Towards Model Repair by Human Opinion-Guided Reinforcement Learning

Kyanna Dagenais

McMaster University – Hamilton, Canada

dagenaik@mcmaster.ca | kyannadagenais.ca

McMaster University

Sustainable Systems and Methods

Context and Research Problem

Model repair

Goal

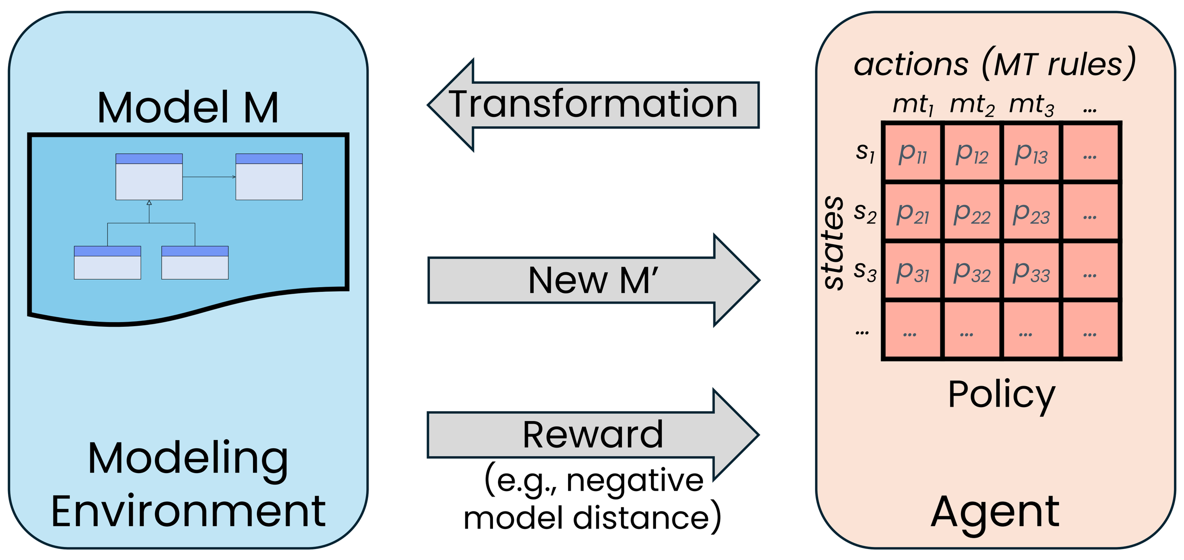
- Repair invalid models by model transformations

Problem

- Complex models → long repair sequences
- Automation is needed

Opportunity

- Modeling projects are longitudinally extensive
- Learn repair patterns *as we go*



Reinforcement learning (RL)

Basics

- Agent learns by trial and error
- Pro: does not require historical data
- Policy: state to MT-rule mapping (probability)

Limitation

- Shallow learning curve (learning takes time)

Goal

- Guide RL by rapidly emerging (uncertain) opinions

Methods

Approach

Benchmark

- Frozen Lake (Open AI)
- Performance metric: cumulative reward

Opinions

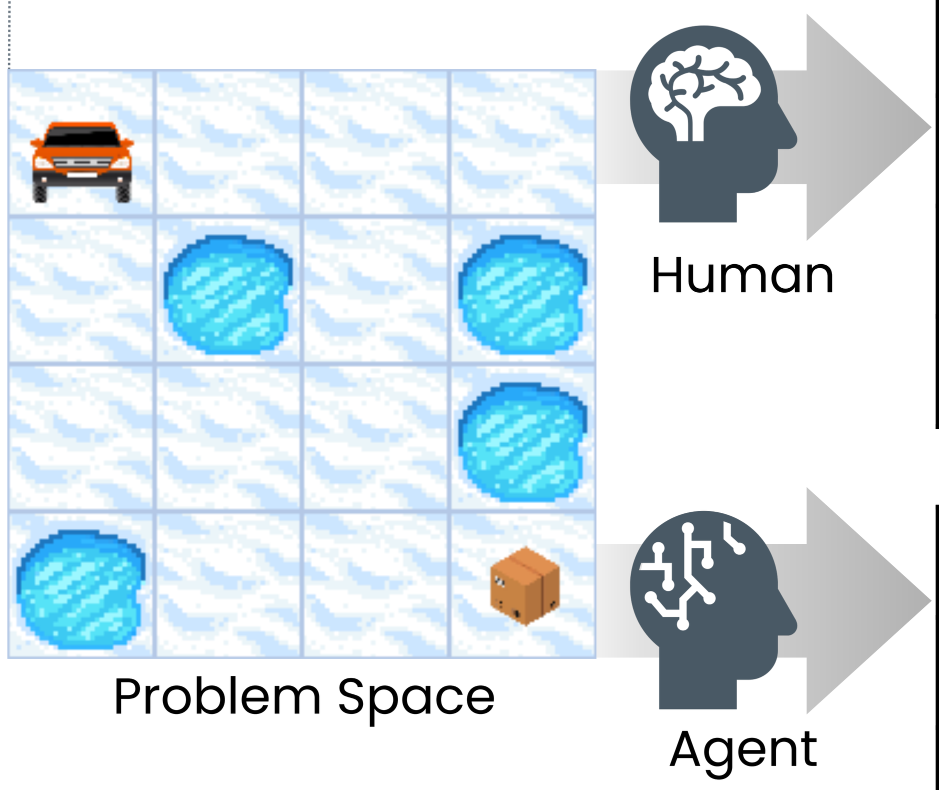
- Trade-off: early insights vs evidence
- By **subjective logic**

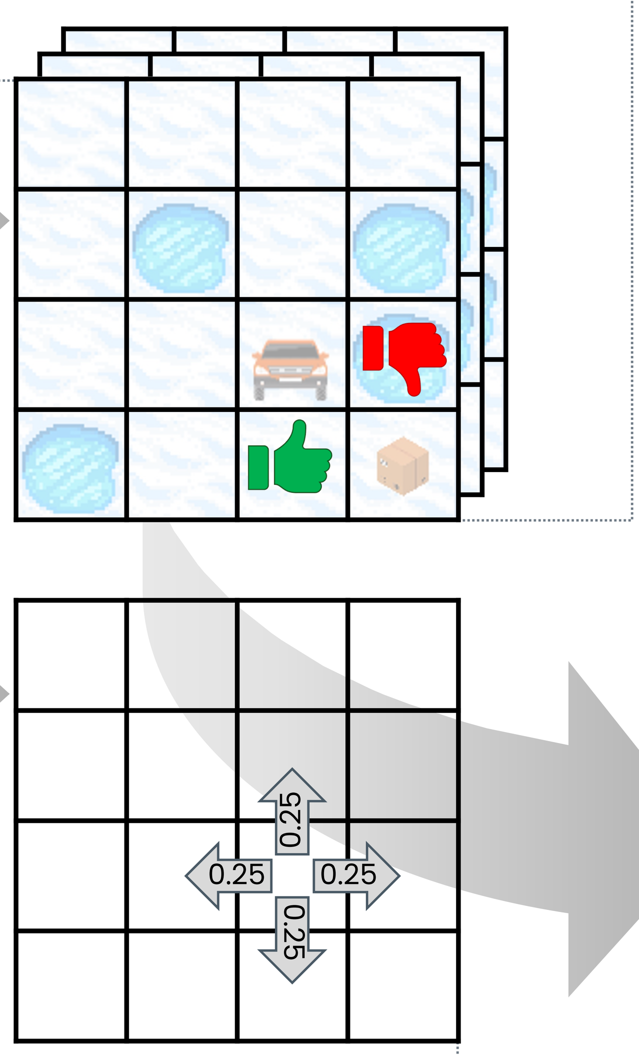
Subjective logic

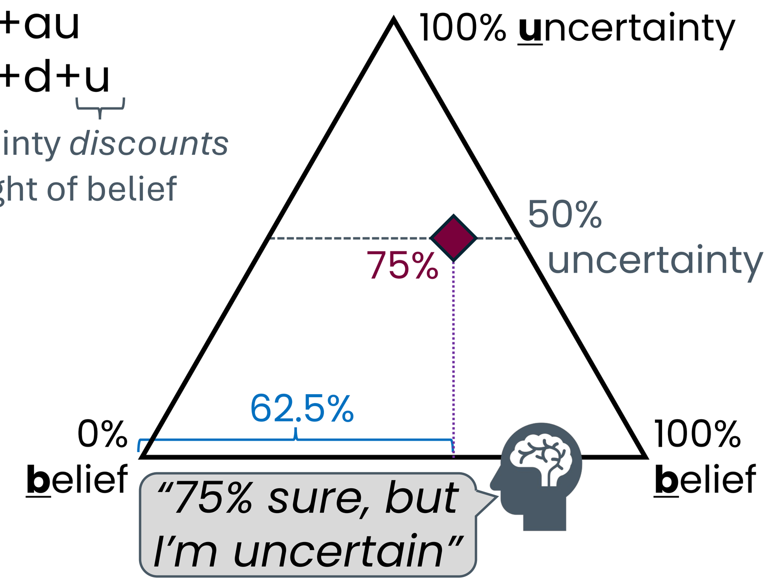
- Probabilistic logic + uncertainty[1]
- $P = b + au$

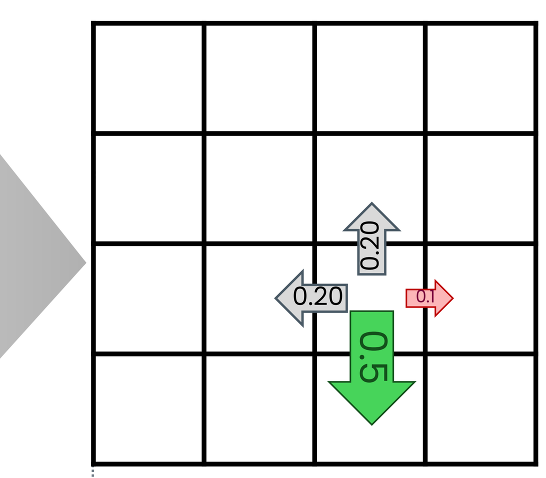
  $1 = b + d + \underbrace{u}$

Uncertainty *discounts* the weight of belief









[1] A. Jøsang. 2016. Subjective Logic. Springer.
[2] K. Dagenais and I. David. 2024. Driving Requirements Evolution by Engineers' Opinions. MPM4CPS'24@MODELS
[3] K. Dagenais and I. David. 2024. Opinion-Guided Reinforcement Learning. Tech. Rep. arXiv:2405.17287 [cs.LG]

Default policy

- Uniform probability distribution of actions

Shaped policy

- Opinions bias the agent's decisions

Mapping between RL and MDE

RL	MDE
Frozen lake	Design space
Agent	Current model state
Action (step)	Model transformation
(0,0) Initial state	Invalid model
Goal state	Valid model
Terminal state	Model beyond repair

Evaluation

- 12x12 version of the Frozen Lake
  - Start→goal: ≥22 steps
  - 20% of states marked terminal
- Agent
  - Algorithm: discrete policy gradient
  - Learning rate: 0.9
  - Discount factor: 1.0
- Learning on 10 000 episodes
- Reward model
  - Terminal state: reward = 0
  - Goal state: reward = +1

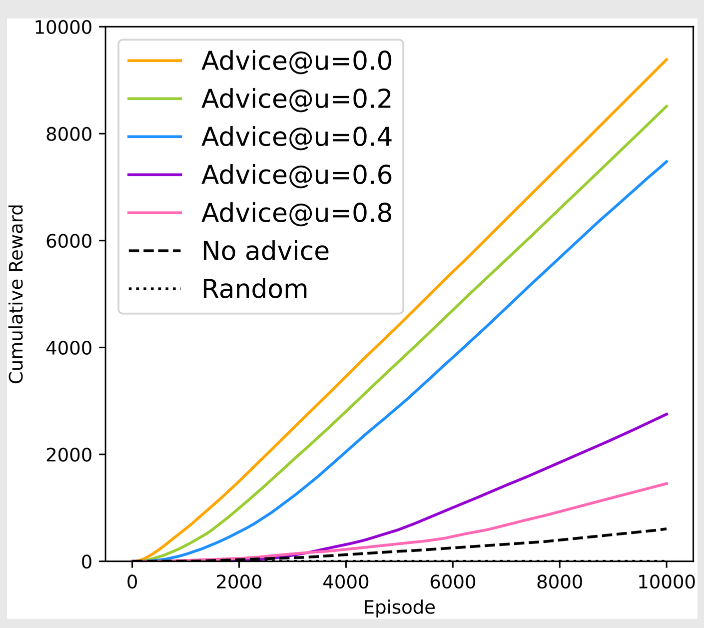
Results and takeaways

Results

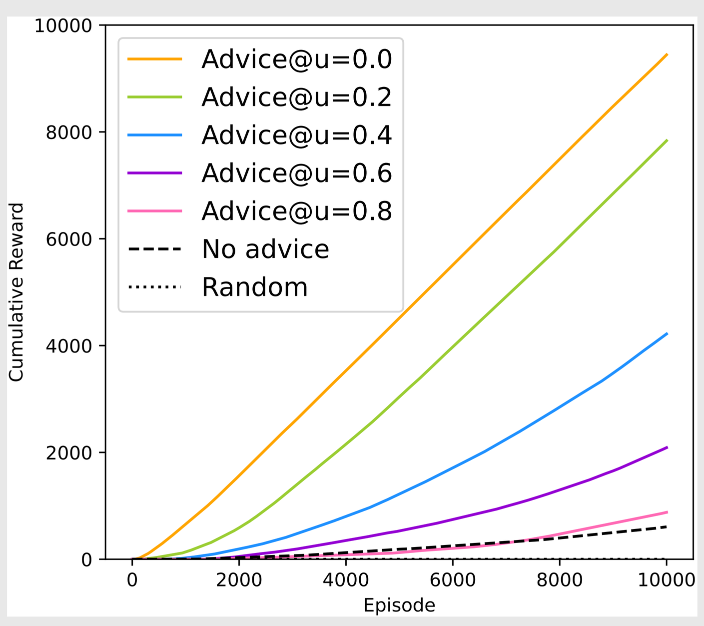
**% denotes opinion quota, i.e., the number of opinions compared to the number of states*

Synthetic Oracle

Idealized setup
- Access to the **whole** problem
- Opinions about **every** state
- Uncertainty is **synthetic**



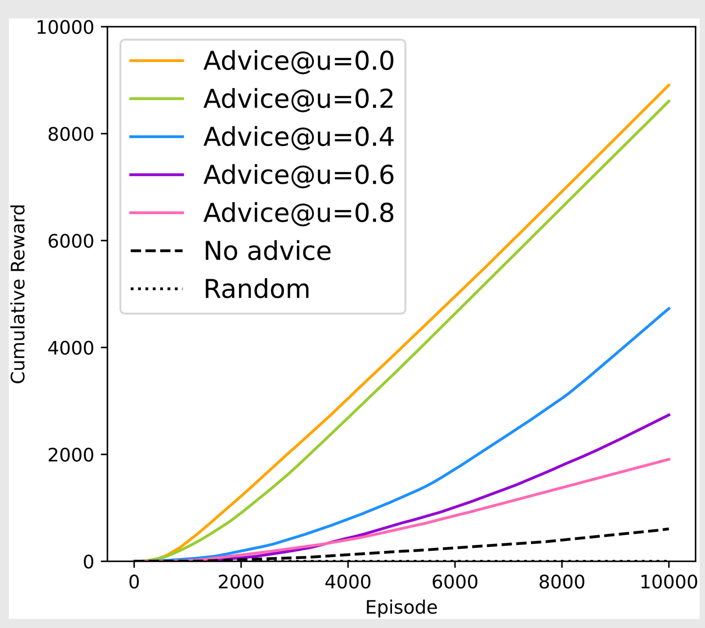
100%



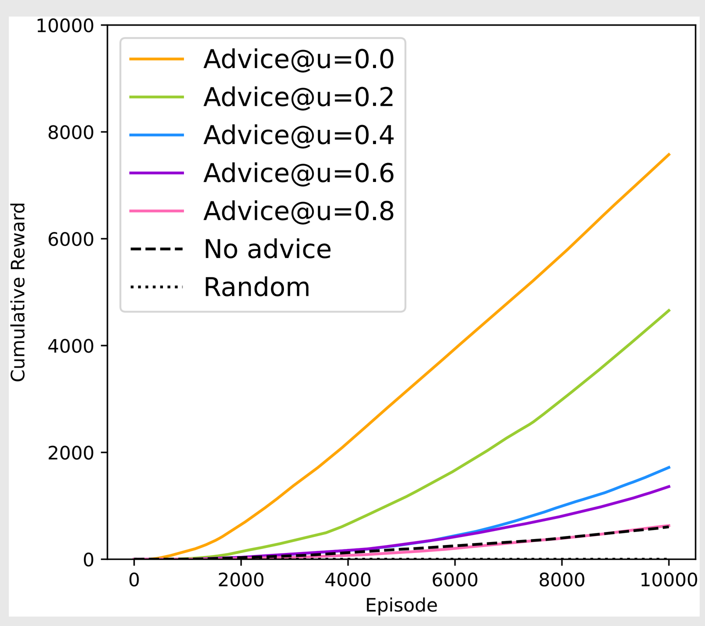
20%

Single Human Advisor

Idealized but humanlike setup
- Access to the **whole** problem
- Opinions about **some** states
- Uncertainty is **synthetic**



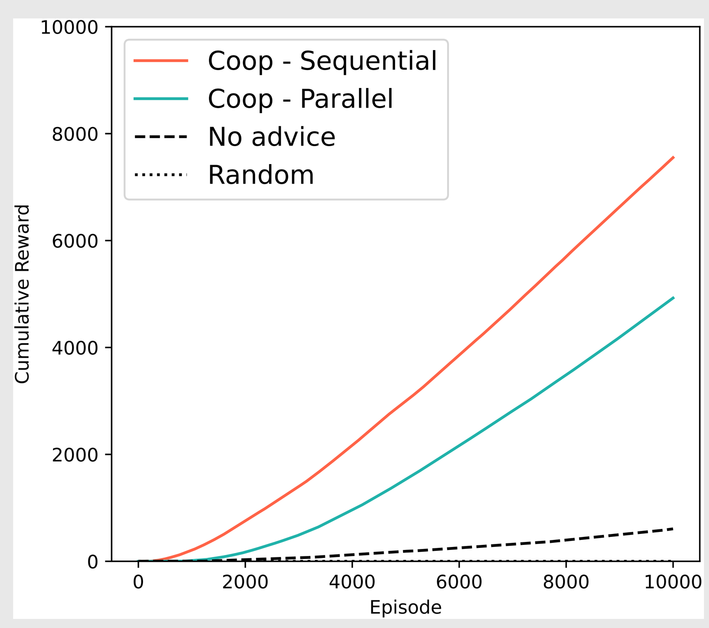
10%



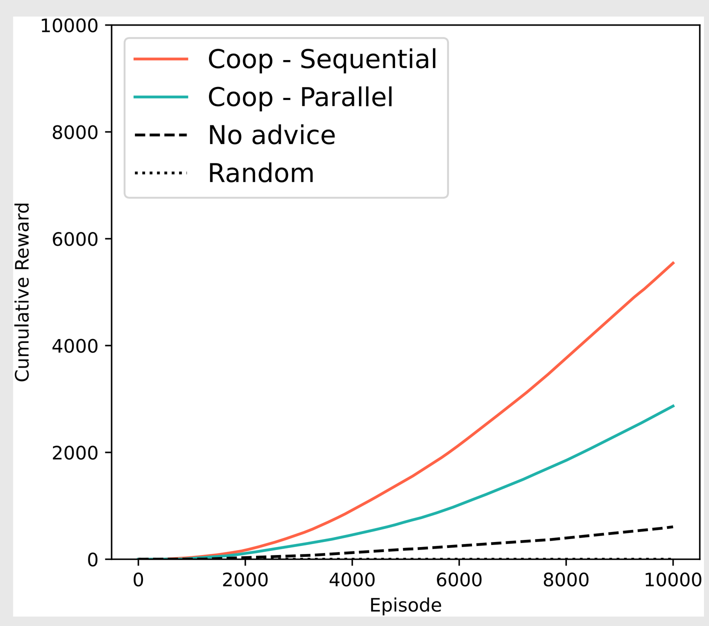
5%

Two Cooperating Humans

Human setup
- Access to **part of** the problem
- Opinions about **some** states
- Uncertainty is **measured**[2, 3]



10%+10%



5%+5%

Takeaways

All opinions, even if uncertain, can be of **high utility**.

In the charts, cumulative reward tends to be significantly higher than *"No advice"* and *"Random"*.

A single human advisor is **as effective** as a synthetic oracle.

In the "Single Human Advisor" charts, cumulative reward tends to be similar to that of the "Synthetic Oracle" charts.

A single human advisor can be **more efficient** than a synthetic oracle.

In the "Single Human Advisor" charts, the same high reward is obtained at lower advice quotas (only 10% and 5% vs 100% and 20%).

Real cooperating humans' performance is **comparable** to that of synthetic advisors.

In the "Two Cooperating Humans" charts, cumulative reward tends to be similar to the other two cases.