IMPERIAL

Argumentative Repair for Neural Networks

Shivangi Kumar, Francesca Toni, Xiang Yin

AIxIA - AI³ 2026

Motivation

Neural networks produce misclassifications

Neural networks are used in high-stakes settings, however even strong performing models still get cases wrong.

1

Widely deployed

Healthcare, finance, security - decisions that carry real consequences.

2

Still misclassify

Even high-accuracy networks make confident errors on individual inputs.

3

Trust at stake

One wrong call in a critical setting undermines confidence in the whole system.

So, the question is: **can we fix the errors a network makes without retraining it from scratch?**

Motivation
Existing repair methods

Repair methods adjust a trained network to fix its mistakes, but they share a limitation.

Existing repair (I-Repair) [1]

Adjust a trained network to fix misclassified inputs, localising the change.

The limitation: contamination

A small weight change can move the boundary sharply, breaking inputs that were already correct.

[1] Henriksen, Leofante & Lomuscio, "Repairing Misclassifications in Neural Networks Using Limited Data", SAC'22.

Research Questions

What can we do differently to existing approaches?

Baseline: I-Repair - a gradient-based method that updates biases and edge weights together.

RQ1

Which components to repair?

We separate bias repair from edge-weight repair and ask whether one is more effective, and whether they fix different errors.

RQ2

Which attribution method?

Does the attribution method matter? We compare gradient-based against removal-based.

Argumentation

How do we answer the research questions?

Computational argumentation models a decision as a debate - some things argue for it, some against, and the verdict follows from the balance. **Prior work shows a neural network can be viewed this way [2]**.

Explanation methods then score how much each part contributed to the verdict, revealing what drove the decision.

Existing work: A Neural Network can be mapped to a QBAF

Neuron	→	Argument
Positive weight	→	A support
Negative weight	→	An attack
Bias	→	Base score
Output neuron	→	Topic argument

Answers RQ1: Arguments and relations separate cleanly

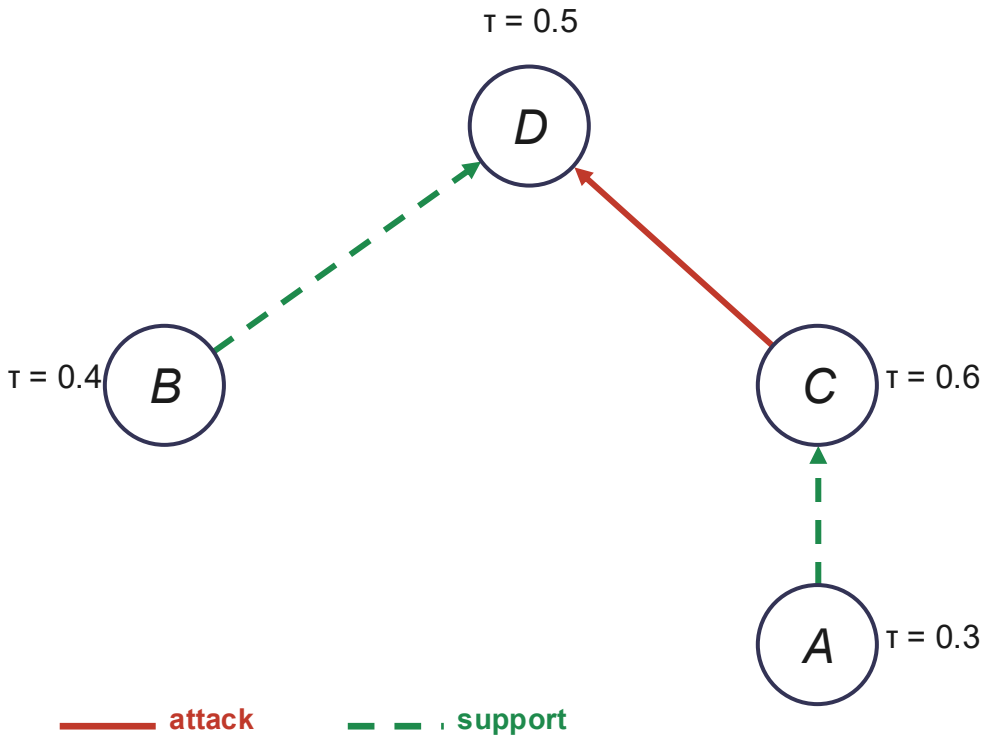
Answers RQ2: Different attribution methods give different explanations, so we can compare them.

[2] Potyka, N. (2021). Interpreting Neural Networks as Quantitative Argumentation Frameworks. AAAI 35, 6463–6470. doi:10.1609/aaai.v35i7.16801

Argumentation
What is a QBAF?

Quantitative Bipolar Argumentation Framework

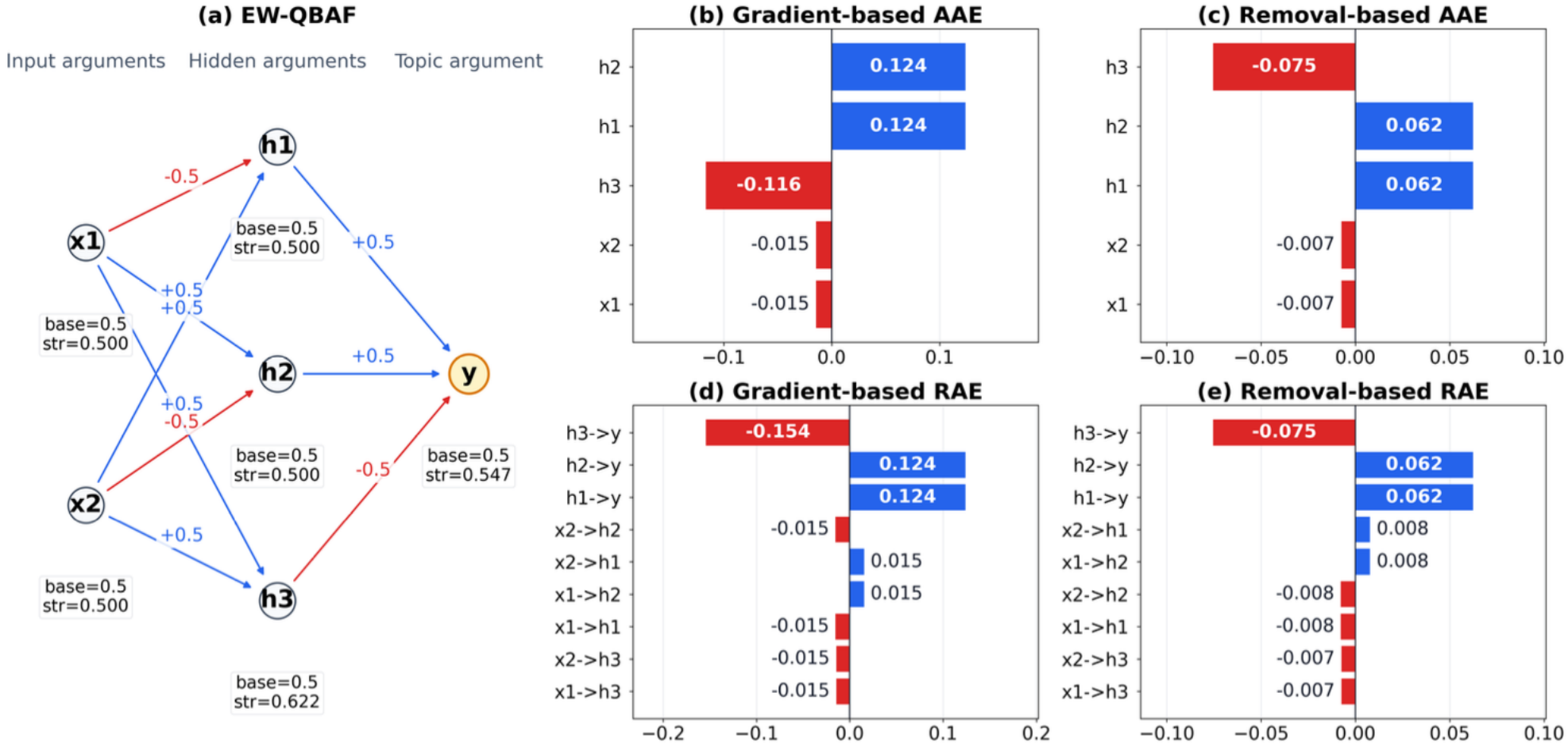
- **Arguments** – the claims in the debate
- **Attack and Support** – relations between them
- **Base score τ** – each argument's initial strength
- **Gradual semantics** – combine the initial strengths and compute the final strength of each argument



A small QBAF: C attacks D (red); B and A give support (green dashed). Nodes annotated with base scores τ.

Argumentation

Attribution Explanations and Scoring Methods

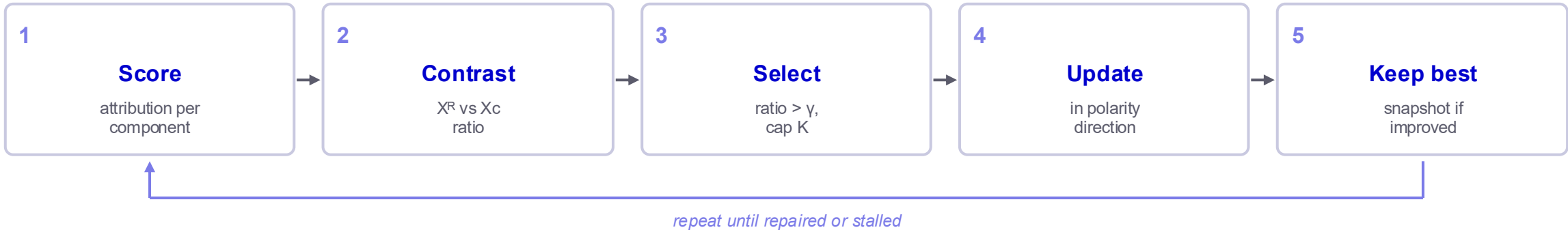


Methods

What was implemented

Bias updates guided by AAE, edge-weight updates guided by RAE, each in gradient and removal variants.

Both algorithms run the same loop. They differ only in what they score and what they change - arguments & biases, or relations & edge weights.

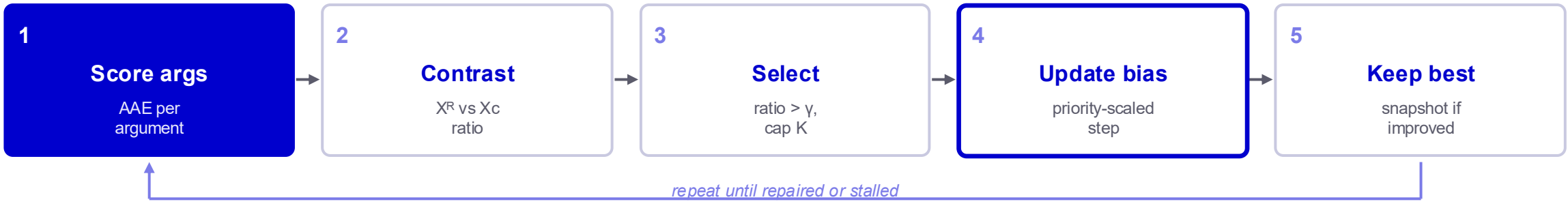


Multi-contestability

We contrast each component's influence on misclassified (X^R) vs correct (Xc) samples. Only those above the ratio γ are repaired - one update for every error at once.

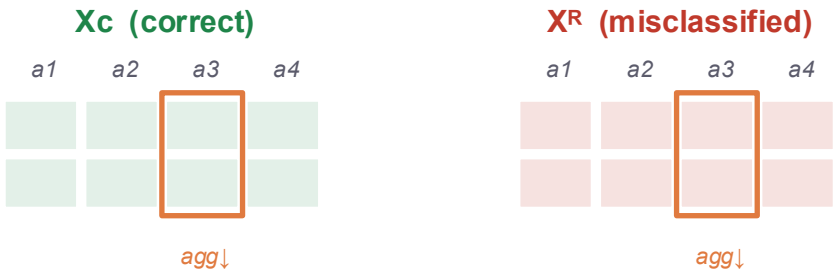
Methods

AAE guided bias update



AAE scores: build one matrix per set, contrast by column

(rows = samples, cols = arguments)



ratio = **agg(X^R)** / **agg(Xc)** > **γ**

*agg = **an aggregator** of each column's |scores|*

→ repair this argument's bias

What & why

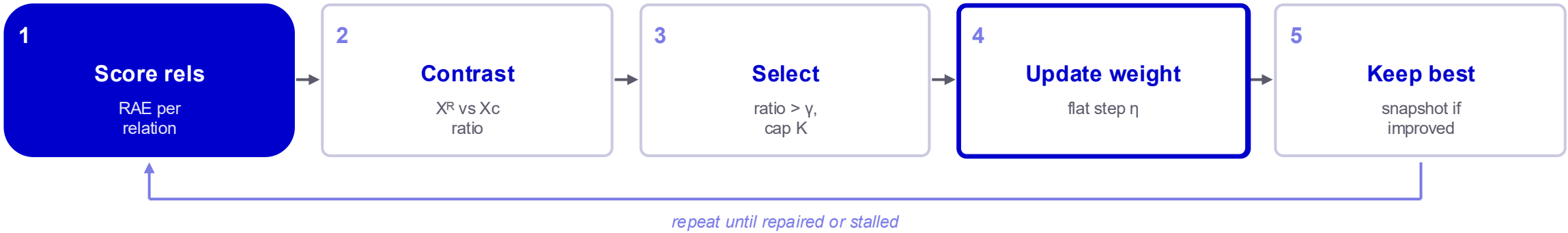
Arguments = neurons, scored by AAE. A bias change shifts a neuron's threshold.

Step is priority-scaled — arguments with stronger error-side influence move more.

Simplest axis: one parameter per neuron, moves the boundary broadly.

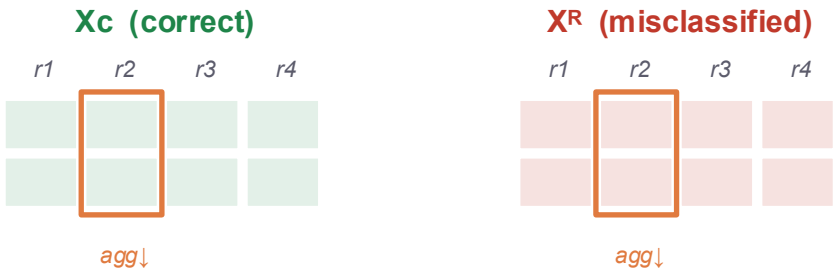
Methods

RAE guided edge weight update



RAE scores: build one matrix per set, contrast by column

(rows = samples, cols = relations)



ratio = $agg(X^R)$ / $agg(Xc)$ > γ

*agg = **an aggregator** of each column's |scores|*

What & why

Relations = weighted edges, scored by RAE. A weight change reshapes one pathway.

Flat step — RAE values are smaller (attenuated through sigmoids), so no priority scaling.

More granular axis: targets pathways, so it stays more contained under pressure.

Experimental Setup

Common setup across all experiments

Three datasets

Pima — diabetes · 768 · 8 features
SAHeart — heart · 462 · 9 features · imbalanced
Ionosphere — radar · 351 · 34 features

Chosen to vary the three things that affect repair: sample size, feature dimensionality, and class balance.

The split (70 / 15 / 15)

Train 70% — trains the initial MLP only.

Repair 15% — what repair acts on. Split into correct (preserve) and misclassified (fix). Disjoint from training — no original data at repair time.

Eval 15% — held out entirely; all generalisation measured here.

Multi-seed (× 10)

The whole pipeline — split, train, repair — repeats over 10 random seeds.

Every result is a mean ± sd.

Essential: repair sets are only tens of samples, so a single split is unreliable — the sd shows how much a result rides on the split.

How we measure repair

Four metrics per run

ΔEval

Change in held-out accuracy (pp)

The metric of record - the only one that says whether a repair helped on unseen data.

Reported throughout as mean ± sd over 10 seeds.

% Repaired

Fraction of the error set X fixed

Necessary but not sufficient: a high repair rate means little if generalisation falls.

Cardinality

$|R'| < |R|$

Did the total number of misclassifications fall after repair? A minimal bar to clear.

Subset soundness

$R' \subset R$

The strict notion: errors fell and not one previously-correct sample was broken.

Results - bias update

Conservative gains are reliable

Bias repair (AAE)

Gradient: conservative wins

Small, reliable gains when applied gently. Pushed aggressive, it repairs more but **overfits the repair set** and generalisation drops.

+1.55 conservative ΔEval, Pima

Removal ≈ gradient here

In the conservative setting the two attribution methods are **effectively the same** - the configuration sets the boundary shift, not the attribution.

≈ same ΔEval, gradient vs removal

Takeaway: **on the bias axis, how conservatively you apply repair matters more than which attribution you use.**

Results - edge-weight update

Gradient edge-weight is the most robust

Edge-weight repair (RAE)

Gradient: the strongest method

Most **robust to aggressiveness** — on Pima ΔEval even rises across configs (+1.29 → +1.47 → +2.59).

+6.42 best ΔEval of any method, Ionosphere

Removal: structurally cautious

Deleting a whole edge is too large a change to discriminate, so it **barely moves the model** - safe, but rarely repairs.

0.8–10% of errors repaired, any setting

Takeaway: the attributions diverge here - gradient edge-weight is the standout, removal stays cautious by design.

Limitations

1 Theoretical evaluation, not real world

ΔEval on benchmarks is a proxy, not a test of trustworthiness in the real high stakes setting.

2 Sound repair is never guaranteed

A global update on a shared boundary can always break some correct samples. We contain the risk, but we cannot remove it.

Conclusions and future work

What we found, and where it goes next

Conclusions

1. **Separating the components pays off**
   Bias and edge-weight repair behave differently — a joint update can't see that.

2. **Conservative matches the baseline**
   At conservative settings, generalisation is on par with I-Repair — while repairing more samples.

3. **Gradient does the work**
   Gradient attribution drives effective repair; removal stays cautious.

Future work

1. **Shapley attribution**
   A third attribution route, well-motivated for QBAFs, to complete the RQ2 comparison.

2. **Scale up**
   Larger, more complex datasets, and extending the idea toward LLMs.