

# An agent can learn who to trust when advised by multiple, potentially unreliable experts



UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

## Harnessing the wisdom of an unreliable crowd for autonomous decision making

Tamlin Love, Ritesh Ajoodha, Benjamin Rosman

**Setting:** Contextual bandit, e.g. a medical diagnosis problem where agent can observe patient symptoms and prescribe any combination of available treatments.

**Problem #1:** Sample efficiency. Especially important when dealing with a real-world environment.

**Solution:** Introduce a domain expert (e.g. a doctor) that can tell the agent what to do when training. Typical assumption involves a single, infallible expert.

**Problem #2:** Experts aren't always perfect. They make mistakes, and sometimes are even malicious...

**Problem #3:** What if we have multiple experts? What if they disagree with each other? How do we work out who to trust?

**Solution:** We introduce CLUE (Cautiously Learning with Unreliable Experts), an algorithm that augments bandit decision-making algorithms with the ability to model the reliability of experts and use these models to aggregate the advice from a panel of experts to better inform decision-making when exploring.

**Model:** Reliability  $\rho = 0$  if expert is always suboptimal,  $\rho = 1$  if always optimal. Estimate expected probability of expert offering optimal advice,  $X \approx E(\rho)$ . Denote evaluation with  $x$ . Assess advice using Q function, setting  $x = 1$  if the action maximises Q and 0 otherwise. Using a recency-weighted average controlled by  $\delta$ , we update the model using:

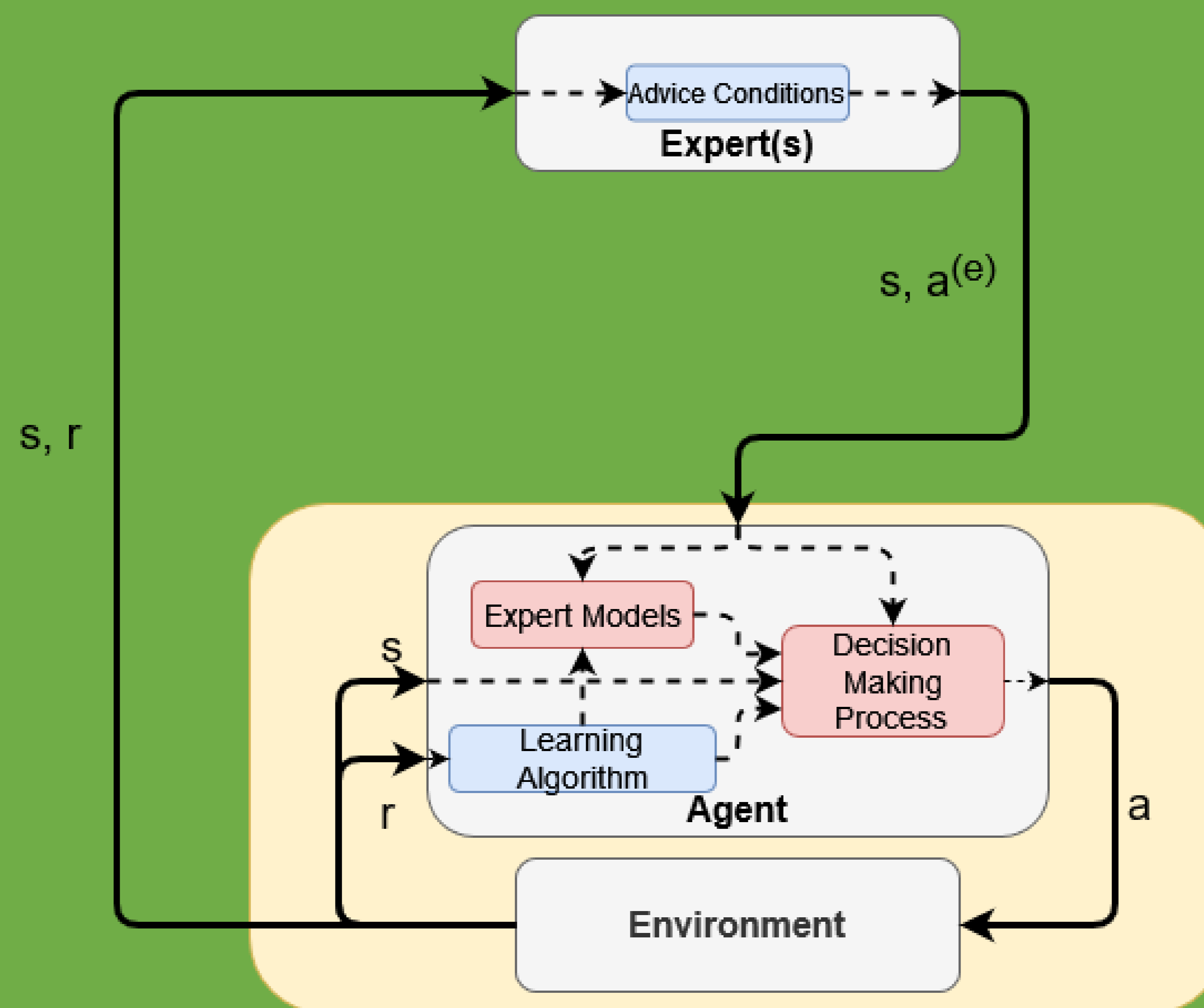
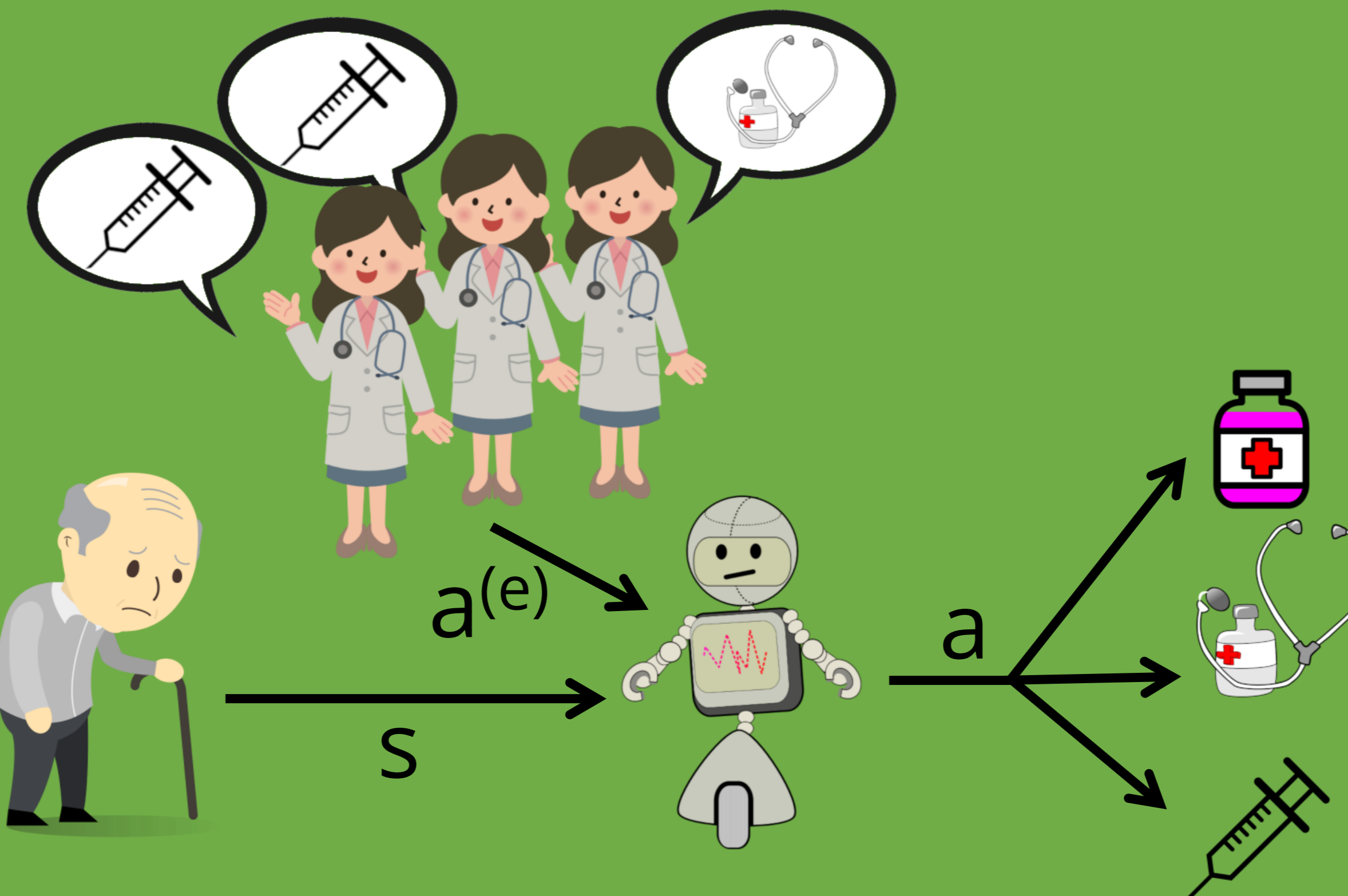
$$X_{t+1} = (1 - \delta)X_t + \delta x_t.$$

**Decision-Making:** Only follow advice when exploring, to allow agent to surpass the experts. Combine all advice received for state  $s_t$  using Bayes rule.  $V_t =$  all advice received,  $v_t^{(e)}$  = advice received from expert  $e$ ,  $E_t =$  set of all experts who advised for  $s_t$ .

$$P(a_j = a^* | V_t) = \frac{\prod_{e \in E_t} P(v_t^{(e)} | a_j = a^*)}{\sum_{k=0}^{|A|} \prod_{e \in E_t} P(v_t^{(e)} | a_k = a^*)}$$

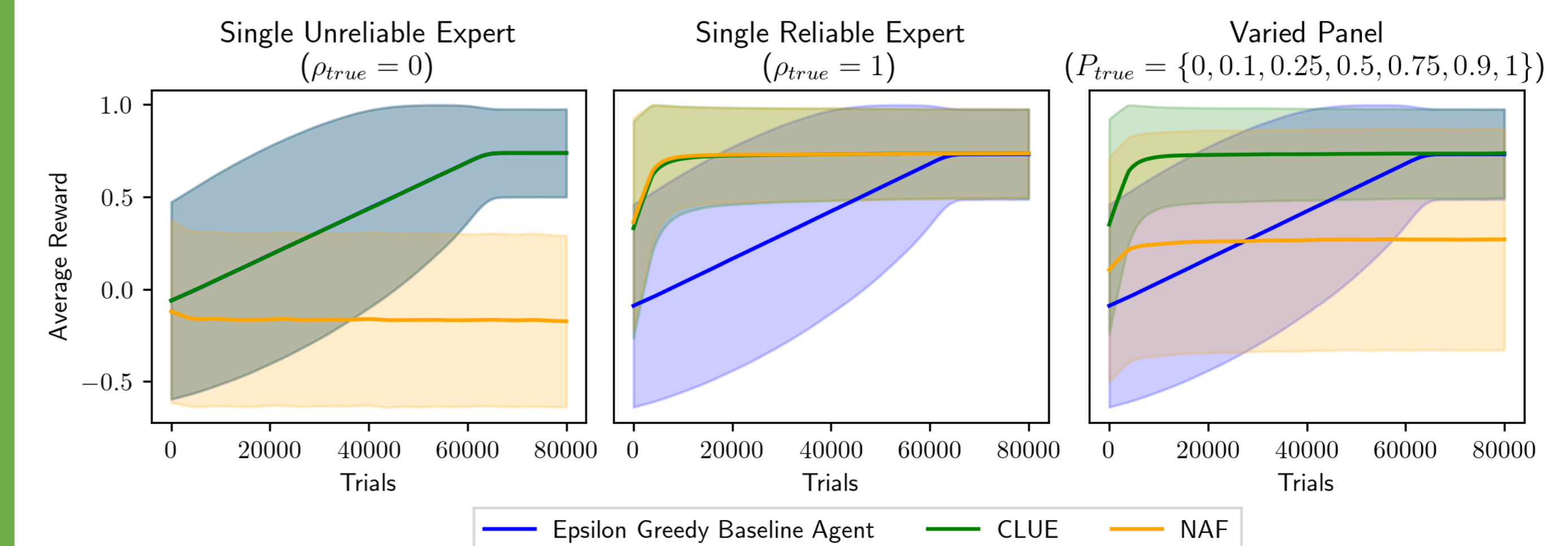
where

$$P(v_t^{(e)} | a = a^*) = \begin{cases} X^{(e)} & \text{if } a = a^{(e)} \\ \frac{1 - X^{(e)}}{|A| - 1} & \text{otherwise} \end{cases}$$

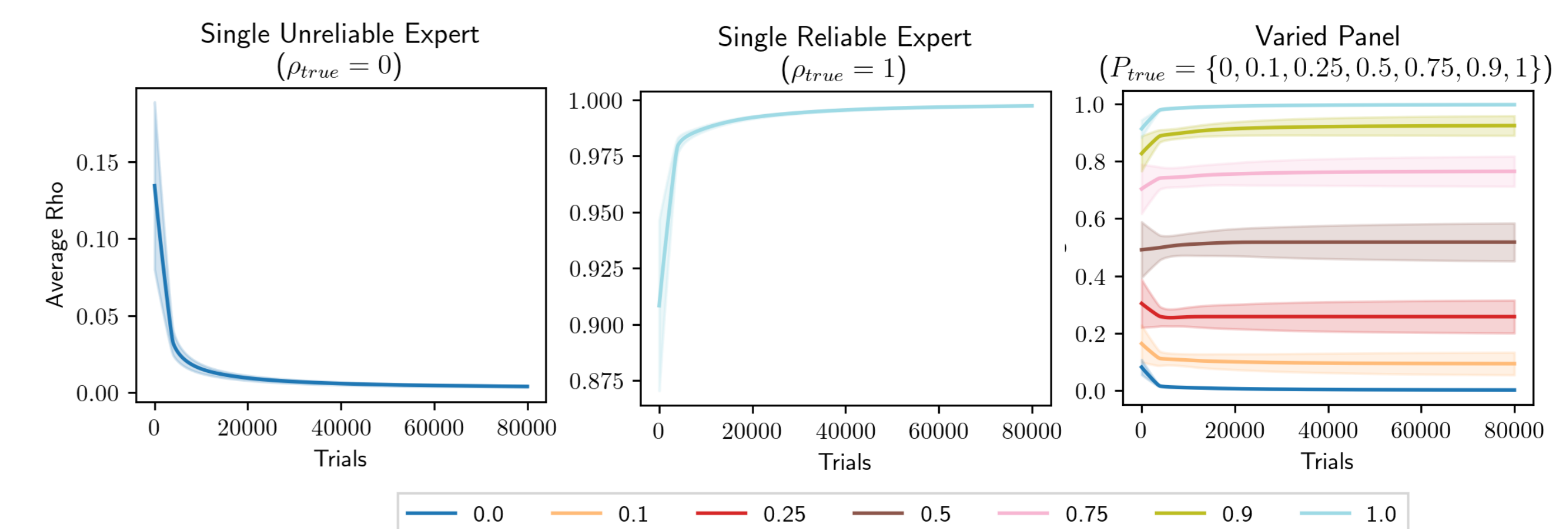


[tamlinlove.github.io](https://github.com/tamlinlove)  
[tamlinlollislove@gmail.com](mailto:tamlinlollislove@gmail.com)

**Results:** Epsilon-Greedy Baseline, 3 panels of experts (1 good, 1 bad, 7 varied). Many random environments. Simulated experts. Compare against baseline and NAF (Naïve Advice Follower), which follows all advice it receives.



Corresponding estimates of reliability:



**Observations:**

- CLUE outperforms baseline when advised by a reliable expert, converging faster
- CLUE is robust to advice from an unreliable expert, defaulting to baseline behaviour
- CLUE can differentiate good experts from bad ones when advised by a mixed panel, and use this to follow good advice while ignoring bad advice

**Conclusion:** CLUE can benefit from increased sample efficiency when advised by a largely reliable expert, but is robust to advice from a largely unreliable expert. CLUE can handle situations with multiple experts, even using their consensus and contradictions to benefit further.

**The Future:** The full RL problem, continuous states/actions, real-world environments (robots!), breaking models into areas of expertise.