Explanations for KLM-style Defeasible Entailment

Lloyd Everett
Department of Computer Science
University of Cape Town
South Africa

ABSTRACT

Explanations are a crucial aspect of reasoning systems but they have not yet been explored in detail for nonmonotonic formalisms such as KLM. We give an overview of KLM-style formalisms and review the current literature on explanation for KLM, finding that prior work points to some possible approaches to advance our understanding of defeasible explanation for rational formalisms and related formalisms such as Relevant Closure. Our survey indicates that there is quite good evidence that algorithmic approaches to justification for Rational Closure can be adapted to Lexicographic Closure and Relevant Closure and on the other hand that it may be possible to characterise justifications for KLM-style defeasible knowledge bases generally in a declarative manner.

CCS CONCEPTS

• **Theory of computation** → *Automated reasoning*; • **Computing methodologies** → **Nonmonotonic, default reasoning and belief revision**; *Causal reasoning and diagnostics.*

KEYWORDS

knowledge representation and reasoning, explainable artificial intelligence, defeasible reasoning, Rational Closure, Relevant Closure

1 INTRODUCTION AND MOTIVATION

Knowledge representation and reasoning using logics is characterised by two key goals: we seek to make it possible to meaningfully encode information about a domain into a knowledge base, and we seek to support reasoning services that enable us to make deductions from the knowledge base [12]. There is sometimes a trade-off between these two goals. Logics that are very expressive facillitate knowledge representation but reasoning services for such logics can be computationally or mathematically intractable. On the other hand, logics that are inexpressive are computationally easier to handle but it can be difficult to represent domain knowledge in these logics. As such, we tend to try to find logics that are just as expressive as they need to be for a given domain, but no more [1].

There are several reasoning services we may desire from a reasoning system. Foremost of these is the service of checking *entailment* [1, 14]. Intuitively, this corresponds to the notion of testing whether a statement can be inferred from a knowledge base. As an example, if we have the knowledge that "Tweety is a bird" and "birds fly", then we are able to verify via the reasoning service that the knowledge *entails* that "Tweety flies". Another basic reaosning service is *satisfiability* which tells us whether a knowledge base or statement is possibly true (or in other words not necessarily false) [1, 3].

A fourth reasoning service, and the focus of this review, is that of *explanation*. Explanation services tell us which statements in a knowledge base are relevant to the entailment between a knowledge base and an entailed statement [8]. Explanations are useful in reasoning systems because they allow the user of a reasoning system to understand which parts of the knowledge base lead to a particular conclusion. This is helpful particularly when the reasoner is giving unexpected entailment results since it allows the user to identify the culprit knowledge base statements and thus explanations can provide a way to debug knowledge bases [13]. Explanations are helpful even when the reasoner is behaving as expected because they can improve knowledge base comprehension, particularly if the user is not familiar with the knowledge base [2, 13]. Explanations have also been shown to improve users' confidence in reasoning systems [4].

One of the simplest and less expressive logics used for knowledge representation and reasoning is *classical propositional logic*, which we will refer to simply as propositional logic. Propositional logic is a well-understood logic that can be seen as the basis for more expressive logics such as modal logics or the popular description logics. Because of its simplicity, and because it serves as a foundational logic, we have sophisticated reasoning services for propositional logic; strictly speaking, the reasoning algorithms here are NP-complete [9], but in practice modern SAT solvers are able to solve propositional logic reasoning problems with very good practical efficiency for real-world knowledge bases [19]. Unfortunately, the simplicity of the logic means that it can be difficult to practically express domain knowledge in the logic.

One of the ways in which propositional logic lacks expressiveness is its inability to describe *typicality*, i.e., it is very difficult to express statements that typically hold, but for which there might be exceptions. The knowledge base we had earlier serves as a good example. In propositional logic, if we have that "Tweety is a bird" and "birds fly" then we can derive entailments such as "Tweety flies" but it becomes very difficult to then represent additional exceptional knowledge such as the statements "Rico is a penguin", "penguins are birds" and "penguins do not fly". What we really want to do is express that "birds *typically* fly" without specifying each and every possible exception upfront. Logics that are able to express typicality in this manner are said to be *nonmonotonic*, and the study of such logics and their associated reasoning services is *defeasible reasoning* [14].

This brings us to the primary focus of this review. Explanation services are relatively well-understood in the classical case [16] but have not yet been explored in detail for defeasible reasoning apart from some foundational work [5, 8]. Although there are many approaches to defeasible reasoning, one approach that has been studied extensively in the literature is that proposed by Kraus, Lehmann & Magidor (KLM) [15]. One of the central contributions [18] of KLM is an axiomatic description of *rational* defeasible entailment

relations; thus, KLM does not define a single notion of defeasible entailment but rather defines a class of defeasible entailment relations that have some interesting theoretical and computational properties [14, 18]. Unlike the classical case, it is generally understood that it is desirable to have a number of formalisms for defeasible entailment that correspond to different reasoning styles [14]. Multiple rational [7] formalisms for defeasible entailment have been described in the literature including *Rational Closure* [18] and *Lexicographic Closure* [17]. Rational Closure corresponds to a more conservative form of reasoning compared to Lexicographic Closure which is much more permissive. Another formalism proposed by Casini, et al. [6] that is not quite rational but is still closely related to KLM is *Relevant Closure*, which lies between Rational and Lexicographic Closure in terms of permissiveness.

In this paper we will look at the literature on KLM defeasible reasoning and explanations with a focus on laying the groundwork for further study that describes an explanation service for KLM-style formalisms of entailment. Such study would make these formalisms more useful as knowledge representation and reasoning systems for the reasons we discussed earlier. Though there are KLM formalisms for both propositional logic [15, 17, 18] and description logic [7], we will focus on the propositional formalism because it is simpler. An algorithm for explanation for Rational Closure has already been described in the literature [8]; we will mainly be looking at Relevant Closure in this paper[1]. We also ultimately wish to determine whether there are some declarative properties characteristic of many reasonable formalisms for defeasible explanation for KLM. This could give evidence that the algorithm given by Chama [8] and any algorithms we propose produce a sensible definition for defeasible justification and could also draw our attention to other forms of defeasible justification distinct from that proposed by Chama.

2 PRESENTATION

2.1 Classical Propositional Logic

2.1.1 Formalism. Classical propositional logic has simple semantics and is the foundation for more complex logics such as KLM propositional logic, so it will be helpful to look at the logic as well as its notion of explanations. The following is a description of classical propositional logic [3]. We begin with a finite set $\mathcal{P} = \{p, q, \cdots\}$ of *propositional atoms*. The binary connectives $\wedge, \vee, \rightarrow, \leftrightarrow$ and the unary negation operator $\neg$ are used recursively to form propositional formulas such as $\neg(p \vee q) \rightarrow p$. The set of all such formulas over $\mathcal{P}$ is called the *propositional language* $\mathcal{L}$.

An *interpretation* or *valuation* is a function $\mathcal{P} \rightarrow \{T, F\}$ that assigns a truth value to each atom in $\mathcal{P}$. We say a formula $A \in \mathcal{L}$ is *satisfied* by an interpretation $\mathcal{I}$, written as $\mathcal{I} \models A$, iff A evaluates to true according to the truth values of the atoms in A and the semantics of operators in A which are defined according to Tarskian semantics. For example, if $\mathcal{I}(p) = T$ and $\mathcal{I}(q) = F$, then $\mathcal{I} \models p \vee q$ but $\mathcal{I} \not\models p \wedge q$. The interpretations that satisfy a formula A are referred to as *models* of A, and the set of models of A is denoted $\text{Mod}(A)$. Finally, we assert that $\top$ is a formula satisfied by every interpretation and that $\bot$ is a formula not satisfied by any interpretation.

[1] My research partner, Emily Morris, is studying Lexicographic Closure.

Table 1: Examples of classical justifications for the example knowledge base $\mathcal{K}$

A	$\mathcal{J}^{\mathcal{K}}(A)$
$t \rightarrow b$	$\{\{t \rightarrow b\}\}$
$t \rightarrow w$	$\{\{t \rightarrow b, b \rightarrow w\}\}$
$(b \wedge f) \rightarrow w$	$\{\{b \rightarrow w\}, \{f \rightarrow w\}\}$

Note that there is an important distinction between statements such as $p \wedge q$ and $\mathcal{I} \models p \wedge q$. Namely, the former is a statement within the propositional language while the latter is a statement in a metalanguage over the propositional language [14]. These concepts are often referred to as the *object level* and the *meta level*. In essence, the object level allows us to make statements about propositional atoms while the meta level allows us to make statements about propositional logic statements. Entities such as $\mathcal{I}$ or $\mathcal{P}$ and relations such as $\models$ therefore belong to the meta level and not the object level.

We now describe some additional semantics at the meta level which will give us the basis for a knowledge representation and reasoning system for propositional logic. A finite set of propositional formulas is called a *knowledge base* $\mathcal{K}$. The models of a knowledge base $\text{Mod}(K)$ are simply $\bigcap \{\text{Mod}(A) \mid A \in \mathcal{K}\}$. For a knowledge base or statement A, we define entailment, satisfiability and validity as follows [1, 3]. We say that A entails a statement B, denoted $A \models B$, iff $\text{Mod}(A) \subseteq \text{Mod}(B)$. A is satisfiable iff $\text{Mod}(A) \neq \emptyset$ and is valid iff $\text{Mod}(A) = \mathcal{W}$ where $\mathcal{W}$ is the set of all interpretations for $\mathcal{P}$.

As a short example, consider the propositional language $\mathcal{L}$ of the atoms $\mathcal{P} = \{b, f\}$. Think of b as meaning that "Tweety is a bird" and f as meaning that "Tweety flies". Now if $\mathcal{K} = \{b \rightarrow f, b\}$ then we have $\mathcal{K} \models f$ since $\text{Mod}(\mathcal{K}) \subseteq \text{Mod}(f)$. We therefore have an example of modus ponens; if Tweety being a bird implies that Tweety flies, and Tweety is a bird, then Tweety flies.

2.1.2 Explanations for Classical Propositional Logic. Perhaps the most common and basic form of explanation for classical logics is what we will refer to as a *justification* [20]. We say that $\mathcal{J}$ is a justification for an entailment $\mathcal{K} \models A$ iff $\mathcal{J}$ is a *minimal* subset $\mathcal{J} \subseteq \mathcal{K}$ such that $\mathcal{J} \models A$ [13]. The meaning of minimal here is that if we remove any statements from $\mathcal{J}$ we should no longer have $\mathcal{J} \models A$; in other words, there is no $\mathcal{J}' \subset \mathcal{J}$ such that $\mathcal{J}' \models A$. Justifications are not necessarily unique since we many have more than one minimal set that entails A [13]. We denote the set of justifications for $\mathcal{K} \models A$ as $\mathcal{J}^{\mathcal{K}}(A)$. Table 1 gives some examples of classical justifications for the example knowledge base $\mathcal{K} = \{t \rightarrow b, b \rightarrow f, f \rightarrow w, b \rightarrow w\}$.

Although justifications can give an indication of which statements are involved in an entailment, they can be a somewhat brittle tool from the user's perspective [20]. One of the problems here is that an entailment may have very many justifications. More sophisticated approaches to explanation have been described that try to alleviate these problems [20], but our focus will be on justifications since they are a simple form of explanation and ultimately we are looking to find the analogue of classical justifications for KLM-style defeasible reasoning.

A brief word on how we obtain these justifications algorithmically. A useful result here is the reduction from checking entailment to checking satisfiability. Specifically, a knowledge base $\mathcal{K}$ entails A iff $\mathcal{K} \cup \{\neg A\}$ is unsatisfiable [1], which means that an algorithm for checking knowledge base satisfiability can also be used to check for entailment. Modern SAT solvers are very efficient at solving these problems [19] and algorithms have been described that can enumerate justifications efficiently for classical entailments [13].

2.2 KLM Defeasible Propositional Logic

2.2.1 Formalism. When Kraus, Lehmann and Magidor initally set out KLM they described defeasible implication as a consequence relation $\vdash$ at the meta level [18]. We will instead describe the approach that is now more commonly used where the propositional logic is extended with an object level defeasible connective $\vdash$ which can be seen as the defeasible analogue of $\rightarrow$ [7, 14]. Statements of the form $p \vdash q$ are read as p *typically implies* q. Unlike $\rightarrow$, we require that when $\vdash$ occurs in a formula it is the outermost operation and certainly $\vdash$ cannot be nested. We denote this extended language $\mathcal{L}_{\mathcal{D}}$. We then define a notion of defeasible entailment $\mathrel{|\!\sim}$ at the meta level which is the defeasible analogue of $\models$. As an example, we can write $\mathcal{K} = \{b \vdash f, p \rightarrow b, p \vdash \neg f\}$ meaning "birds typically fly", "penguins are birds" and "penguins typically do not fly". Now a reasonable definition of $\mathrel{|\!\sim}$ will allow us to defeasibly conclude $\mathcal{K} \cup \{t \rightarrow b\} \mathrel{|\!\sim} t \vdash f$ ("if Tweety is a bird then Tweety flies") and $\mathcal{K} \cup \{r \rightarrow p\} \mathrel{|\!\sim} r \vdash \neg f$ ("if Rico is a penguin then Rico does not fly"). The idea here is that we want $\mathrel{|\!\sim}$ to favour the most specific rules in the knowledge base that are applicable [14]. In the case of Rico, $r \rightarrow p, p \vdash \neg f$ is more specific than $r \rightarrow p, p \rightarrow b, b \vdash f$, so the former rule is favoured.

Of course, we have not yet defined the meaning of $\mathrel{|\!\sim}$. As we said earlier, KLM defines a set of properties for an entailment relation to be *rational*. An entailment relation $\mathrel{|\!\sim}$ is rational iff it obeys the following postulates [7]:

(1) *Left logical equivalence (LLE). If* $\mathcal{K} \mathrel{|\!\sim} A \leftrightarrow B$ *and* $\mathcal{K} \mathrel{|\!\sim} A \vdash C$ *then* $\mathcal{K} \mathrel{|\!\sim} B \vdash C$.
(2) *Right weakening (RW). If* $\mathcal{K} \mathrel{|\!\sim} A \rightarrow B$ *and* $\mathcal{K} \mathrel{|\!\sim} C \vdash A$ *then* $\mathcal{K} \mathrel{|\!\sim} C \vdash B$.
(3) *Reflexivity (Ref).* $\mathcal{K} \mathrel{|\!\sim} A \vdash A$.
(4) *And. If* $\mathcal{K} \mathrel{|\!\sim} A \vdash B$ *and* $\mathcal{K} \mathrel{|\!\sim} A \vdash C$ *then* $\mathcal{K} \mathrel{|\!\sim} A \vdash B \wedge C$.
(5) *Or. If* $\mathcal{K} \mathrel{|\!\sim} A \vdash C$ *and* $\mathcal{K} \mathrel{|\!\sim} B \vdash C$ *then* $\mathcal{K} \mathrel{|\!\sim} A \vee B \vdash C$.
(6) *Cautious Monotonicity (CM). If* $\mathcal{K} \mathrel{|\!\sim} A \vdash C$ *and* $\mathcal{K} \mathrel{|\!\sim} A \vdash B$ *then* $\mathcal{K} \mathrel{|\!\sim} A \wedge B \vdash C$
(7) *Rational Monotonicity (RM). If* $\mathcal{K} \mathrel{|\!\sim} A \vdash C$ *and* $\mathcal{K} \not\mathrel{|\!\sim} A \vdash \neg B$ *then* $\mathcal{K} \mathrel{|\!\sim} A \wedge B \vdash C$.

These postulates are not sufficient to ensure that an entailment relation is sensible for defeasible reasoning [14], and in fact relations such as Relevant Closure may be useful even though they are not rational, so they do not necessarily describe a minimal set of characteristics for a useful $\mathrel{|\!\sim}$ either. Nevertheless, these axioms are useful as a starting point for describing what we expect from a defeasible entailment relation. Furthermore, these entailment relations have some compelling characteristics both from a theoretical standpoint and from an algorithmic standpoint [10, 14, 18]. A full exploration of these features is out of the scope of this paper so we will instead focus on the elements that are most relevant to our study of explanations for KLM defeasible reasoning.

In that vein, some of the most important results here show that rational entailment relations can be described from other angles that intuitively seem very different to the axiomatic description above. These include ranked interpretations, preferential interpretations and base ranks on propositional formulas [7, 14, 18]. The last of these three is the most useful from a computational perspective [14] and will be our main focus, both because current work on defeasible explanation for Rational Closure is described using it [8] and because it can also be used to describe entailment relations that are not rational such as Relevant Closure [6, 14] (which is not the case for ranked or preferential interpretations). We will define both Rational Closure [18] and Relevant Closure [6] in these terms.

2.2.2 Rational Closure. Rational closure is the form of defeasible reasoning that Lehmann and Magidor [18] proposed for KLM and is the most conservative (prototypical) of all the rational entailment relations [14]. In this section, we define the defeasible entailment relation $\mathrel{|\!\sim}_{\text{RC}}$ for Rational Closure using the concept of *base ranks*. First, we need to discuss some preliminary ideas. Earlier, we allowed statements in $\mathcal{L}_{\mathcal{D}}$ to be classical statements not of the form $A \vdash B$ (in other words, we had $\mathcal{L} \subset \mathcal{L}_{\mathcal{D}}$). For the sake of simplicity, we are now going to assume that all statements in $\mathcal{L}_{\mathcal{D}}$ are defeasible implications $A \vdash B$ where $A \in \mathcal{L}, B \in \mathcal{L}$. This does not restrict our ability to express certain or categorical information in the language because the entailment relations we are going to define will allow us to express any classsical statement C as $\neg C \vdash \bot$ in a defeasible knowledge base; this is true for all rational entailment relations [14] as well as Relevant Closure [6]. Now define the *materialisation* $\overline{\mathcal{K}}$ of a defeasible knowledge base $\mathcal{K}$ as $\{A \rightarrow B \mid A \vdash B \in \mathcal{K}\}$ [18]. Thus $\overline{\mathcal{K}}$ gives a classical knowledge base of classical implications for each statement in $\mathcal{K}$.

We say that a propositional formula $A \in \mathcal{L}$ is *exceptional* for $\mathcal{K}$ iff $\overline{\mathcal{K}} \models \neg A$ [18]. The intution behind this is that exceptional formulas are false in the most typical valuations for $\mathcal{K}$ but may be true for more specific sets of valuations. For our penguin example earlier, we would have $\overline{\mathcal{K}} = \{b \rightarrow f, p \rightarrow b, p \rightarrow \neg f\}$ with $\overline{\mathcal{K}} \models \neg p$ and thus p is exceptional for $\mathcal{K}$. We also define $\varepsilon(\mathcal{K})$ to give us the set of statements in $\mathcal{K}$ whose antecedents are exceptional for $\mathcal{K}$:

$$\varepsilon(\mathcal{K}) = \left\{A \vdash B \mid A \vdash B \in \mathcal{K} \text{ with } \overline{\mathcal{K}} \models \neg A\right\}.$$

We are now in a position to define a sequence of knowledge bases $\mathcal{E}_0^{\mathcal{K}}, \mathcal{E}_1^{\mathcal{K}}, \cdots, \mathcal{E}_n^{\mathcal{K}}$ [14] such that knowledge bases earlier in the sequence contain, in addition to the statements in later knowledge bases, statements that are more defeasible or retractable than those in later knowledge bases. Let $\mathcal{E}_0^{\mathcal{K}} = \mathcal{K}$ and $\mathcal{E}_{i+1}^{\mathcal{K}} = \varepsilon\left(\mathcal{E}_i^{\mathcal{K}}\right)$. The last knowledge base $\mathcal{E}_n^{\mathcal{K}}$ is the first $\mathcal{E}_i^{\mathcal{K}}$ where

$$\varepsilon\left(\mathcal{E}_i^{\mathcal{K}}\right) = \mathcal{E}_i^{\mathcal{K}}.$$

We will often refer to $\mathcal{E}_n^{\mathcal{K}}$ as $\mathcal{E}_\infty^{\mathcal{K}}$ as a convenience. The statements in $\mathcal{E}_\infty^{\mathcal{K}}$ are not retractable [7], and note that we may have $\mathcal{E}_\infty^{\mathcal{K}} = \emptyset$ in the event that $\mathcal{K}$ does not contain any non-retractable information [14]. We also define $\mathcal{R}_i^{\mathcal{K}} = \mathcal{E}_i^{\mathcal{K}} \setminus \mathcal{E}_{i+1}^{\mathcal{K}}$ for $0 \leq i \leq n-1$ and $\mathcal{R}_\infty^{\mathcal{K}} = \mathcal{E}_\infty^{\mathcal{K}}$ since sometimes it is more convenient to work with a

Figure 1: Ranking $\mathcal{R}_0^{\mathcal{K}}, \cdots, \mathcal{R}_\infty^{\mathcal{K}}$ for the example knowledge base $\mathcal{K}$

∞	$p \rightarrow b, r \rightarrow p$
1	$p \vdash \neg f$
0	$b \vdash f, b \vdash w$

Table 2: Examples of $\approx_{\text{RC}}$ entailment for the example knowledge base $\mathcal{K}$

$A \vdash B$	$br^{\mathcal{K}}(A)$	$br^{\mathcal{K}}(A \wedge \neg B)$	$\mathcal{K} \approx_{\text{RC}} A \vdash B$?
$b \vdash \neg f$	0	0	No
$r \vdash \neg f$	1	∞	Yes
$p \vdash w$	1	1	No

ranking of the statements in $\mathcal{K}$. Figure 1 gives this ranking for the example knowledge base $\mathcal{K} = \{b \vdash f, b \vdash w, p \rightarrow b, p \vdash \neg f, r \rightarrow p\}$. (Strictly, classical statements such as $p \rightarrow b$ should be written as $\neg(p \rightarrow b) \vdash \bot$ as discussed earlier; the simpler notation is used only as a shorthand.)

Define the *base rank* $br^{\mathcal{K}}(A)$ of a propositional formula $A \in \mathcal{L}$ as the minimum r such that A is not exceptional in $\mathcal{E}_r^{\mathcal{K}}$ [7]:

$$br^{\mathcal{K}}(A) = \min\left\{r \mid \overline{\mathcal{E}_r^{\mathcal{K}}} \not\models \neg A\right\}.$$

Now we can define the entailment relation $\approx_{\text{RC}}$ for Rational Closure:

$$\mathcal{K} \approx_{\text{RC}} A \vdash B \text{ iff } br^{\mathcal{K}}(A) < br^{\mathcal{K}}(A \wedge \neg B) \text{ or } br^{\mathcal{K}}(A) = \infty.$$

A result by Giordano, et al. [11] shows that that the relation $\approx_{\text{RC}}$ defined here is the same as that initially given by Lehmann and Magidor [18]. Table 2 gives some examples of $\approx_{\text{RC}}$ entailment for the knowledge base $\mathcal{K} = \{b \vdash f, b \vdash w, p \rightarrow b, p \vdash \neg f, r \rightarrow p\}$. These results can be obtained using the ranking in Figure 1. The intuition here is that we use the ranking to eliminate more typical or defeasible ranks that contradict more specific ranks with respect to the antecedent, and once these ranks have been eliminated, we rely on classical tools, namely classical entailment, to reason about the knowledge base.

This approach seems reasonable given our initial stated goal of always using the most specific information in the knowledge base when specific and typical information disagree. This does however result in a rather conservative form of reasoning since we always retract entire ranks of more typical statements even though only a handful of statements in a rank may disagree with statements in higher, more specific ranks [6]. We will see later that Relevant Closure tries to address this problem by only retracting the statements that are actually involved in the exceptionality of the antecedent [6].

One of the advantages of this representation of Rational Closure is that the definition admits an algorithm with time complexity equal to that of checking classical entailment. Expanding the definition above, we have $\mathcal{K} \approx_{\text{RC}} A \vdash B$ iff

$$\min\left\{r \mid \overline{\mathcal{E}_r^{\mathcal{K}}} \not\models \neg A\right\} < \min\left\{r \mid \overline{\mathcal{E}_r^{\mathcal{K}}} \not\models A \rightarrow B\right\},$$

which we can compute as long as we have a solver for classical entailment. These algorithms are summarised here as Algorithm 1 and 2; the former computes the sequence $\mathcal{R}_0^{\mathcal{K}}, \mathcal{R}_1^{\mathcal{K}}, \cdots, \mathcal{R}_n^{\mathcal{K}}$ for a knowledge base $\mathcal{K}$ and the latter computes whether it is true given $\mathcal{K}$ and $A \vdash B$ that $\mathcal{K} \approx_{\text{RC}} A \vdash B$. A result from Freund [10] proves that these algorithms are indeed equivalent to the definition of rational closure above.

Algorithm 1 Rank

Input: A knowledge base $\mathcal{K}$
Output: $\left(\mathcal{R}_0^{\mathcal{K}}, \cdots, \mathcal{R}_n^{\mathcal{K}}, n\right)$
$i \coloneqq 0; E_0 \coloneqq \mathcal{K}$;
while $E_i \neq \varepsilon(E_i)$
    $E_{i+1} \coloneqq \varepsilon(E_i)$;
    $R_i \coloneqq E_i \setminus E_{i+1}$;
    $i \coloneqq i + 1$;
$R_i \coloneqq E_i$;
return $(R_0, \cdots, R_i, i)$

Algorithm 2 RationalClosure

Input: A knowledge base $\mathcal{K}$ and a query $A \vdash B$
Output: TRUE iff $\mathcal{K} \approx_{\text{RC}} A \vdash B$
$(\mathcal{K}_0, \cdots, \mathcal{K}_n, n) \coloneqq \text{Rank}(\mathcal{K})$;
$i \coloneqq 0; \mathcal{K}' \coloneqq \mathcal{K}$;
while $i < n$ **and** $\overline{\mathcal{K}'} \models \neg A$
    $\mathcal{K}' \coloneqq \mathcal{K}' \setminus \mathcal{K}_i; i \coloneqq i + 1$;
return $\overline{\mathcal{K}'} \models A \rightarrow B$;

2.2.3 Relevant Closure. We mentioned that one of the potential problems of Rational Closure is that it represents a very conservative style of reasoning, and that the reason for this is that it retracts more information than intuitively seems necessary. As a solution to this problem, Casini, et. al. [6] propose *Relevant Closure* which adapts Rational Closure so that we only retract the statements in a more typical rank that actually disagree with more specific statements in higher ranks with respect to the antecedent of the query. Casini, et al. in fact describe two forms of Relevant Closure, *Basic Relevant Closure* and *Minimal Relevant Closure*, where the former is more conservative than the latter. Neither form of Relevant Closure is rational, which can arguably be seen as a weakness; however, Relevant Closure allows for significantly more permissive reasoning compared to Rational Closure and it is well-behaved from a computational perspective [6]. Though initially presented in terms of the description logic $\mathcal{ALC}$, the ideas here are applicable to propositional logic as well and here we give a definition in terms of $\mathcal{L}_{\mathcal{D}}$.

We begin by defining a notion of justification distinct from, but very much related to, the classical justifcations we discussed in Section 2.1.2. We will refer to these justifications as *ε-justifications* to distinguish from the classical case. For a knowledge base $\mathcal{K}$ and $A \in \mathcal{L}$, we define that $\mathcal{J}_\varepsilon$ is an ε-justifcation for the pair $(\mathcal{K}, A)$ iff $\mathcal{J}_\varepsilon$ is a minimal subset $\mathcal{J}_\varepsilon \subseteq \mathcal{K}$ such that A is exceptional for $\mathcal{J}_\varepsilon$. By minimal we mean that there is no $\mathcal{J}_\varepsilon{}' \subset \mathcal{J}_\varepsilon$ such that A is exceptional for $\mathcal{J}_\varepsilon{}'$. Denote the set of ε-justifications for $(\mathcal{K}, A)$ as $\mathcal{J}_\varepsilon^{\mathcal{K}}(A)$. It should be clear that ε-justifications are closely related to classical justifications; in fact, $\mathcal{J}_\varepsilon$ is an ε-justification for $(\mathcal{K}, A)$ iff $\overline{\mathcal{J}_\varepsilon}$ is a classical justification for $\overline{\mathcal{K}} \models \neg A$.

Now given a knowleddge base $\mathcal{K}$ and $A \vdash B$ we say that the statements *basically relevant*[2] to $A \vdash B$ are those that appear in an ε-justification for $(\mathcal{K}, A)$, i.e., $\bigcup \mathcal{J}_\varepsilon^{\mathcal{K}}(A)$. On the other hand, the statements *minimally relevant* to $A \vdash B$ are $\bigcup \left\{\min_{br} \mathcal{J}_\varepsilon \mid \mathcal{J}_\varepsilon^{\mathcal{K}}(A)\right\}$ where $\min_{br} \mathcal{J}_\varepsilon$ gives the statements in $\mathcal{J}_\varepsilon$ that have antecedents with the smallest base rank:

$$\min_{br} \mathcal{J}_\varepsilon = \Big\{C \vdash D \mid C \vdash D \in \mathcal{J}_\varepsilon \text{ and, for all } E \vdash F \in \mathcal{J}_\varepsilon,\ br^{\mathcal{K}}(C) \leq br^{\mathcal{K}}(E)\Big\}.$$

Given some $A \vdash B$, this allows us to split a knowledge base $\mathcal{K}$ into (R, R^-) w.r.t. either basic or minimal relevance where R is the set of statements relevant to $A \vdash B$ and R^- is the set of statements not relevant to $A \vdash B$. We now define entailment for Relevant Closure in terms of Algorithm 3 which applies to both Basic and Minimal Relevant Closure. For either Basic or Minimal Relevant Closure, let $\mathcal{K} \approx A \vdash B$ iff Algorithm 3 returns TRUE given $\mathcal{K}$, $A \vdash B$ and (R, R^-) for $\mathcal{K}$ w.r.t. the relevance for the closure at hand.

Algorithm 3 `RelevantClosure`

Input: A knowledge base $\mathcal{K}$, a query $A \vdash B$ and the partition (R, R^-) for $\mathcal{K}$
Output: TRUE iff $\mathcal{K} \approx A \vdash B$ for the closure of the partition (R, R^-)

```
(K_0, ..., K_n, n) := Rank(K);
i := 0; R' := R;
while i < n and the materialisation of R^- ∪ R' does not entail ¬A
    R' := R' \ (K_i ∩ R); i := i + 1;
return the materialisation of R^- ∪ R' entails A → B;
```

The essence of what we are doing here is similar to the algorithm for Rational Closure, and the difference is that while for Rational Closure we retract an entire rank $\mathcal{K}_i$ each iteration, here we retract only the statements in the rank responsible for making the antecedent A of a query $A \vdash B$ unsatisfiable. For Basic Relevant Closure, we retract all statements in $\mathcal{K}_i \cap \bigcup \mathcal{J}_\varepsilon^{\mathcal{K}}(A)$. For Minimal Relevant Closure, we retract the statements that are in $\mathcal{K}_i$ but also are in the smallest base rank of their respective ε-justifications. Minimal Relevant Closure is therefore less prototypical than Basic Relevant Closure and both are less prototypical than Rational Closure [6].

As an example, we will consider the query $p \vdash w$ ("Penguins have wings") for the example knowledge base we had earlier for Rational Closure since it will illustrate the difference between Rational and Relevant Closure. We gave the ranking $\mathcal{R}_0^{\mathcal{K}}, \cdots, \mathcal{R}_\infty^{\mathcal{K}}$ for this knowledge base in Figure 1 and Table 2 shows that $\mathcal{K} \not\approx_{\text{RC}} p \vdash w$. The reason for this is that we retracted the entirety of $\mathcal{R}_0^{\mathcal{K}}$ when we had $\overline{\mathcal{E}_0^{\mathcal{K}}} \models \neg p$ even though $b \vdash w \in \mathcal{R}_0^{\mathcal{K}}$ has no role in the exceptionality of the antecedent p. For Relevant Closure (in this case either Basic or Minimal), we only retract the relevant statement $b \vdash f \in \mathcal{R}_0^{\mathcal{K}}$ and $b \vdash w$ remains under consideration. We are left with $\mathcal{K}' = \{b \vdash w, p \to b, p \vdash \neg f, r \to p\}$ which has $\overline{\mathcal{K}'} \not\models \neg p$ and $\overline{\mathcal{K}'} \models p \to w$ and thus $\mathcal{K} \approx p \vdash w$ for either Basic or Minimal Relevant Closure.

This example illustrates that Relevant Closure is more permissive than Rational Closure, but it also illustrates one of its potential pitfalls: penguins, after all, do not have wings. This example makes it seem like Relevant Closure is too permissive, but on the other hand we can easily construct examples where Rational Closure seems too conservative [6]. The style of reasoning that is optimal ultimately depends on context and the characteristics of the knowledge base in question.

2.2.4 Explanations for KLM. There is limited work on explanations for KLM; however, one of the main works of interest here is that of Chama [8] which gives an algorithmic definition of justifications for Rational Closure. The key idea here is to take the algorithm for Rational Closure and make use of classical justifications for $A \to B$ w.r.t. a knowledge base taken to be the statements we have not retracted, i.e., $\mathcal{E}_r^{\mathcal{K}}$ with $r = br^{\mathcal{K}}(A)$. This mirrors the approach we had for defeasible entailment: after eliminating more typical ranks, we rely on classical tools to reason about the knowledge base, only in this case we use classical justification instead of classical entailment. Algorithm 4 summarises this result formally for $\mathcal{L}_{\mathcal{D}}$ where $\underline{\mathcal{K}} = \{A \vdash B \mid A \to B \in \mathcal{K}\}$ for a knowledge base $\mathcal{K}$.

Algorithm 4 `JustifyRationalClosure`

Input: A knowledge base $\mathcal{K}$ and a query $A \vdash B$
Output: The justifications for $\mathcal{K} \approx_{\text{RC}} A \vdash B$

```
(K_0, ..., K_n, n) := Rank(K);
i := 0; K' := K;
while i < n and the materialisation of K' does not entail ¬A
    K' := K' \ K_i; i := i + 1;
return {underline(J) | J ∈ J^{materialisation of K'}(A → B)};
```

As an example, consider the query $r \vdash \neg f$ for the example knowledge base $\mathcal{K}$ we used for Rational Closure. In Figure 1 and Table 2 we found that $br^{\mathcal{K}}(r) = 1$. When testing for $\approx_{\text{RC}}$ entailment, we would then evaluate whether $\overline{\mathcal{E}_1^{\mathcal{K}}} \models r \to \neg f$. Now instead we find the classical justifcations for this entailment (denoted J):

$$J = \mathcal{J}^{\overline{\mathcal{E}_1^{\mathcal{K}}}}(r \to \neg f).$$

Materialising $\mathcal{E}_1^{\mathcal{K}}$ gives the following:

$$\overline{\mathcal{E}_1^{\mathcal{K}}} = \{\neg(p \to b) \to \bot, \neg(r \to p) \to \bot, p \to \neg f\}.$$

And hence $J = \{\{\neg(r \to p) \to \bot, p \to \neg f\}\}$. We are not done quite yet because the justifications in J are subsets of the materialisation $\overline{\mathcal{K}}$ and not $\mathcal{K}$. This is not an issue because we can find defeasible counterparts for each of these statements, which gives us the final result of $J' = \{\{\neg(r \to p) \vdash \bot, p \vdash \neg f\}\}$. This example, though simple, illustrates that the algorithm given by Chama finds a set of statements that defeasibly entail the query while still respecting the overall reasoning process for Rational Closure.

Another idea here is the notion of *strong explanations* given by Brewka and Ulbricht [5], which is proposed not for KLM specifically but generally for nonmonotonic formalisms. A summary of this

[2] These terms do not appear in the literature but are here to aid explanation.

result as applied to propositional logic is that we take the requirement that $\mathcal{J} \not\approx A \vdash B$ for a minimal $\mathcal{J} \subseteq \mathcal{K}$ where $\mathcal{K} \not\approx A \vdash B$ and additionally require that the conclusion $A \vdash B$ is never retracted when using additional statements in $\mathcal{K}$, i.e., for $\mathcal{J}$, there is no $\mathcal{J}' \subseteq \mathcal{K}$ such that $\mathcal{J} \cup \mathcal{J}' \not\approx A \vdash B$. This addition is important; naively adapting the definition for classical justification produces unsound results [5] and in fact the added requirement addresses the key problem here by requiring the use of the most specific information justifications when typical and specific statements disagree with respect to a particular query. Intuitively, this property also seems to have some correspondence with the algorithm given by Chama [8] as the latter also does not to permit justifications that use more typical information which disagrees with more specific information.

3 DISCUSSION

The result given for Rational Closure by Chama [8] is promising in that it suggests that we may be able to produce similar justification algorithms for Relevant and Lexicographic Closure. It seems quite likely that the same approach might prove useful for these formalisms. To our knowledge, such algorithms have not yet been described in the literature.

The question of finding declarative properties for defeasible justifications is perhaps more difficult. As a useful starting point, it would be interesting to see if we can relate the algorithm for justifications of Rational Closure [8] with the definition of strong explanations [5] applied to propositional logc. Overall, the likely approach here is to strengthen the condition for justifications of classical entailments and to do so in a way that has no effect for knowledge bases of purely classical information. Intuitively, we would expect the concept of defeasible justification to extend the concept of classical justification.

4 CONCLUSIONS

We have discussed a number of essential features of classical propositional logic, classical justification and KLM-style defeasible propositional logic which we then used as the basis to explore current literature on explanation for KLM formalisms. While limited work has been done in this area, current results suggest some avenues to advance our understanding of defeasible explanation for KLM, both in terms of modifying algorithms to produce justifications and in terms of describing justification for KLM axiomatically. Such advancements would be welcome given that explanation services are a crucial aspect of reasoning systems and these ideas are not currently well-understood for KLM; in fact, such results may work toward cemeting KLM formalisms as practical reasoning systems.

REFERENCES

[1] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider. 2010. *The Description Logic Handbook: Theory, Implementation and Applications* (2nd ed.). Cambridge University Press, USA.

[2] S. P. Bail. 2013. *The justificatory structure of OWL ontologies.* Ph.D. Dissertation. University of Manchester.

[3] Mordechai Ben-Ari. 2012. *Propositional Logic: Formulas, Models, Tableaux.* Springer London, London, 7–47. https://doi.org/10.1007/978-1-4471-4129-7_2

[4] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, Vol. 8. 8–13.

[5] Gerhard Brewka and Markus Ulbricht. 2019. *Strong Explanations for Nonmonotonic Reasoning.* Springer International Publishing, Cham, 135–146. https://doi.org/10.1007/978-3-030-22102-7_6

[6] Giovanni Casini, Thomas Meyer, Kodylan Moodley, and Riku Nortjé. 2014. Relevant Closure: A New Form of Defeasible Reasoning for Description Logics. In *Logics in Artificial Intelligence*, Eduardo Fermé and João Leite (Eds.). Springer International Publishing, Cham, 92–106.

[7] Giovanni Casini, Thomas Meyer, and Ivan Varzinczak. 2019. Taking Defeasible Entailment Beyond Rational Closure. In *Logics in Artificial Intelligence*, Francesco Calimeri, Nicola Leone, and Marco Manna (Eds.). Springer International Publishing, Cham, 182–197.

[8] V. Chama. 2020. *Explanation for defeasible entailment.* Master's thesis. University of Cape Town, Cape Town, South Africa. http://hdl.handle.net/11427/32206

[9] Stephen A. Cook. 1971. The Complexity of Theorem-Proving Procedures. In *Proceedings of the Third Annual ACM Symposium on Theory of Computing* (Shaker Heights, Ohio, USA) *(STOC '71)*. Association for Computing Machinery, New York, NY, USA, 151–158. https://doi.org/10.1145/800157.805047

[10] Michael Freund. 1998. Preferential Reasoning in the Perspective of Poole Default Logic. *Artificial Intelligence* 98, 1–2 (Jan. 1998), 209–235. https://doi.org/10.1016/S0004-3702(97)00053-2

[11] L. Giordano, V. Gliozzi, N. Olivetti, and G.L. Pozzato. 2015. Semantic characterization of rational closure: From propositional logic to description logics. *Artificial Intelligence* 226 (2015), 1–33. https://doi.org/10.1016/j.artint.2015.05.001

[12] C. Grosan and A. Abraham. 2011. *Knowledge Representation and Reasoning.* Springer Berlin Heidelberg, Berlin, Heidelberg, 131–147. https://doi.org/10.1007/978-3-642-21004-4_6

[13] Matthew Horridge. 2011. *Justification Based Explanation in Ontologies.* Ph.D. Dissertation. University of Manchester.

[14] A. Kaliski. 2020. *An overview of KLM-style defeasible entailment.* Master's thesis. University of Cape Town, Cape Town, South Africa. http://hdl.handle.net/11427/32743

[15] Sarit Kraus, Daniel Lehmann, and Menachem Magidor. 1990. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence* 44, 1 (1990), 167–207. https://doi.org/10.1016/0004-3702(90)90101-5

[16] Markus Krötzsch and Daria Stepanova. 2019. *Reasoning Web. Explainable Artificial Intelligence: 15th International Summer School 2019, Bolzano, Italy, September 20–24, 2019, Tutorial Lectures.* Vol. 11810. Springer Nature.

[17] Daniel Lehmann. 1995. Another perspective on default reasoning. *Annals of mathematics and artificial intelligence* 15, 1 (1995), 61–82.

[18] Daniel Lehmann and Menachem Magidor. 1992. What does a conditional knowledge base entail? *Artificial Intelligence* 55, 1 (1992), 1–60. https://doi.org/10.1016/0004-3702(92)90041-U

[19] Olga Ohrimenko, Peter J. Stuckey, and Michael Codish. 2007. Propagation = Lazy Clause Generation. In *Principles and Practice of Constraint Programming – CP 2007*, Christian Bessière (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 544–558.

[20] Stefan Schlobach and Ronald Cornet. 2003. Non-Standard Reasoning Services for the Debugging of Description Logic Terminologies. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence* (Acapulco, Mexico) *(IJCAI'03)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 355–360.