Modal Logic Neural Networks

Differentiable Kripke semantics with a *learnable*, *inspectable* accessibility relation

Antonin Sulc Noor Naddour

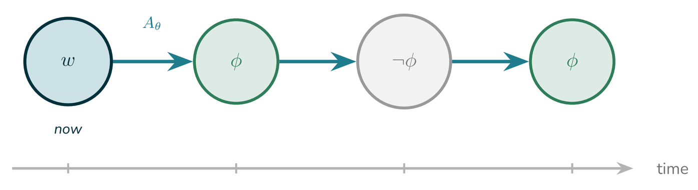
Lawrence Berkeley National Laboratory • The University of Queensland • NeSy 2026

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY Office of Science

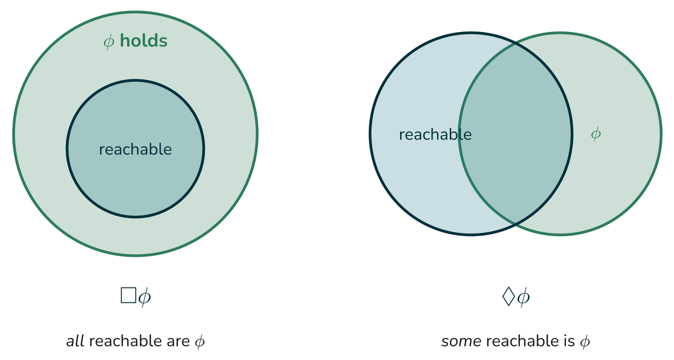
The accessibility relation is the *unknown*: one differentiable □/◊ engine **learns** it, **bounds** truth as $[L, U]$, and **exposes** it: the matrix that computes the answer is the matrix you read.

A frame: worlds in time



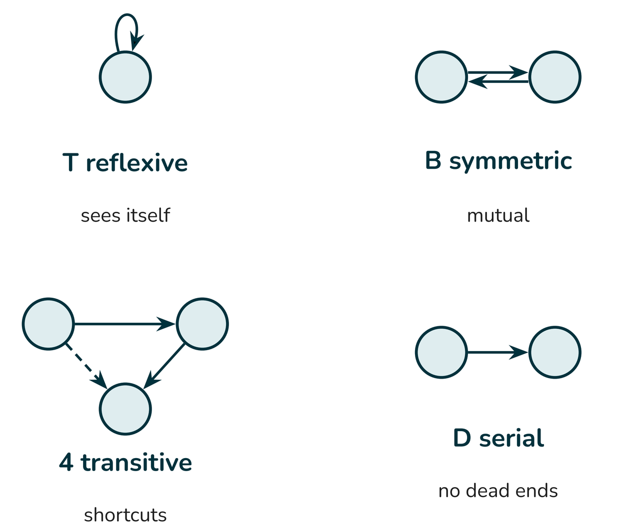
- Worlds are **moments**; from *now* (w) the future is reachable via A_θ.
- $\Box\phi$ = *always*: does *every* future moment satisfy ϕ? **(no)**
- $\Diamond\phi$ = *eventually*: does *some*? **(yes)**
- A_θ (which world bears on which) is the **unknown** we learn.

□ = all, ◊ = some



- **Necessity** $\Box\phi$: reachable $\subseteq \{\phi\}$.
- **Possibility** $\Diamond\phi$: reachable $\cap \{\phi\} \neq \varnothing$.
- Truth at w lands in $[\Box\phi, \Diamond\phi]$.

Axioms are the shape of A_θ

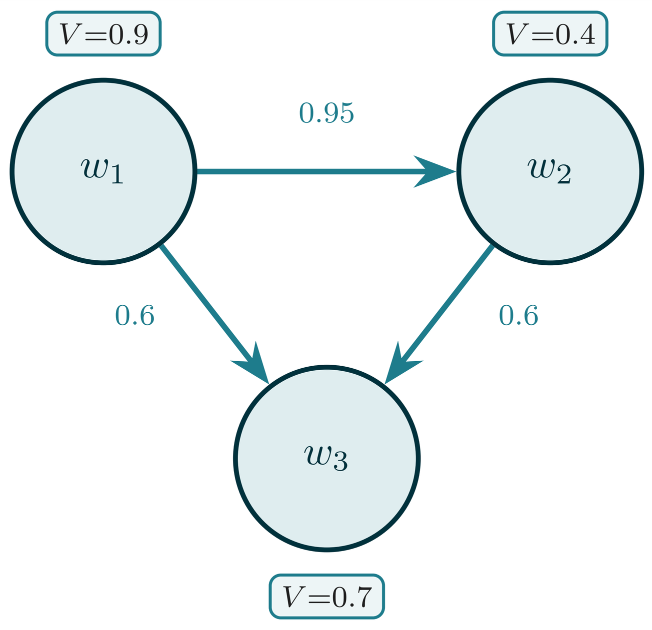


- The modality is set by the relation's *shape*.
- **T** reflexive, **B** symmetric, **4** transitive, **D** serial.
- We **measure** them after training, not impose them.

Where MLNN sits

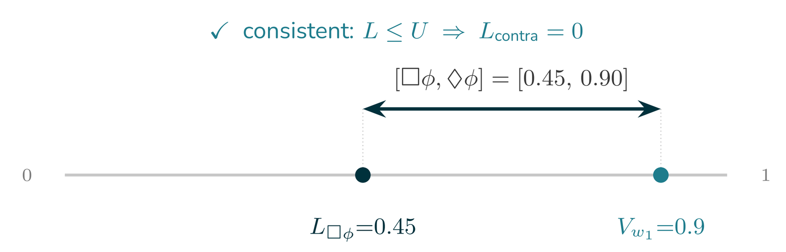
- **Generalises LNNs**: classical weighted logic is the zero-one case.
- vs *DeepProbLog* / *Semantic Loss*: adds *cross-world* modality.
- vs *SATNet* / *RRN*: solves from axioms *and* learns the relation itself.

Differentiable Kripke



- Worlds + a **learnable** relation $A_\theta \in [0, 1]$ (the arrow weights).
- Each world carries a valuation $V_w(\phi)$; truth is a bound $[L, U]$, not a scalar.

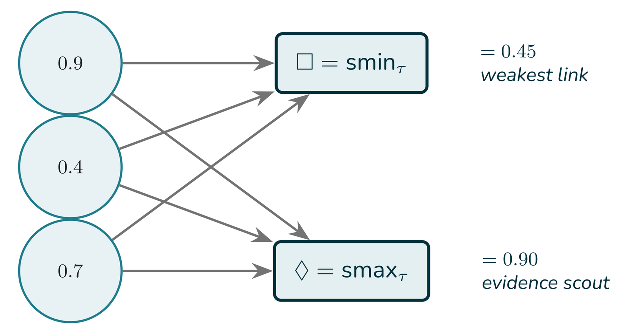
The truth bracket $[\Box\phi, \Diamond\phi]$



- Firing □/◊ at w_1 **brackets** its value $[0.45, 0.90]$; here $L \leq U$, so **consistent** (L_{contra}=0).
- A crossed bound $L > U$ is a **contradiction**: the training signal.

Numerical example: a source with $[L, U]$=$[0.70, 0.40]$ has $L > U$, so L_{contra} = $\max(0, L{-}U) = \mathbf{0.30}$.

Two modal neurons

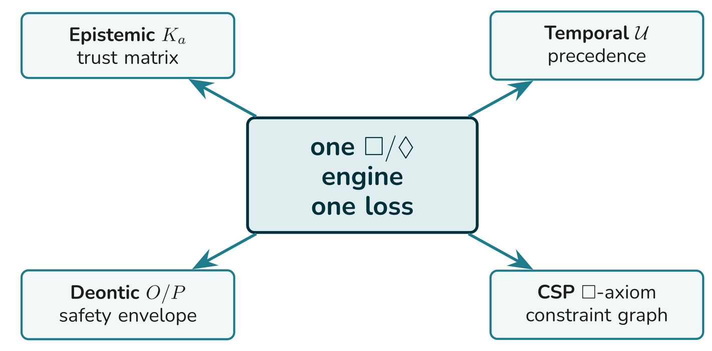


- □ = soft *min* (weakest link).
- ◊ = soft *max* (evidence scout); both *exact* as $\tau \to 0$.
- Train by minimising $L_{\text{contra}} = \sum \max(0, L{-}U)$.

Every axiom is a few lines

```
# temporal "until" (ROCStories)
s = F.until(chain, is_last, A)
# doxastic trust (coherence)
loss = (1 - F.necessity(coh, A))**2
# deontic safety (C-MAPSS)
Ld = (1-Uf)**2 + (1-Ls)**2
# frame-axiom audit & top-k mask
S = F.audit(A); A = F.topk(A, k=8)
```

One engine, four modalities



- One □/◊ engine, one contradiction loss: epistemic, temporal, deontic & constraint reads.
- **Sound** bounds; **scalable** to $20{,}000$ worlds; axioms audited vs. a **matched null**.

The `torchmodal` library

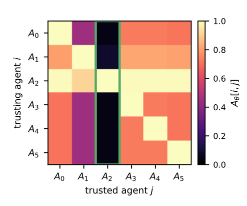
```
$ pip install torchmodal
from torchmodal import
    functional as F
box = F.necessity(prop, A)
dia = F.possibility(prop, A)
```

- One engine; A_θ is a first-class, *inspectable* object (see recipe, left).

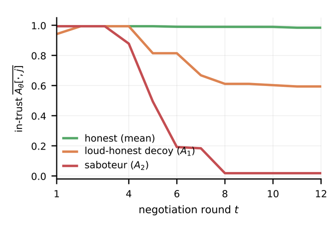
Doxastic: a trust matrix

6 LLM agents debate a $50K budget; **agent 2** saboteurs. An NLI head scores each utterance into $[L_j, U_j]$; a flip-flopper drives $L_j{>}U_j$. **Axiom** $A_\theta[i, j] \to$ consistent(j)=$1{-}$relu$(L_j{-}U_j)$: maximising $\Box_i$coherent drains the saboteur's *column* to 0.

× **agent 2 flip-flops** *real utterances*
"I concur with the $50K budget" *(entail 1.0)*
"we need a minimum of $75K" *(contradict 1.0)*



trust matrix A_θ



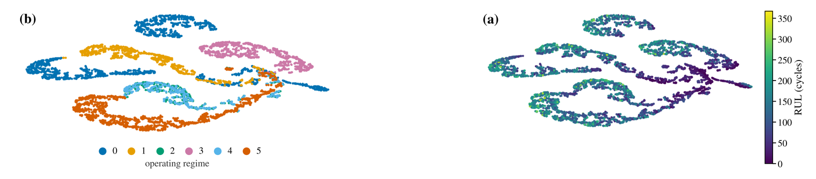
in-trust per round

12/15 vs judge 3/15 (subtle); **15/15** vs 5/15 (inject).

- Detection is the **NLI front-end**; the modal layer adds the **audit trail** & **injection immunity**.
- **Belief-revision knob:** the memory window sets a forgiveness policy, persistent, or forgiving once the saboteur reforms.

Deontic: C-MAPSS turbofans

One deontic loss organises the sensor embedding by **regime** *and* **wear** at once:



embedding: **regime** degradation: **wear (RUL)**

- $48{,}299$ (engine, cycle) worlds; both heads read **sensors only**.
- Regime decodes at 0.91; a ridge probe to RUL reaches ρ=0.71; no RUL at inference.

Temporal: ROCStories

Score each ordering σ of the 5 sentences with *one* **Until**:

$$(\phi\,\mathcal{U}\,\psi)_t = \psi_t \vee (\phi_t \wedge (\phi\,\mathcal{U}\,\psi)_{t+1})$$

ϕ_t=$A_\theta(\sigma_{t-1}, \sigma_t)$ ("next follows"), ψ_t="is last"; the top permutation wins.

✓ **recovered** *illustrative*
(1) Mia sowed a seed · (2) she watered it · (3) a sprout rose · (4) it bloomed · (5) she beamed
order: $1 \to 2 \to 3 \to 4 \to 5$

× **swaps a pair** *illustrative*
(1) Sam packed · (2) he checked the sky · (3) he grabbed an umbrella · (4) he left · (5) rain fell
order: $1 \to 3 \to 2 \to 4 \to 5$: (2) & (3) swap.

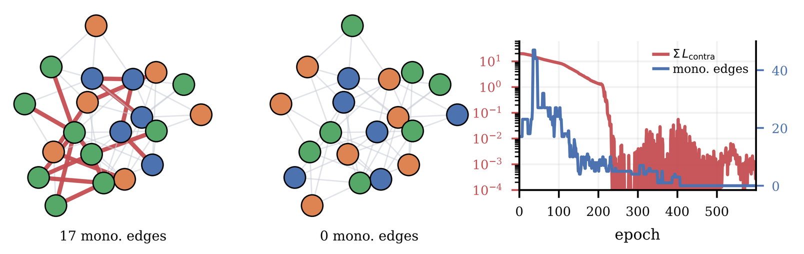
ties 7B LLM pair 0.77 / exact 0.24, at $\sim 10^{-4}$ params.

CSP: graph colouring

Adjacency **is** accessibility: "adjacent nodes differ" is exactly the modal axiom:

$$\bigwedge_{c=1}^{k} (p_c \to \neg\Diamond p_c)$$

"if I am colour c, no accessible (adjacent) node is c." Solve with R fixed and L_{task}=0.



29/30 solved the modal operator, not the gradient, carries it.

- Ties an RRN-GNN; beats non-modal ($18/30$) & *Semantic Loss* ($26/30$).
- **Complementary** to exact solvers: exposes L_{contra} & recovers the relation.

Future work

Sub-quadratic A_θ via approximate max-inner-product search.

Enforced, not just audited frames when a domain mandates a system.

Scale the doxastic test to 30B+ backbones.

Richer fragments: *since/until*, graded, probabilistic.

Framework available at `github.com/sulcantonin/torchmodal` • `pip install torchmodal`