BEST-RESPONSE BAYESIAN REINFORCEMENT LEARNING WITH BAYES-ADAPTIVE POMDPS FOR CENTAURS

Mustafa Mert Çelikok[1], Frans A. Oliehoek[2], Samuel Kaski[1,3]
[1] Aalto University, [2] Delft University of Technology, [3] University of Manchester
mustafamert.celikok@aalto.fi

Human-Machine Centaur (HuMaCe)

Definition: Half-human half-AI decision-maker that appears as a single agent to outsiders. The AI's goal is to automate things and help the human half make better decisions. Critical design decisions for a centaur are: **(I) Interaction protocol between human and the AI, (II) AI's assumptions about the human partner, (III) The reward function of the AI.**
Interaction Protocol: A sequential game where, at the environment state s:

1. The AI chooses an action a_m.
2. The human observes a_m, and decides whether to allow a_m to get executed or override it with a_h.
3. If $a_h =$ noop, a_m is executed. If not, a_h gets executed.
4. Both receive the same observation o and base reward r from the environment.
5. **If $a_h \neq$ noop: human pays a *cost of effort* $c_h(s, a_h)$ on top of r. AI also receives an additional term to their reward, $c_m(s, a_m)$.**

Assumptions of the AI:

1. Both the human and the AI have their own subjective models $STM_i = (M_i, OPT_i)$ of the task, where M_i is their subjective POMDP, and OPT_i is their optimality criterion.
2. The disagreements between the AI and the human on what action to take, stems from the differences of their subjective models.
3. The AI models the human using a model space $\mathcal{M}_h = \{m_h \mid m_h = (\bar{\mathcal{A}}_h, \bar{\Omega}_h, \bar{O}_h, \mathcal{I}_h, \bar{\pi}_h, \beta_h, I_h)\}$ where $\bar{\mathcal{A}}_h = \mathcal{A}_h \cup \{noop\}$, $\bar{\pi}_h(\bar{a}_h \mid I_h)$ is the human policy, β_h is the belief update function, $\bar{\Omega}_h$ includes $\mathcal{A}_m$, $\bar{O}_h$ provides full observability of the machine's actions. The human's internal states $I_h \in \mathcal{I}_h$ include the observed action of the machine. Since this model space is very wide, we focus on a particular class of human models called the *Machine-optimistic Human Model* (MoH). Let $Q^{\pi_h^*}(b, a)$ and $V^{\pi_h^*}(b)$ be the value functions π_h^*, the optimal policy for STM_h. Given b_h and a_m, the MoH overrides the machine if and only if $V^{\pi_h^*}(b_h) - Q^{\pi_h^*}(b_h, a_m) > \mathbb{E}_{s\sim b_h}[c_h(s, \pi_h^*(b_h))]$. The subjectively rational choice is to override with $\pi_h^*(b_h)$.
4. Once this model space is given to the AI, the AI's subjective model becomes a Bayesian Best-response model (BA-BRM), BA-BRM$_m(BRM_m, \mu) = (\acute{\mathcal{S}}_m, \mathcal{A}_m, \acute{D}_m, \bar{\Omega}_m, R_m)$, where $\mu(STM; \theta)$ is an appropriately parameterized distribution over the set of possible STM_hs to capture the machine's uncertainty over the MoH's subjective task model. $\acute{\mathcal{S}}_m = \mathcal{S} \times \mathcal{I}_h \times \Theta$ includes the parameter space of μ as Θ. The expected dynamics of the BA-BRM at a given augmented state are defined as:
$\acute{D}_{m,\theta}(\acute{s}'_m, \bar{o}_m \mid \acute{s}_m, a_m) = \sum_{\bar{o}_h} T_m(s' \mid s, a_c)\bar{O}_m(\bar{o}_m \mid \vec{s}'_m, a_c)\beta_h(I'_h \mid I_h, a_c, \bar{o}_h) \int \bar{O}_h(\bar{o}_h \mid \vec{s}'_m, a_c, STM)\bar{\pi}_h(\bar{a}_h \mid I_h, STM)d\mu(STM; \theta)$.

Learning as Planning in HuMaCe: We use a specific adaptation of the *root sampling of the model* variant of the BA-POMCP algorithm to solve the resulting BA-BRM. For partially observable state-spaces, each simulation starts by sampling (s, θ) from the current belief, and in fully-observable settings the current (s, θ) is known. Then, an STM from $\mu(STM; \theta)$ is sampled, and from thereon the simulation proceeds with the STM fixed.

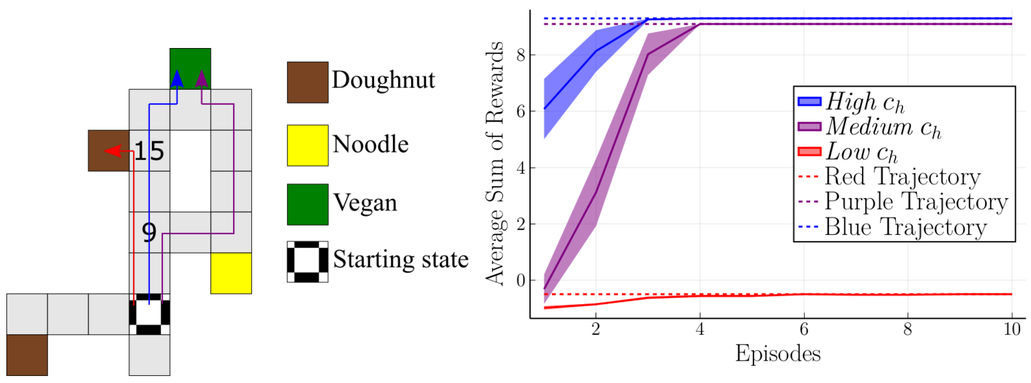
EXPERIMENT 1: Time-inconsistency

Setup:

- Food truck environment. Reward vectors: *doughnut* $= (10, -10)$, *noodle* $= (0, 0)$, *vegan* $= (-10, 20)$. When an agent goes into a restaurant state, it receives the first reward and transitions into a terminal state at the next time-step, where it receives the second.
- Time-inconsistent human behaviour can lead to the red trajectory (getting tempted by nearby doughnuts). Hyperbolic discounting of rewards with a discount factor given by $d(t; \gamma) = \frac{1}{1+t\gamma}$ explains time-inconsistency. We can model this setting as: OPT_h is the hyperbolically discounted sum of rewards, while OPT_m is exponential. The model space of human is parameterized by γ.

Results

- High c_h: Human never overrides. AI infers this quickly after 3 episodes.
- Low c_h: Human overrides at *any* disagreement. AI quickly infers to perform the red trajectory.
- Medium c_h: Purple trajectory is admissible. AI infers this after 4 episodes. Performance is much better than red for everyone.



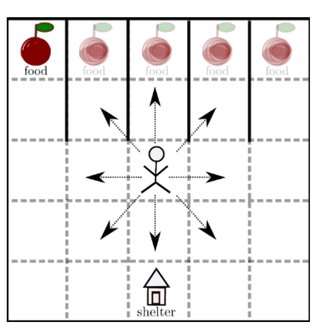
EXPERIMENT 2: Maladaptive Avoidance

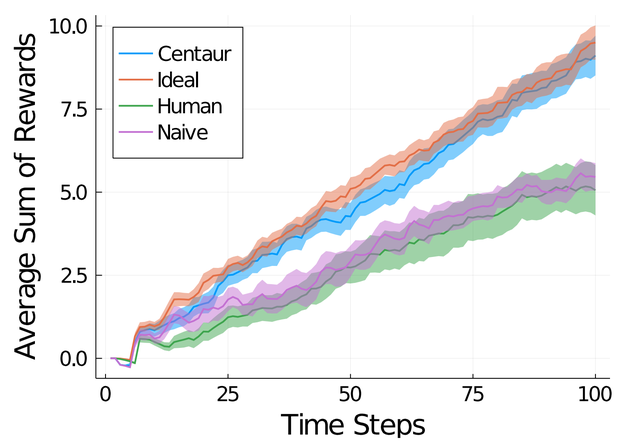
Setup:

- In OTM and STM_m, all actions have noise 0.1, leading to uniform transitions. However, the human believes that the diagonal action noise is $0.1 + 2\epsilon$ and the rest are $0.1 + \epsilon$. So $T = T_m \neq T_h$. Thus, the human model space is parameterized by $\epsilon \in [0, 0.45]$.

Results

- Results with $(\epsilon = 0.45, c_h = 0.05, c_m = 0.2)$. The *centaur* shows the mean return for our method with unknown (ϵ, c_h), *naive* is when the machine executes the policy MCTS gives with STM_m, ignoring the human. *Ideal* is the machine's MCTS solution to known true (ϵ, c_h). The shades represent one standard error.





THEORETICAL RESULT: THE BELIEF ALIGNMENT PROBLEM

Belief Alignment Problem: If $O_m \neq O_h$, when the human and the AI has different beliefs $b_h \neq b_m$, these beliefs may never come closer. Different beliefs will lead to disagreement on optimal action. We present a structural result on when can the beliefs contract depending on the underlying POMDP dynamics and observation probabilities.
Theorem 6.4: For $STM_h = (\mathcal{S}, \mathcal{A}, T, \Omega, R, O_h, OPT)$ and $STM_m = (\mathcal{S}, \mathcal{A}, T, \Omega, R, O_m, OPT)$ with $O_m = O$, let $KL(O_m(. \mid s, a) \mid\mid O_h(. \mid s, a)) \leq \epsilon_O$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Let $\mathbf{T}^{a_c}$ be the transition matrix of the hidden Markov model induced by a fixed centaur policy π_c, with entries $T(s' \mid s, a_c)$ where a_c denotes the executed centaur action. The belief updates satisfy the inequality: $\mathbb{E}_{o\sim O_m(.|b_m,a_c)}[KL(b'_m \mid\mid b'_h)] \leq (1 - \alpha(\mathbf{T}^{a_c}))\, KL(b_m \mid\mid b_h) + \gamma(O_m)3\sqrt{\epsilon_O} - \left(\frac{\gamma(O_m)KL(b_m||b_h)}{\sqrt{2}\log\frac{1}{\mu}}\right)^2$, where $\gamma(O_m) > 0$ is the induced HMM's value of observation, the $\alpha(\mathbf{T}^{a_c})$ is its minimal mixing rate, and μ is a constant such that $b_h(s), b_m(s) \geq \mu$ for all s.
Intuition of Theorem 6.4: Shows that if the dynamics are deterministic and the STM_m is unobservable, the beliefs will not expand. When the STM_m is fully observable, the contraction of beliefs depend on how bad the human's approximate observation model is (i.e. ϵ_O). If the minimal mixing rate cannot be influenced (e.g. deterministic dynamics), the only thing the machine can do to align beliefs is to increase $\gamma(O_m)$.

CONCLUSION

- Designers of collaborative AI agents cannot assume that human users will share the same view of the world with the AI. We formalized a general multiagent Bayesian RL framework for modelling the decision-making of half-human half-AI agents, *centaurs*, and showed that when equipped with an expressive model space for the human behaviour, the AI can learn how to improve a human's decisions, or improve its own decisions with the help of a human.
- Cognitive science can provide us with models useful for learning from human decisions. Sufficient statistics can be drawn from these models and used in multiagent reinforcement learning for assisting humans. Finally, we identified a novel trade-off for partially observable cases which highlights the importance of future work on partial observability for centaurs.
- **Full paper at: arXiv:2204.01160 or DOI: 10.5555/3535850.3535878 (AAMAS 2022)**