

Explanations Using Defeasible Rule Sets

PRESENTER

Antony Baker, Daniel Schlesinger, Augustine Mochoeneng

DATE

October 2026

EDRS



UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

IT | SCHOOL OF IT

Where do AI models explanations come from?

Classifiers make **high-stakes** calls: loans, diagnoses.

People need to know **why**, but most models are black boxes.

Rule-based classifiers learn **IF-THEN rules** from data, so every prediction comes with the rule that made it.

The problem

Real data has exceptions, so rules **conflict**. ECCRS settles it with a procedure: the most specific rule wins.

QUESTION

Is that sound reasoning, or just a convenient tie-breaker?

Defeasible reasoning is the formal logic of exceptions. Madzime & Meyer proved the two agree under **three assumptions**. We built a workbench to test it on real rule sets.

ECCRS: the explanation is the computation

MOST SPECIFIC WINS

Applicable rules

$$\mathcal{A}(F) = \{(B, \ell) \in \mathcal{R} : F \models B\}$$

Inclusion-maximal rules

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

MSW operator

$$\text{MSW}_{\mathcal{R}}(F) = \begin{cases} \ell & \text{if all rules in } \mathcal{M}(F) \text{ have label } \ell \\ \perp & \text{if } \mathcal{M}(F) = \emptyset \quad (\text{abstain}) \\ \text{halt} & \text{if } \mathcal{M}(F) \text{ has both labels} \end{cases}$$

EXPLANATION TRACE

r1: odour = none $\Rightarrow$ 1 (edible)

r2: odour = none $\wedge$ spore = green $\Rightarrow$ 0

r3: odour = foul $\Rightarrow$ 0

F : odour = none, spore = green

$\mathcal{A}(F) = \{r_1, r_2\}$

$B_{r_1} \subset B_{r_2}$, so r1 is overridden

$\mathcal{M}(F) = \{r_2\}$

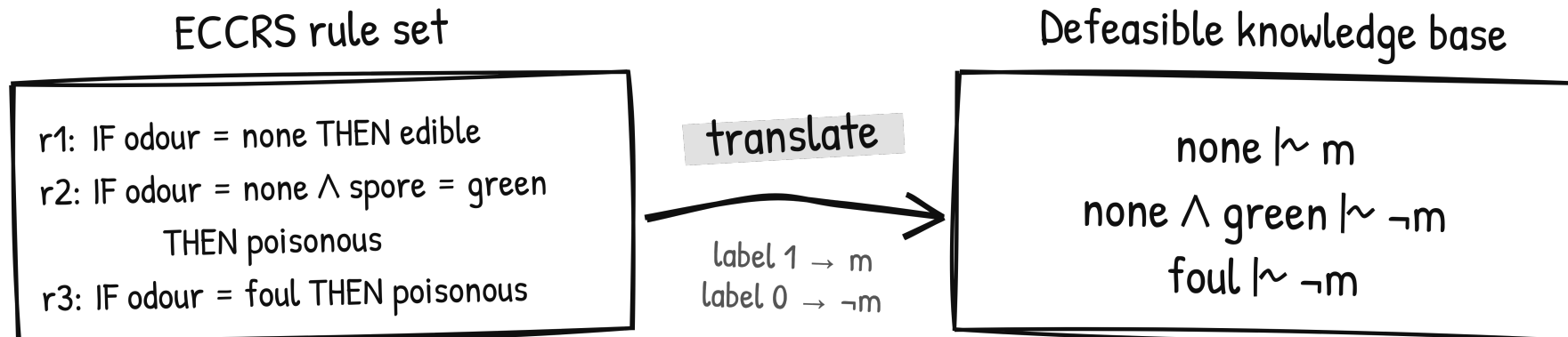
$\text{MSW}(F) = 0$ **poisonous**

*The steps that compute the answer **are** the explanation.*

From rules to defeasible reasoning

DEFEASIBLE REASONING (KLM)

bird $\sim$ flies penguin $\rightarrow$ bird penguin $\sim \neg$ flies the specific statement wins



Three closures, one question

RATIONAL CLOSURE

Discards whole ranks of statements, most general first.

Daniel Schlesinger

LEXICOGRAPHIC CLOSURE

Keeps as many statements from the clashing rank as it can.

Augustine Mochoeneng

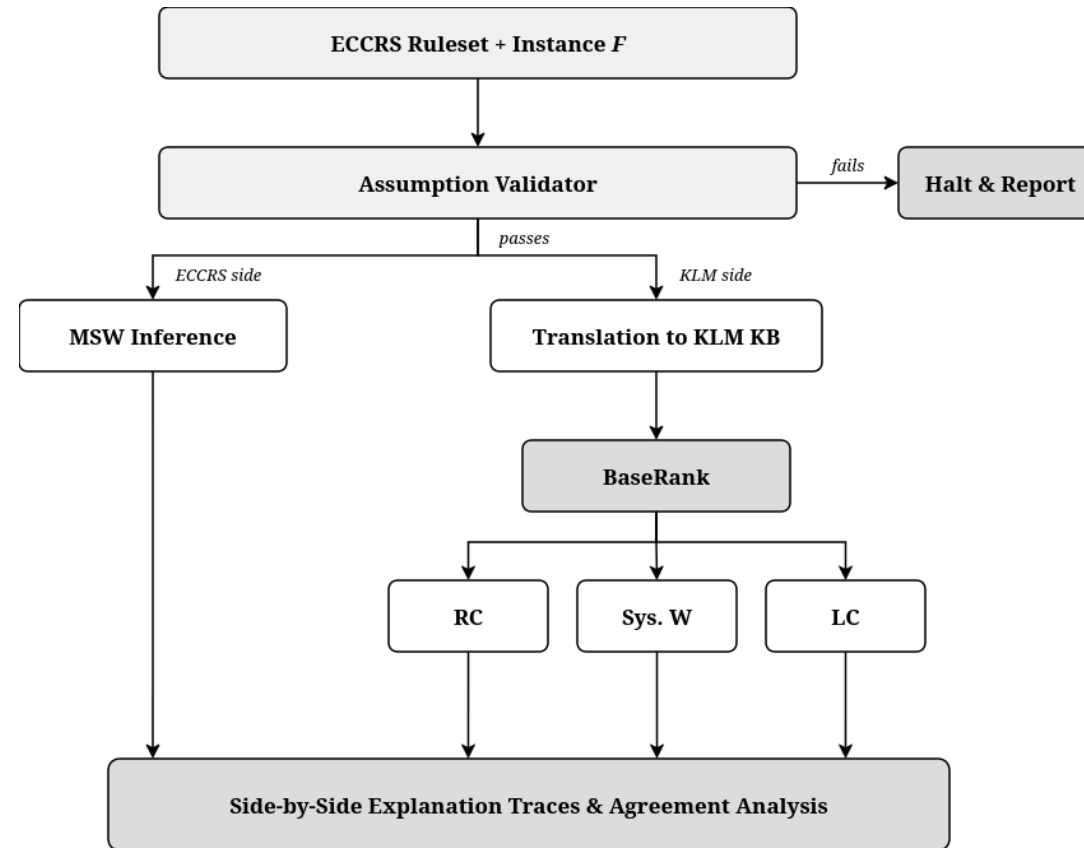
SYSTEM W

Compares worlds by the set of statements they break.

Antony Baker

For every instance F : does $F \models m$ or $F \models \neg m$?

The workbench



The workbench cont.

```

ECCRS workbench demo: cancer

12 rules, 6 rule atoms, 175 CSV instances

== Validation (alignment-theorem conditions) ==
PASS sanity
PASS strict closure
PASS no total override

== BaseRank (one-off cost: 23 SAT calls) ==
rank 0: 0, 1, 6, 10, 11
rank 1: 2, 4, 8, 9
rank 2: 3, 7
rank 3: 5

== Comparison ==

```

mode	population	evaluated	SAT calls	time	n	agree	disagree	both-abstain	agreement
naive	pattern space (2 ⁶)	64 patterns	360	0.008s	64	64	0	0	100.0%
naive	realisable subset	-	-	-	64	64	0	0	100.0%
naive	CSV instances	175 instances	808	0.021s	175	175	0	0	100.0%
optimised	pattern space (2 ⁶)	10 classes	64	0.001s	64	64	0	0	100.0%
optimised	realisable subset	-	-	-	64	64	0	0	100.0%
optimised	CSV instances	175 instances	183	0.005s	175	175	0	0	100.0%

```

agreement = (agree + both-abstain) / n
pattern space = every complete assignment over the rule atoms; realisable = those one-hot encoding allows
Speed-up over the pattern space: SAT calls 360 -> 64 (5.6x), wall-clock 0.008s -> 0.001s (6.9x)

== Explanation traces ==

--- pattern 405 ---
ECCRS side (pattern ~a(1) & a(2) & a(4) & ~a(5) & ~a(7) & ~a(8))
11: a(2) => 0 [overridden by 4, 9]
9: a(2) & ~a(5) => 1 [overridden by 4]
4: ~a(1) & a(2) & a(4) & ~a(5) => 1 [MAXIMAL]
MSW: label = 1

RC side
query 1: F |~ m (supports class 1)
query 2: F |~ ~m (supports class 0)
br(F) = 1, br(F & m) = 1, br(F & ~m) = 2
br(F) < br(F & ~m) holds -> RC entails class 1
decisive conditionals at rank 1, from ECCRS rules: 4, 9
=> AGREE

```

Results

ASSUMPTIONS HOLD			
Instances	RC	LC	System W
13,509	100%	100%	100%
Bank, Breast Cancer and Mushroom rule sets.			

ASSUMPTIONS BROKEN (MSW VS LC)				
Broken	Agree	Abstain	Conflict	Wrong
Closure	13,395	114	0	0
Sanity	12,019	0	1,490	0
Total override	13,374	0	135	0
13,509 instances per row.				

CONCLUSION

**ECCRS explanations are not just readable.
They are backed by formal reasoning.**