception must appear before the default it overrides) which means the structural relationship between a default and its exception is implicit rather than visible. Madzime et al. [15] address this with DHCE, which groups defaults and exceptions into naturally readable normally-unless statements. DHCE produces human-readable explanation traces but lacks formal grounding in a non-monotonic framework, meaning its explanations cannot be formally verified. ECCRS [14] closes this gap by embedding the same design goals within the KLM framework, encoding the default-exception relationship directly in the rule bodies and resolving conflicts through the Most Specific Wins operator, making predictions both transparent and formally verifiable.

2.6 Translation Tools and Traces

Darwiche and Hirth [5] establish that the input-output behaviour of a classifier can be encoded as a propositional formula, enabling formal reasoning about decisions through complete and sufficient reasons. This provides the theoretical basis for translating ECCRS rulesets into KLM conditional knowledge bases, where each rule becomes a defeasible conditional and classification queries are answered through entailment. Chama [4] demonstrates that defeasible conclusions can be presented through natural language explanation templates, confirming that formal and human-readable explanations are complementary. This project builds on both contributions, producing explanation traces on two sides simultaneously.

2.7 Gap in the Literature

Existing symbolic classifiers either lack formal grounding or fail to capture the default-exception structure of human reasoning. ECCRS addresses both limitations, and an alignment theorem [14] guarantees that its MSW predictions agree exactly with Rational Closure under three structural conditions. However, no prior work has verified whether this agreement extends to Lexicographic Closure and System W, which capture default-exception reasoning with greater permissiveness than Rational Closure. Furthermore, no existing work produces a two-sided explanation trace that connects the human-readable justifications of a symbolic classifier to the formal semantics of defeasible entailment. This project addresses both gaps.

3 Problem Statement

Symbolic classifiers demonstrate that predictors with traceable justification are achievable, but traceability alone does not guarantee that the reasoning behind a prediction reflects how humans think about decisions. ECCRS addresses this by encoding the default-exception structure of human reasoning directly in its rule bodies, producing explanation traces that show which defaults applied and which exceptions overrode them. Its predictions are determined by the Most Specific Wins operator, which selects the most specific applicable rules and uses their common label as the classification.

The alignment theorem [14] guarantees that MSW predictions agree exactly with Rational Closure entailment on the translated KLM knowledge base, under three structural conditions: strict global exception closure, sanity, and no total override. This establishes a formal correspondence between the symbolic classifier

and defeasible reasoning for the base case. What remains unverified is whether this correspondence holds under Lexicographic Closure and System W, which are strictly more permissive than Rational Closure and capable of drawing conclusions that Rational Closure cannot. Since these methods handle the inheritance of generic properties by exceptional subclasses more faithfully than Rational Closure, any divergence between their outputs and MSW predictions would reveal a limitation of one side. This project builds the verification workbench needed to answer these questions, and uses it to produce two-sided explanation traces that make the relationship between ECCRS predictions and defeasible entailment legible.

3.1 Research Questions

- Under the conditions required for the alignment theorem, do Lexicographic Closure and System W entailments agree with MSW predictions on the translated KLM knowledge base, or do the additional inferences permitted by these methods reveal limitations of the MSW operator?
- Do the structural conditions imposed by strict global exception closure already render Lexicographic Closure and System W redundant on any well-formed ECCRS ruleset, or do divergences arise on rulesets that satisfy the conditions but expose edge cases in the MSW operator?
- Are the two-sided explanation traces produced by the workbench, one grounded in the ECCRS default-exception narrative and one in formal defeasible entailment, consistent, and do they together provide stronger justification for ECCRS predictions than either side alone?

3.2 Aims

- Translate the given ECCRS ruleset into a KLM conditional knowledge base.
- Implement Rational Closure entailment on the translated knowledge base, and verify its agreement with MSW predictions.
- Extend the workbench to Lexicographic Closure and System W, and use all three entailment methods to evaluate the scope and limits of the alignment theorem beyond the Rational Closure case.
- Produce two-sided explanation traces for each classification query, reporting the ECCRS default-exception chain and the formal defeasible entailment side by side, and evaluate whether the traces are consistent across all four methods.

4 Methodology

4.1 Overview

The initial stages of this project will be separated into components that relate to translation, implementation of algorithms, and the corresponding explanations. We will develop tools to validate assumptions according to strict global closure, sanity, and no total override [14]. Once we have validated all assumptions, we can translate the ECCRS rulesets into KLM style knowledge bases. The next step of the design will be to develop a naive version of Rational Closure that will form the foundation of the project. We will verify and compare the Rational Closure entailment algorithm presented

by Kraus et al. [11] with the ECCRS MSW predictions [15]. Subsequently, we expect to develop extensions of Rational Closure using Lexicographic [3] and System W algorithms [2]. Furthermore, we aim to verify the relevant assumptions for these entailments. The final feature set will produce traces for all four of the entailed labels.

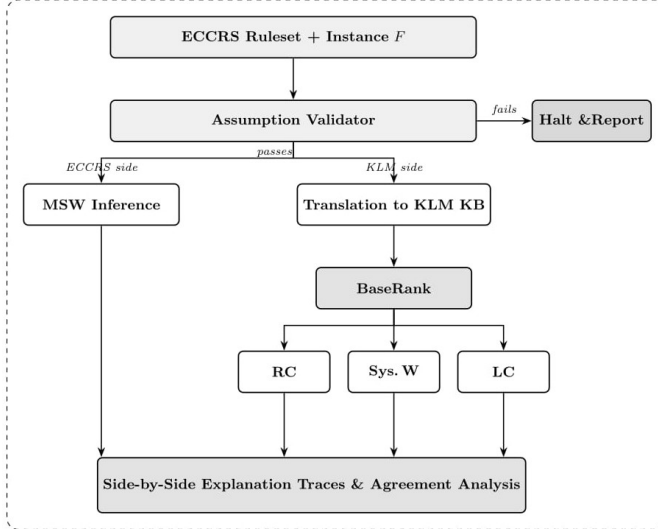


Figure 1: Representation of System Design

4.2 Approach

Before we can extend Rational Closure, we need to familiarise ourselves with the more permissive entailment algorithms given in the literature [3, 9]. We may need to read and understand any other relevant topics that will aid in their implementations. To assess the correctness of our approach, we will need to engage in discussions with our supervisors, presenting our work for evaluation. If any issues arise, we will need to re-evaluate our approach and present revised editions.

4.3 Software Environment

The workbench will be implemented in python. All data structures and algorithms are implemented from scratch. The ruleset is provided as plain text file in a defined format and parsed at runtime.

4.4 Hardware Environment

No specialized hardware is required.

4.5 Implementation Phases

The workbenches will consist of four phases:

Phase 1: Assumption Validation

The ruleset is parsed and checked against the three assumptions required for the alignment theorem to hold: Strict global closure, no total override, and Sanity. If any of these assumptions are violated, the workbench should halt and report.

Phase 2: Translation

The validated rule sets are translated into a KLM defeasible knowledge base by mapping each rule to a defeasible conditional.

Phase 3 : Algorithms Implementation

Four algorithms are implemented. ECCRS Most Specific Wins is implemented directly on the original ruleset. A naive Rational Closure is implemented as a baseline defeasible entailment algorithm and is a group contribution shared across all three defeasible entailment systems. System W and Lexicographical Closure are then built on top of RC, each refining the querying procedure in the way described [8]. An optimized version of Rational Closure

Phase 4: Explanations Traces and Output

Each workbench will produce not only a prediction but a structured explanation traces that records the reasoning steps leading to a conclusion. The explanations are generated on both sides, ECCRS Side and the translated KLM KB, using one of the three entailments, making the correspondence between the ECCRS reasoning and KLM reasoning directly visible. ECCRS trace, for a given input pattern, first, list all applicable rules for F. Second, mark rules which are strictly more specific than others. Third, the inclusion maximal applicable rules are identified. Fourth, the common label among the inclusion-maximal applicable rules is output as the prediction, or the classifier abstains if no rule applies. KLM traces, for same input F in the translated knowledge base, first the two candidate queries are stated. Second, the workbench reports which query is entailed by the entailment relation being used, RC, System W or LC or neither if the instance is uncovered. Third, we record the comparison driving the result. Fourth, the winning label is related back to the most specific ECCRS rule whose translation produced the surviving conditional.

The algorithms that we will use are specified by Kaliski [8].

Agreement analysis.: Once traces are generated for all complete feature patterns covered by the ruleset, a comparative analysis is conducted. For each pattern, the predictions of all four algorithms namely ECCRS MSW, RC, System W, and LC, are recorded. This recording will help us further analyze what percentage of the time do all four algorithms reach the same conclusions

4.6 Testing Methodology

4.6.1 Testing Data. Testing will use the ECCRS rule sets exported from an already learned classifier. The ruleset from the Ionosphere, Breast Cancer Wisconsin, and German Credit datasets, the same benchmarks used by Madzime et al [15] to evaluate DHCE will be used as test cases. These datasets provide rulesets of varying size and complexity, allowing the workbench to be evaluated across different scales.

5 Ethical , Professional and Legal Issues

This project is primarily theoretical and does not involve human participants, personal data, or sensitive information. The ECCRS ruleset used for verification is derived from a publicly available medical dataset that has been widely used as a benchmark in symbolic classifier research [7]. No private or proprietary data is used at any stage of the project.

The broader context of this work carries ethical significance. Explainability in high-stakes AI decision-making is a matter of public concern. By demonstrating that ECCRS predictions are formally certifiable through standard defeasible reasoners, this project

contributes to the trustworthiness and accountability of symbolic classifiers in these settings.

6 Anticipated Outcomes

The central goal of this project is to analyze and critically evaluate ECCRS and Rational Closure with its extensions (Lexicographic and System W). The idea is to implement already established defeasible algorithms for entailment and compare them to ECCRS MSW predictions. The anticipated outcomes for this project include:

- Three Python/C workbenches that verify assumptions for ECCRS to KLM Knowledge base. It will use the relevant alignment theorems, translates to KLM KBs, and each will run the relevant entailment algorithms.
- Produce explanation traces that show a side-by-side trace demonstrating the default-exceptions rule override narrative from both ECCRS and KLM perspectives.
- Provide a quantitative report that evaluates what proportion of the predictions (RC, LC and SW) agree with ECCRS MSW predictions?

7 Project Plan

7.1 Risks

Inadequate Understanding of the theoretical material This project is positioned at an intersection of symbolic machine learning and defeasible reasoning, which are technically demanding areas. Misunderstanding the alignment theorem assumptions or the computational differences between the three entailments relations could lead to incorrect implementation. To mitigate this, we will use Kaliski et al [8] and the supervisor’s notes on ECCRS as theoretical and technical prerequisites. Progress is monitored by testing each algorithm against the worked examples in literature before integrating into the workbench. Any conceptual uncertainty is raised with the supervisor.

Team member availability Different schedules among team members may affect the pace of shared components, especially the planned naive Rational closure implementation, which is a group contribution. To mitigate this, the RC implementation is prioritised early in the project timeline so that the delays do not cascade into the System W and LC implementations that depend on it.

See Table 2

7.2 Timeline

See Figure 2

7.3 Resources

The resources required for this project involve a computing device with a code-editor and the relevant libraries for the implementation of the various algorithms in either C or Python.

7.4 Deliverables

Deliverables	Due Date
Literature Review	18 March
Project Proposal	24 April
Project Proposal Presentation	28-30 April
(Revised Proposal: First Reader)	5 May
(Revised Proposal: Second Reader)	5 May
Project Draft Demonstration	27 July
Project Paper Scaffold	10 August
Final Paper Draft	24 August
Final Paper Submission	31 August
Project Code Final Submission	14 September
Project Demonstration	16-23 September
Poster	18 September
Website	5 October

7.5 Milestones and Tasks

See Figure 1

7.6 Work Allocation

Student	Work Allocation
Daniel, August, Antony	Running ECCRS and assumption validator, the translation from ECCRS to KLM, the side-by-side explanation trace output, and the comparative analysis and write-up.
Daniel	Daniel is responsible for optimising the RC implementation beyond the naive baseline, improving its efficiency and ensuring it handles edge cases correctly across all test datasets.
Antony	Antony is responsible for implementing the System W entailment algorithm, building on the shared RC foundation and extending the query procedure and producing explanation traces
August	Augustine is responsible for implementing the Lexicographic Closure entailment algorithm, including the weakening procedure that distinguishes LC from RC and produces the explanation traces.

8 AI Usage Statement

Generative AI was used in the writing to improve the clarity and overall coherence of the writing. The initial drafts were written manually and specific grammar and word choices were further improved through the use of AI. The use of Overleaf AI writing checker and the Typst AI grammar checker.

Tasks	START	END
Literature Review	01/03	18/03
Project Proposal	13/04	05/05
Draft	13/04	24/04
Presentation	24/04	30/04
Final	01/05	05/05
Preliminary System Design	06/05	13/05
Outline of System Design	06/05	11/05
Review with Supervisor	12/05	13/05
Development of Naive Rational Closure	14/05	24/05
Draft of implementation	14/05	18/05
Finalize algorithm implementations	19/05	24/05
Development of Defeasible Algorithms	25/05	20/06
Draft of implementation	25/05	06/06
Feedback and Revision	07/06	08/06
Finalize algorithm implementations	08/06	20/06
Explanation Traces and Agreement Analysis	21/06	04/08
Draft of implementation	21/06	20/07
Feedback and refinement of design	20/07	21/07
Implementation of design	21/07	04/08
Project Draft Demonstration	24/07	27/07
Final Paper	15/08	31/08
Project Code Final Submission	01/09	14/09
Project Demonstration	16/09	23/09
Project Poster	16/09	18/09
Project Website	16/09	05/10

Table 1: Milestones and Tasks

Risk Scenario	Probability	Impact	Mitigation	Monitoring	Management
Supervisor Unavailable	Unlikely	High	Frequent checks on supervisor's schedule and communication about availability	Email and communication with supervisor regularly	Stockpile questions to be asked in meetings
Team member drops out	Possible	Moderate	Ensuring each project is a standalone project	Communicate with team members with regards to milestones	Standalone projects ensure that team members are able to complete project
Poor team communication	Possible	High	Group chat as well as regular check-ins with each other.	Active communication and that everyone engages with check-ins	Can be managed by meetings with supervisor to resolve any issues
Difficulty understanding theoretical background	Likely	Moderate	Review relevant literature and try to grasp fundamental concepts	Regular meetings where we explain components we are working on and illustrate any difficulties in understanding underlying concepts.	Consult supervisor to assist in understanding.
Team member fails to complete initial RC implementation	Unlikely	Moderate	break the implementation into small subtask with clear deadlines	Review progress continuously and flag delays	Assist the struggling member
Feature Creep	Possible	Moderate	Have clearly defined project scope and list of features	Compare what we have implemented with project scope and review alignment	Any features that are non-essential to the final workbench are to be labelled clearly in the future-work section

Table 2: Risk Identification, Mitigation and Monitoring

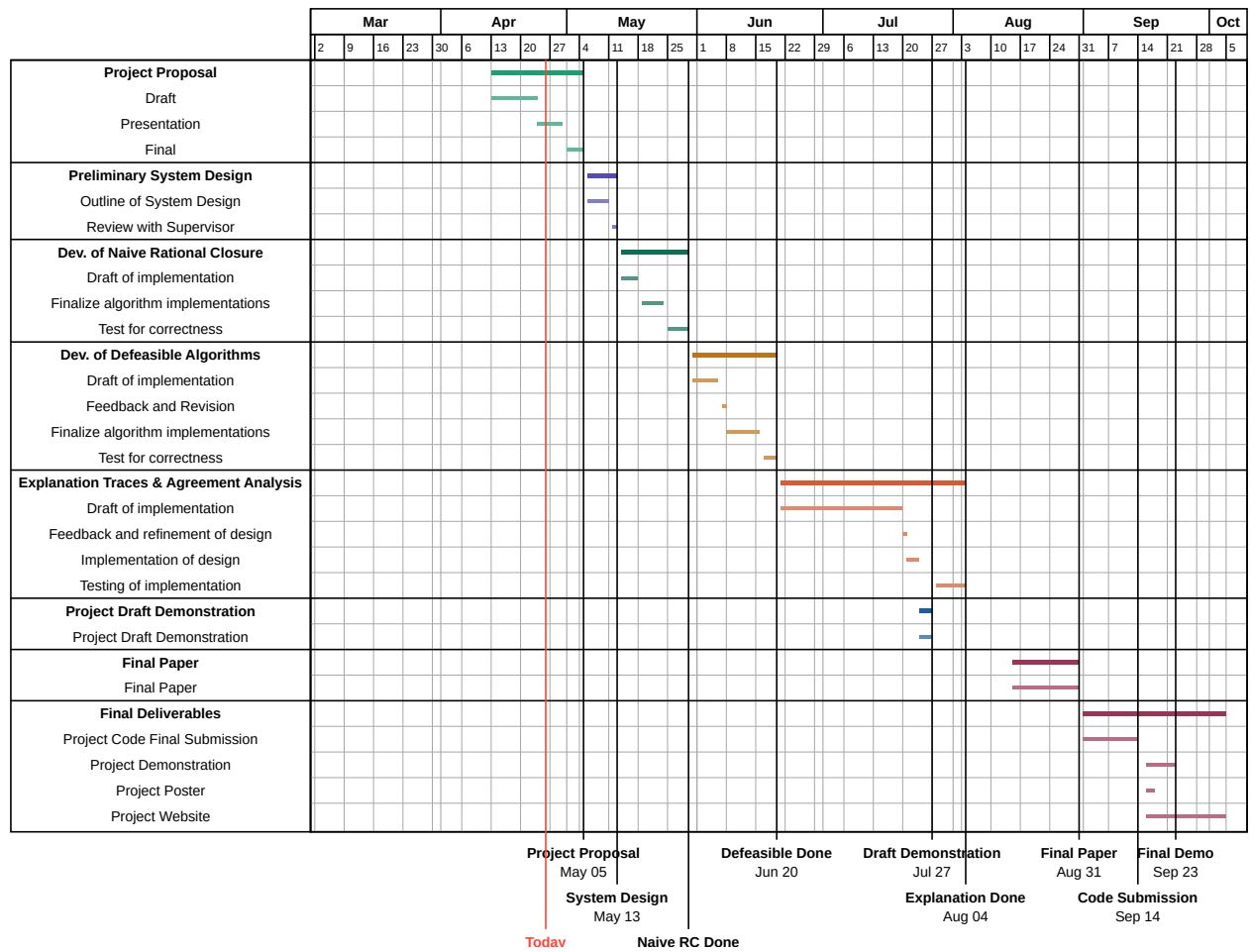


Figure 2: Gantt Chart

References

- [1] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. 2018. Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research* 18, 234 (2018), 1–78.
- [2] Christoph Beierle, Aron Spang, and Jonas Haldimann. 2024. A Partial MaxSAT Approach to Nonmonotonic Reasoning with System W. *The International FLAIRS Conference Proceedings* 37, 1 (May 2024). doi:10.32473/flairs.37.1.135330
- [3] Giovanni Casini, Thomas Meyer, and Ivan Varzinczak. 2019. *Taking Defeasible Entailment Beyond Rational Closure*. 182–197. doi:10.1007/978-3-030-19570-0_12
- [4] Victoria Chama. 2020. Explanation for defeasible entailment. (2020). Master’s Thesis, University of Cape Town.
- [5] Adnan Darwiche and Auguste Hirth. 2023. On the (complete) reasons behind decisions. *Journal of Logic, Language and Information* 32, 1 (2023), 63–88.
- [6] Jonas Haldimann, Christoph Beierle, Gabriele Kern-Isberner, and Thomas Meyer. 2023. Conditionals, Infeasible Worlds, and Reasoning with System W. *The International FLAIRS Conference Proceedings* 36, 1 (May 2023). doi:10.32473/flairs.36.133268
- [7] Reijo Jaakkola, Tomi Janhunen, Antti Kuusisto, Masood Fezbakhsh Rankooh, and Miikka Vilander. 2023. Short Boolean Formulas as Explanations in Practice. In *Logics in Artificial Intelligence*, Sarah Gaggl, Maria Vanina Martinez, and Magdalena Ortiz (Eds.), Springer Nature Switzerland, Cham, 90–105. doi:10.1007/978-3-031-43619-2_7
- [8] Adam Kaliski. 2020. An overview of KLM-style defeasible entailment. <http://hdl.handle.net/11427/32743>
- [9] Christian Komo and Christoph Beierle. 2022. Nonmonotonic reasoning from conditional knowledge bases with system W. *Annals of Mathematics and Artificial Intelligence* 90, 1 (2022), 1–60. doi:10.1007/s10472-021-09777-9
- [10] Sarit Kraus, Daniel Lehmann, and Menachem Magidor. 1990. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence* 44, 1 (1990), 167–207. doi:10.1016/0004-3702(90)90101-5
- [11] Sarit Kraus, Daniel Lehmann, and Menachem Magidor. 2002. Nonmonotonic Reasoning, Preferential Models and Cumulative Logics. *arXiv:cs/0202021 [cs.AI]* <https://arxiv.org/abs/cs/0202021>
- [12] Daniel Lehmann. 1995. Another perspective on default reasoning. *Annals of Mathematics and Artificial Intelligence* 15 (1995). doi:10.1007/BF01535841
- [13] Daniel Lehmann and Menachem Magidor. 1992. What does a conditional knowledge base entail? *Artificial Intelligence* 55, 1 (1992), 1–60. doi:10.1016/0004-3702(92)90041-U
- [14] Ruvarashe Madzime. 2026. ECCRS in a Nutshell for the Honours Project: Semantics, Most-Specific-Wins Operator, and the Rational Closure Alignment. (2026). Unpublished technical note, University of Cape Town.
- [15] Ruvarashe Shalom Madzime, Thomas Meyer, and Louise Leenen. 2025. An Override-Aware Classifier for Transparent AI. In *Proceedings of the Southern African Conference for Artificial Intelligence Research* (Cape Town, South Africa), Aurore Gerber and Anban Pillay (Eds.).
- [16] Fernando E. B. Otero and Alex A. Freitas. 2013. Improving the Interpretability of Classification Rules Discovered by an Ant Colony Algorithm. In *Proceedings of the Fifteenth Annual Conference on Genetic and Evolutionary Computation (GECCO '13)*. ACM, 73–80.
- [17] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *arXiv:1811.10154 [stat.ML]* <https://arxiv.org/abs/1811.10154>