June 2021

Best-Response Bayesian Reinforcement Learning with Bayes-adaptive POMDPs for Centaurs

Mustafa Mert CELIKOK [a,1], Frans A. OLIEHOEK [b] and Samuel KASKI [c]
[a] *Aalto University, Finland*
[b] *Delft University of Technology, the Netherlands*
[c] *Aalto University, Finland; University of Manchester, the United Kingdom*

Abstract. Centaurs are half-human, half-AI decision-makers where the AI's goal is to complement the human. To do so, the AI must be able to recognize the goals and constraints of the human and have the means to help them. We present a novel formulation of the interaction between the human and the AI as a sequential game where the agents are modelled using Bayesian best-response models. We show that in this case the AI's problem of helping bounded-rational humans make better decisions reduces to a Bayes-adaptive POMDP. In our simulated experiments, we consider an instantiation of our framework for humans who are subjectively optimistic about the AI's future behaviour. Our results show that when equipped with a model of the human, the AI can infer the human's bounds and nudge them towards better decisions. We discuss ways in which the machine can learn to improve upon its own limitations as well with the help of the human. We identify a novel trade-off for centaurs in partially observable tasks: for the AI's actions to be acceptable to the human, the machine must make sure their beliefs are sufficiently aligned, but aligning beliefs might be costly. We present a preliminary theoretical analysis of this trade-off and its dependence on task structure.

Keywords. Bayesian Reinforcement Learning, Multiagent Learning, Hybrid Intelligence, Computational Rationality

1. Introduction

Humans and AI systems have different computational bounds and biases, and these differences lead to unique strengths and weaknesses. Using this insight, hybrid intelligence aims to combine human and machine intelligence in a complementary way in order to augment the human intellect [1]. From an agent-based perspective, the hybrid intelligence can be seen as a centaur: a part human, part AI decision-maker. Even though essentially a multiagent team, a distinguishing feature of centaurs is that they appear to others as a single entity, two agents acting as one, when acting in an environment.

Cognitive science research has been providing empirically verified computational models of human decision-making that can be used as forward and inverse models in control and reinforcement learning settings [2]. Specifically, the theory of *computational*

[1] Corresponding Author: mustafamert.celikok@aalto.fi

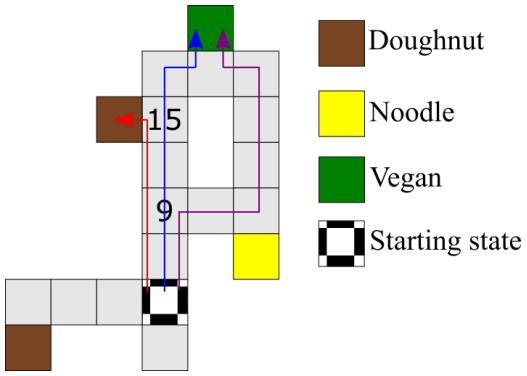

Figure 1. The Food Truck environment has a vegan restaurant, a noodle café, and two identical doughnut shops. When the preferences are *vegan* $\succ$ *doughnut* $\succ$ *noodle*, the blue trajectory is optimal, but the time-inconsistency bias might lead to the red trajectory. In this case, the purple trajectory would help a time-inconsistent human avoid getting tempted by the doughnut shop.

rationality [3,4] focuses on developing models of decision-making under computational bounds, resource constraints, and biases. It argues that agents who seem irrational behave rationally according to their subjective models and constraints. This implies that in a centaur, the AI and the human may disagree on optimal behaviour.

Consider the grid-world in Figure 1, a variant of the *Food Truck* environment of [5] where a human's restaurant preferences are *vegan* $\succ$ *doughnut* $\succ$ *noodle*. The optimal trajectory is the blue line as the shortest path to the most preferred restaurant. However, the human in question follows the red trajectory: they think they can resist the temptation of a doughnut but when the doughnut shop is too close, they fail to do so, and after the dust settles, they regret their decision. Behavioural sciences explain such behaviour in humans as having preferences that are not consistent over a period of time [6,7]. In decision-making with delayed rewards, human time-inconsistency is well-modelled by discounting the rewards with hyperbolic functions of the form $d(t;\gamma) = \frac{1}{1+t\gamma}$ [8,9,10]. This is due to the fact that unlike in the exponential discounting of the form $d(t;\lambda) = \lambda^t$, in hyperbolic discounting the ratio $\frac{d(t;\gamma)}{d(t+k;\gamma)}$ depends on t as well as k which can lead to preference reversals. Now imagine that a time-inconsistent human has recruited an AI to help them eat healthily, and asked the AI to autonomously drive them to the nearest vegan restaurant. Since the typical AI agent has been trained with exponential discounting, it would attempt to follow the blue trajectory, and the human would override the AI at grid 15 by taking control of the car to stop for a doughnut. However, if the AI was able to predict this, it could try to follow the purple trajectory instead by attempting a detour at grid 9. The purple trajectory costs two time-steps extra, but the human may allow this detour if they decide saving two time-steps is not worth overriding the autonomous driver. In this paper, we formalize these intuitions by developing a decision-theoretic multiagent model for centaurs.

2. The Human–Machine Centaur (HuMaCe) Model

The interaction between the human (h) and the machine (m) is modelled as a sequential game. At time t, first the machine chooses an action a_m, and the human observes this choice. Then, the human either lets a_m get executed by playing a special *no operation* action $a_h = noop$ or overrides it with another action $a_h \neq noop$. If the human overrides, they pay an additional cost, $c_h(s, a_h)$, which represents the human's internal incentive to automate the task and delegate things to the machine. When overridden, the machine also receives an additional cost (or reward), $c_m(s, a_h)$, that determines its incentives about getting overridden. The specification of c_m is part of the machine's design and offers significant flexibility. For instance in an autonomous car it makes sense to avoid forcing the human to take control, thus c_m might penalize overrides just like c_h. In other cases such as autonomous flight, we may want to incentivize the machine with c_m to safely trigger an override if the human operator is losing attention. The underlying task performed by the centaur is single-agent: either the human or the machine execute an action in the real environment. The executed action, called the *centaur action*, is defined as a function $a_c(a_m, a_h) = \mathbb{I}[a_h \neq noop]a_h + \mathbb{I}[a_h = noop]a_m$, and it is observed by both agents. This leads to specific structure where the underlying transition dynamics and rewards are essentially single-agent as $T_i(s' \mid s, a_c(a_m, a_h))$ and $R_i(s, a_c(a_m, a_h))$ for $i \in \{h, m\}$. In the end, the human's decision to override or not influences the transition dynamics the centaur experiences. Therefore, given a parametric model space for the human's behaviour, the machine's problem of inferring the model of the human reduces to learning this new transition dynamics. Once the machine learns the human-influenced dynamics, it can predict whether it will get overridden or not, and also perform an optimal policy in terms of helping the human. This learning problem is modelled as a Bayes-adaptive best-response model (BRM), which is detailed in our paper [11]. In the end, the Bayesian BRMs are a special class of Bayes-adaptive POMDPs. We apply a novel adaptation of a Monte Carlo planning algorithm which maintains a set of plausible human models as particles.

2.1. Time-Inconsistent Preferences

In the introduction, we have given an example for time-inconsistent behaviour: the red trajectory in Figure 1. We simulate the interaction in this setting with six behavioural classes of the human based on (c_h, γ) pairs, where γ represents how time-inconsistent the human is. When γ is high (≥ 2.0), the human behaviour leads to the time-inconsistent red trajectory, and if low (≤ 0.5) the human and the machine agree, producing the blue trajectory. If the c_h is too low (≤ 0.2) the AI is overridden whenever the human disagrees, and if high enough (≥ 0.4) the AI can execute the blue trajectory regardless of γ. For medium c_h ($[0.21, 0.4]$) and high γ, the AI cannot execute the blue, but can get the purple accepted. Figure 2 shows the machine's undiscounted sum of rewards for the three cases when γ is high. The dashed lines show the same for trajectories shown in Figure 1. When c_h is low, the centaur learns that following the red trajectory is the only option, whereas if the c_h is high, it learns to follow the blue. In the case of medium c_h, the blue trajectory is not admissible, but the centaur learns that the purple is. Thus, the machine improves the human's utility drastically. In experiments with low γ, the centaur follows the blue trajectory since the machine and the human fully agree, and the machine never gets overridden.

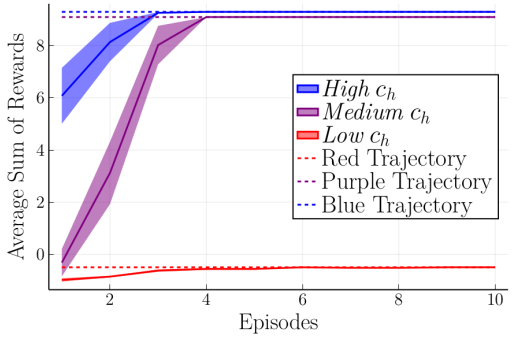

Figure 2. The machine's return in the Food Truck environment with a time-inconsistent human ($\gamma = 7.5$). Low c_h (0.01): the human always overrides, so the machine cannot help avoid the red trajectory. Medium c_h (0.21): machine follows the purple trajectory and increases the human's performance drastically. Dashed lines show the undiscounted return of the three trajectories from Figure 1 without any overrides. The shades represent one standard error.

2.2. Belief alignment problem

Our theoretical analysis revealed that, in partially observable environments, differences between the way the machine and the human model the world can lead to a novel issue: the *belief alignment problem*. In short, if the human and the machine have different observation or transition models, they will interpret the same observation differently, and end up with different beliefs. Difference in beliefs can lead to disagreements on what action to take, crippling the centaur permanently. We have derived a contraction bound, describing in which cases the machine and the human's beliefs can come closer over time, even if their world models are different. This result indicates that whether beliefs can be aligned depends heavily on the structure of the task and the incentives of the human.

3. 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 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. Our formulation can capture various human factors in automation, such as automation misuse where the human is over-reliant on the AI (very high c_h) or automation disuse (very low c_h) [12]. 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. An adaptive human can also learn and improve their model of the world through interaction, which opens further theoretical questions such as how the learning rate of the human and the machine affect the belief contraction.

References

[1] Akata Z, Balliet D, de Rijke M, Dignum F, Dignum V, Eiben G, et al. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. Computer. 2020;53(8):18-28.

[2] Ho MK, Griffiths TL. Cognitive science as a source of forward and inverse models of human decisions for robotics and control. CoRR. 2021;abs/2109.00127.

[3] Lewis RL, Howes A, Singh S. Computational Rationality: Linking Mechanism and Behavior Through Bounded Utility Maximization. Topics in Cognitive Science. 2014;6(2):279-311.

[4] Gershman SJ, Horvitz EJ, Tenenbaum JB. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. Science. 2015;349(6245):273-8.

[5] Evans O, Stuhlmüller A, Goodman ND. Learning the Preferences of Ignorant, Inconsistent Agents. In: Schuurmans D, Wellman MP, editors. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA. AAAI Press; 2016. p. 323-9.

[6] Loewenstein G, Prelec D. Anomalies in intertemporal choice: Evidence and an interpretation. The Quarterly Journal of Economics. 1992;107(2):573-97.

[7] Frederick S, Loewenstein G, O'Donoghue T. Time discounting and time preference: A critical review. Journal of Economic Literature. 2002;40(2):351-401.

[8] Green L, Fristoe N, Myerson J. Temporal discounting and preference reversals in choice between delayed outcomes. Psychonomic Bulletin & Review. 1994;1(3):383-9.

[9] Kirby KN, Herrnstein RJ. Preference reversals due to myopic discounting of delayed reward. Psychological Science. 1995;6(2):83-9.

[10] Kleinberg JM, Oren S. Time-inconsistent planning: a computational problem in behavioral economics. In: Babaioff M, Conitzer V, Easley DA, editors. ACM Conference on Economics and Computation, EC '14, Stanford , CA, USA, June 8-12, 2014. ACM; 2014. p. 547-64.

[11] Çelikok MM, Oliehoek FA, Kaski S. Best-Response Bayesian Reinforcement Learning with BA-POMDPs for Centaurs. In: Proceedings of the Twenty-First International Conference on Autonomous Agents and Multiagent Systems (AAMAS); 2022. .

[12] Parasuraman R, Riley V. Humans and automation: Use, misuse, disuse, abuse. Human factors. 1997;39(2):230-53.