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

Mustafa Mert Çelikok
Aalto University
Finland
mustafamert.celikok@aalto.fi

Frans A. Oliehoek
Delft University of Technology
The Netherlands
frans.oliehoek@tudelft.nl

Samuel Kaski
Aalto University
Finland
University of Manchester
The United Kingdom
samuel.kaski@aalto.fi

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 [2]. 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
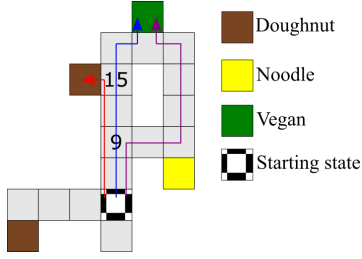
a single entity, two agents acting as one, when acting in an environment. In this work, we introduce two important design factors for centaurs: I) The AI must be able to learn a model of the human's decision-making, and II) The interaction protocol between the AI and the human must never prohibit an option for the human.

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 [21]. Specifically, the theory of *computational rationality* [14, 28] 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 Evans et al. [8] where a human's restaurant preferences are *vegan > doughnut > 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 [13, 29]. In decision-making with delayed rewards, human time-inconsistency is well-modelled by discounting the rewards with hyperbolic functions of the form $d(t; \gamma) = \frac{1}{1+t\gamma}$ [17, 23, 25]. 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. Our main contributions from the most general to specific are: I) We formulate the interaction protocol between the human and the AI as a sequential game. First, the AI proposes to execute an action, and then the

Figure 1: The Food Truck environment has a vegan restaurant, a noodle café, and two identical doughnut shops. When the preferences are *vegan > doughnut > 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.**

human responds by either allowing or overriding it. The human's internal incentive to delegate tasks to the AI is modelled as a cost the human pays when they override an action. We model the agents in a centaur with Bayesian best-response models (defined in Section 3), where learning about the human and nudging them is cast into a Bayesian reinforcement learning problem. (II) We identify a novel challenge in centaurs for partially observable cases, called the *belief alignment* problem, and analyse some of its theoretical properties. (III) As an instance of our framework, we model a class of humans who are optimistic about the machine and compute their best-response approximately. Adopting computational models of human bounded-rationality provided by behavioural economics [13, 25, 29] and cognitive science [17, 20, 23, 40], we simulate two cases of human bounded-rationality: time-inconsistent preferences and over (or under) estimation of probabilities. Our experiments show that if a human's incentives allow it, the AI is able to nudge the human into making better decisions. The mathematical and computational framework we present opens up a new research direction where various hybrid intelligence problems can be cast into centaurs and more complex model spaces for humans can be investigated in the future.

2 BACKGROUND

In this section we give a brief introduction to Bayesian reinforcement learning (BRL) for partially observable Markov decision processes (POMDPs), and Bayesian best response models (BA-BRM).

POMDPs and Bayesian Reinforcement Learning. We define a discrete POMDP by a 6-tuple $(\mathcal{S}, \mathcal{A}, T, R, \Omega, O)$ where $(\mathcal{S}, \mathcal{A}, T, R)$ are the sets of states and actions, transition dynamics, and the reward function of a Markov decision process (MDP); Ω is a set of observations, and $O : \Omega \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the observation function that defines the conditional observation probabilities which can be represented as a set of $|\mathcal{S}| \times |\Omega|$ matrices $\{O^a\}_{\forall a \in \mathcal{A}}$ with the i-th row, j-th column giving the observation probability $O^a_{(ij)} = Pr(o_j \mid s_i, a)$. The Bayes-adaptive POMDP (BA-POMDP) is a BRL model where the unknown transition and observation probabilities are modelled with Dirichlet distributions, and the $\mathcal{S}$ is augmented to include the

Dirichlet parameters which transition according to the Bayes rule [35]. The solution to the BA-POMDP is the Bayes-optimal policy in terms of the exploration/exploitation trade-off, with respect to the prior.

Bayesian Best Response Models. In the same vein as interactive POMDPs (I-POMDPs) [16] and the recursive modelling method [15], Bayesian best-response models (BRM) [31] fall under the *subjective perspective* approach to multi-agent systems, where the systems are modelled from the point of view of a protagonist agent [32]. In Bayesian BRMs, the protagonist (denoted by i) is considered in its subjective view of the multi-agent environment defined as $MAE_i \triangleq (\mathcal{S}, \{\mathcal{A}_i, \mathcal{A}_j\}, T, \{\Omega_i, \Omega_j\}, O, R_i)$ where j denotes the antagonist, $\mathcal{A}_j$ and Ω_j are the sets of actions and observations of j, and R_i is the protagonist's reward function. The transition dynamics and observation function are defined for the joint set of actions $\mathcal{A} = \mathcal{A}_i \times \mathcal{A}_j$ and the observations $\Omega = \Omega_i \times \Omega_j$. To compute a best-response, the protagonist also needs a model of j as described below, and an *optimality criterion* OPT_i that defines how i accumulates reward over time, such as the exponentially discounted sum of rewards.

A *model* of an agent j is a tuple $m_j = (\mathcal{A}_j, \Omega_j, \mathcal{I}_j, \pi_j, \beta_j, I_j)$ that is consistent with the MAE_i and fully specifies how the agent will respond to future observations. The $\mathcal{I}_j$ denotes the finite set of internal states of j, I_j is its current internal state, $\pi_j : \mathcal{I}_j \to \Delta(\mathcal{A}_j)$ is a policy mapping the internal states to action probabilities, and $\beta_j : \mathcal{I}_j \times \mathcal{A}_j \times \Omega_j \to \Delta(\mathcal{I}_j)$ is a belief update function describing how j maintains beliefs over $\mathcal{I}_j$. In a BRM, the agent i is uncertain about some elements of m_j such as I_j or π_j, however it has a set of models $\mathcal{M}_j$ that are consistent with the known elements of it. In the end, the best-response model combines these three elements, denoted with $BRM_i = (MAE_i, \mathcal{M}_j, OPT_i)$, where the state-space gets augmented to $\bar{\mathcal{S}} = \mathcal{S} \times \mathcal{I}_j$. The dynamics of the BRM_i is defined as:

$$D_i(\bar{s}'_i, o_i \mid \bar{s}_i, a_i) = \sum_{a_j} \sum_{o_j} O(o_j, o_i \mid s', a_i, a_j) T(s' \mid s, a_i, a_j)$$
$$\beta_j(I'_j \mid I_j, a_j, o_j)\pi_j(a_j \mid I_j). \quad (1)$$

Oliehoek and Amato [31] have shown that BRMs are a class of POMDPs which, when solved, provides i's best response to j. Therefore, if the T and O are unknown to i, they can be modelled with Dirichlet distributions as in BA-POMDPs, which leads to Bayes-adaptive best-response models (BA-BRMs). The BA-BRM of i is defined as $BA\text{-}BRM_i = (\bar{\mathcal{S}}, \mathcal{A}_i, \Omega_i, \bar{D}_i, \bar{R}_i)$ where $\bar{\mathcal{S}} = \mathcal{S} \times \mathcal{I}_j \times \Theta$ is the augmented set of states with Θ denoting the parameter space of Dirichlet distributions; $\bar{R}_i(\bar{s}, a_i) = \sum_{a_j} R_i(s, a_i, a_j)\pi_j(a_j \mid I_j)$ is the augmented reward function; and $\bar{D}_i$ is the joint transition and observation dynamics defined as:

$$\bar{D}_\theta(\bar{s}', o_i \mid \bar{s}, a_i) \triangleq \frac{\theta^{\bar{s}',o_i}_{\bar{s},a_i}}{\sum_{\bar{s}',o_i} \theta^{\bar{s}',o_i}_{\bar{s},a_i}}$$
$$\bar{D}_i(\bar{s}', o_i \mid \bar{s}, a_i) = \bar{D}_\theta(\bar{s}', o_i \mid \bar{s}, a_i)\mathbb{I}[\theta' = \theta + \delta^{\bar{s}',o_i}_{\bar{s},a_i}]. \quad (2)$$

Here $\theta^{\bar{s}',o_i}_{\bar{s},a_i}$ is the count for $(\bar{s}', o_i, \bar{s}, a_i)$, $\delta^{\bar{s}',o_i}_{\bar{s},a_i}$ is the one-hot encoding of $(\bar{s}', o_i, \bar{s}, a_i)$, and $\mathbb{I}$ is the indicator function. Intuitively, $BA\text{-}BRM_i$ models the i's uncertainty about j as part of its uncertainty about the dynamics.

3 THE HUMAN–MACHINE CENTAUR (HUMACE) MODEL

In this section we formalize our general framework for human–machine centaurs (HuMaCe), and then present empirical and theoretical results on an instantiation of our framework. The HuMaCe model represents the human and the machine's behaviour as BRMs with different subjective models of the same task, which will be defined explicitly. BRMs inherently assume some model about the other agent, and if that other agent itself employs a BRM this leads to a recursion of beliefs, as in I-POMDPs. Here, we cut this short by considering the BRMs from both agents in isolation, assuming that a sufficiently useful model of the other is available.

The interaction protocol. 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 switching-controller property of the protocol leads to MAEs with 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 $i \in \{h, m\}$.

The objective and subjective task models. As described in the interaction protocol, the underlying dynamics of the agents' multiagent environments is single-agent. Therefore, we represent an agent's sequential decision-making task with a POMDP $M = (\mathcal{S}, \mathcal{A}, T, R, \Omega, O)$ and an *objective optimality criterion OPT*. The *objective task model* is defined as $OTM \triangleq (M, OPT)$. When solved exactly, the OTM provides the best behaviour possible (e.g. the blue trajectory in Figure 1), and represents the ideal problem definition a rational agent can hope for. However, in general the OTM is unknown even to the designer of the machine. Therefore, the human and the machine each have their own subjective model of the task, STM_i, which consists of their subjective POMDP M_i and optimality criterion OPT_i for $i \in \{h, m\}$. The STMs represent the agents' subjective surrogates that approximate the OTM, and since they can be different, they may lead to disagreements on what is optimal behaviour. For instance, in our introductory example the disagreement is due to differences in OPT_i: OPT_h is the hyperbolically discounted sum of rewards whereas OPT_m and OPT are exponentially discounted.

Since the dynamics of MAEs are fully determined by STMs here, an STM and a model of the human suffice to derive the best-response model of the machine.

3.1 Best Response Model of the Machine

In this section, we define the machine's model for computing its best response to the human. The model space of the other agent in BRMs is quite general, since it can represent a wide range of agent types, including POMDP-based agents and any policy that can be represented as a finite state controller.

The machine's BRM. We define the model space of the human as $\mathcal{M}_h = \{m_h \mid m_h = (\bar{\mathcal{A}}_h, \bar{\Omega}_h, \bar{O}_h, \mathcal{I}_h, \bar{\pi}_h, \beta_h, I_h)\}$ where $\bar{\mathcal{A}}_h = \mathcal{A}_h \cup \{noop\}$; the augmented observation space $\bar{\Omega}_h$ includes $\mathcal{A}_m$, and the $\bar{O}_h$ 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. With the $\mathcal{M}_h$, the machine's BRM is defined as $BRM_m(STM_m, \mathcal{M}_h) = (\bar{S}_m, \mathcal{A}_m, \bar{D}_m, \bar{\Omega}_m, \bar{R}_m)$, where $\bar{S}_m = S \times I_h$, $\bar{\Omega}_m = \Omega_m \times \bar{\mathcal{A}}_h$, and $\bar{R}_m(\bar{s}_m, a_m) = R_m(s, a_c(a_m, \bar{a}_h)) - \mathbb{I}[\bar{a}_h \neq noop]c_m(s, \bar{a}_h)$. The $\bar{\Omega}_m$ captures the fact that the machine can observe the human's override, but only after it had happened. The $\bar{D}_m$ is defined as follows, where the a_c overloads $a_c = a_c(a_m, \bar{a}_h)$:

$$\bar{D}_m(\bar{s}'_m, \bar{o}_m \mid \bar{s}_m, a_m) = \sum_{\bar{a}_h} \sum_{\bar{o}_h} \bar{O}(\bar{o}_h, \bar{o}_m \mid \bar{s}'_m, a_c)T_m(s' \mid s, a_c)$$
$$\beta_h(I'_h \mid I_h, a_c, \bar{o}_h)\bar{\pi}_h(\bar{a}_h \mid I_h). \quad (3)$$

The form of $\bar{D}_m$ captures the fact that once the machine chooses a_m, $\bar{\pi}_h$ determines the centaur action a_c by overriding or not. Then the dynamics, observations, and the human's belief update evolve according to the a_c. The machine cannot observe $\bar{a}_h$ before it chooses a_m, and must predict that using $\bar{\pi}_h$ and I_h. If these two are unknown, then the machine's BRM is augmented to a BA-BRM by placing a prior over them, as described in Section 2.

In practice, the machine will need a more concrete model space for the human in order to infer the human's model and assist them. Next, we define a bounded-rational model space for humans as an instantiation of our framework.

3.2 The Machine-Optimistic Human

It is not realistic to assume the human will have a perfect model of the machine and will override it optimally with respect to the STM_h, due to cognitive bounds. Optimism is one of the most prevalent and ubiquitous approximations humans use to make bounded-rational decisions under cognitive bounds [19, 27, 34, 37]. We apply this approximation to our setting and propose the *machine-optimistic human (MoH) model*, where the human assumes that the machine's behaviour will agree with them from $t + 1$ onward, so after t they will not have to override. This model is one of many possible instantiations of our framework, and future work will focus on enriching the space of human models under the guidance of cognitive sciences as proposed in Ho and Griffiths [21].

The following proposition states that a MoH compares the value of the machine's action a_m to the action they think is optimal, using the optimal value function of their STM_h: the $Q^{\pi_h^*}$.

PROPOSITION 3.1. *Let $Q^{\pi_h^*}(b, a)$ and $V^{\pi_h^*}(b)$ be the value functions of the subjectively optimal policy π_h^* computed according to the OPT_h*

for M_h. Given their current belief b_h and the machine's action a_m, the MoH overrides the machine if and only if $V^{\pi_h^*}(b_h) - Q^{\pi_h^*}(b_h, a_m) > \mathbb{E}_{s \sim b_h}[c_h(s, \pi_h^*(b_h))]$, and if they decide to override a_m, the subjectively rational choice is to override with $\pi_h^*(b_h)$.

In Supplementary Section 1.1, we prove that Proposition 3.1 follows directly if the human is solving a BRM of their own to decide when to override, using their optimal policy for STM_h, the π_h^*, for predicting how the machine will behave in the future. This result provides a decision-theoretic grounding for the MoH model.

For the machine, Proposition 3.1 implies that the $Q^{\pi_h^*}$, b_h, and c_h are sufficient to predict the MoH's response to a_m, where b_h is the MoH's belief in M_h over $\mathcal{S}$. Next, we define the model space for MoH and the machine's BA-BRM for MoH models using this result.

The space of machine-optimistic human models. A MoH's internal state is defined as $I_h = (b_h, z, a_m)$, where b_h is its belief state in M_h and a_m is the machine's action. The z combines $Q^{\pi_h^*}$ and c_h to a binary function indicating if the MoH will override as: $z(a_m) = \mathbb{I}[V^{\pi_h^*}(b_h) - Q^{\pi_h^*}(b_h, a_m) > \mathbb{E}_{b_h}[c_h(s, \pi_h^*(b_h))]]$. The definition of $\bar{\pi}_h$ follows directly from Proposition 3.1 as $\bar{\pi}_h(\bar{a}_h \mid I_h) = z(a_m)\pi_h^*(b_h) + (1 - z(a_m))\delta(\bar{a}_h, noop)$ where δ is the Kronecker delta function. The π_h^* is an optimal policy to a POMDP and assumed to be deterministic without loss of generality, and thus $\bar{\pi}_h$ is also deterministic.

The machine's BA-BRM for machine-optimistic humans. In our setting the machine's uncertainty over $\mathcal{M}_h$ is due to unknowns $\bar{\pi}_h$ and I_h. However, the human affects the machine's dynamics only through a_c and therefore the machine's uncertainty over $\mathcal{M}_h$ still manifests itself as uncertainty over the dynamics. The most general Bayes-adaptive approach would be to place a Dirichlet prior on the parameters of the categorical distribution $\bar{D}_m$ as in BA-POMDPs. However, this would ignore the structure in $\bar{D}_m$: if the machine had known $STM_h = (M_h, OPT_h)$ and the c_h, it could compute $Q^{\pi_h^*}$ which is enough to fully determine $\bar{D}_m$ for MoHs. Therefore, in this case it is more sample-efficient to maintain a posterior over STM_h. We will choose an appropriately parameterized distribution $\mu(STM; \theta)$ over the set of possible STMs to capture the machine's uncertainty over the MoH's subjective task model. Together with μ, the machine's BA-BRM can be defined as:

$$BA\text{-}BRM_m(BRM_m, \mu) = (\acute{S}_m, \mathcal{A}_m, \acute{D}_m, \bar{\Omega}_m, R_m) \qquad (4)$$

where the augmented state space $\acute{S}_m = \mathcal{S} \times I_h \times \Theta$ includes the parameter space of μ as Θ.
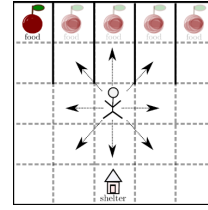
The definition of $\acute{D}_{m,\theta}(\acute{s}'_m, \bar{o}_m \mid \acute{s}_m, a_m)$ is as follows, where $\bar{O}_h$ and $\bar{\pi}_h$ are fully determined by the corresponding STM:

$$\sum_{\bar{o}_h} T_m(s' \mid s, a_c) \bar{O}_m(\bar{o}_m \mid \bar{s}'_m, a_c) \beta_h(I'_h \mid I_h, a_c, \bar{o}_h)$$

$$\int \bar{O}_h(\bar{o}_h \mid \bar{s}'_m, a_c, STM) \bar{\pi}_h(\bar{a}_h \mid I_h, STM) d\mu(STM; \theta). \qquad (5)$$

The $\acute{D}_m$ is simply the filtration of $\acute{D}_{m,\theta}$ similar to the equation 2's $\bar{D}_i$, which makes sure θ transitions according to the Bayes rule.

4 LEARNING AS PLANNING IN HUMACE

Any planning algorithm designed for BA-POMDPs can be applied to BA-BRMs, and therefore can solve the machine's problem in

Figure 2: Food Shelter environment taken from Dimitrakakis et al. [5]. Food randomly appears and gives positive reward when consumed. The agent receives a negative reward for each time-step the shelter remains collapsed, and can re-build it by visiting its grid.

HuMaCe [31]. In this section, we present a specific adaptation of the *root sampling of the model* variant of the BA-POMCP algorithm proposed in Katt et al. [22], for the case of machine-optimistic human models.
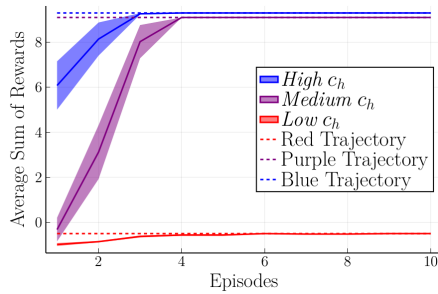
An adaptation for machine-optimistic humans. For partially observable state-spaces, each simulation starts by sampling (s, θ) from the current belief, and in fully-observable settings the current (s, θ) is known. Then, an STM from $\mu(STM; \theta)$ is sampled, and from thereon the simulation proceeds with the STM fixed.

Updating our belief about the human's STM $\mu(STM; \theta)$ is intractable, since computing the likelihood requires computing $\bar{\pi}_h$ for infinitely many STMs. However, the (PO)MDPs of STMs are discrete, thus the set of possible $\bar{\pi}_h$s is finite and discrete for machine-optimistic human models. This implies that for MoH, there are finitely many behavioural equivalence classes over the space of STMs. We take advantage of this implication and discretize μ with a weighted set of particles $\hat{\mu} = \{STM^{(k)}\}_{k=1,\dots,N}$. The $\hat{\mu}$ is initialized by sampling STMs according to a prior $\mu(STM; \theta_0)$, and $\bar{\pi}_h$ is computed and stored offline for each particle. The $\mu(STM; \theta)$ therefore is updated by filtering the particles in $\hat{\mu}$.

Maintaining particles. In some cases, we may need to reinvigorate the particle set $\hat{\mu}$. New STMs can be added by perturbing the existing ones. Unfortunately this will require us to solve the new STMs in the online phase. However, if the perturbations applied to existing particles are bounded, it is reasonable to assume the newly added STMs will lead to $\bar{\pi}_h$s that are close to pre-existing ones thanks to behavioural equivalence. In that case, this can be amortized by training a neural network parameterized by ϕ, $\hat{\pi}_h(\bar{a}_h \mid I_h, STM; \phi)$, which takes (I_h, STM) as input and provides a distribution over $\bar{a}_h$. Further details of this amortization procedure are discussed in Supplementary Section 3.

5 EXPERIMENTS AND RESULTS

We consider two cases of bounded-rationality when the humans want the machine to help them make better decisions, and apply our framework instantiated with the machine-optimistic human models. In both cases, the machine's STM_m is a better model of OTM compared to the human's STM_h. For the Food Shelter environment, the converse of this case, where the human can correct the machine's behaviour and help it learn a better policy, is also presented. All experiments are run with 19 seeds and both the $c_h(s, a) = c_h$ and

Figure 3: The machine's return in the Food Truck environment with a time-inconsistent human (γ = 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.

$c_m(s, a) = c_m$ are assumed to be non-negative constants. However, the machine does not know c_h and must infer it from interaction. The details of each setting are provided in Supplementary Section 2.

5.1 Time-Inconsistent Preferences

Motivation and setup. In the introduction, we have given an example for time-inconsistent behaviour: the red trajectory in Figure 1. We implement the gridworld shown in Figure 1 as an MDP M and the objective optimality criterion OPT is the undiscounted sum of rewards. The subjective task models of agents are $STM_i = (M, OPT_i), i \in \{h, m\}$, where OPT_h is the hyperbolically discounted sum of rewards with discount function $d(t; \gamma) = \frac{1}{1+t\gamma}$, and the OPT_m is exponentially discounted with factor $\lambda = 0.95$. Both γ and c_h are unknown to the machine. If γ is small enough, $d(t; \gamma)$ becomes flat and behaves like exponential discounting, thus this model space is expressive enough to capture both time-inconsistent and consistent behaviour.

In this setting, we do not need the amortization from $\hat{\pi}_h$. Let Q_λ^* denote an optimal Q-function computed with exponential discounting $d(t; \lambda) = \lambda^t$. Any optimal Q-function Q_γ^* computed with hyperbolic discounting $d(t; \gamma) = \frac{1}{1+t\gamma}$ is equal to $\int_0^1 w(\gamma, \lambda) Q_\lambda^* d\lambda$ where the weights are computed in closed form as $w(\gamma, \lambda) = \frac{1}{\gamma} \lambda^{\frac{1}{\gamma}-1}$ [11]. We discretize the belief space with a grid of (c_h, γ) values, and approximate the integral with a weighted Riemannian sum over a grid of λs and their corresponding Q_λ^*s offline. The details are deferred to Supplementary.

Results. We simulate the interaction with six behavioural classes of the human based on (c_h, γ) pairs: When γ is high (≥ 2.0), the human's solution to STM_h produces the time-inconsistent red trajectory, and if low (≤ 0.5) the hyperbolic function becomes flat and the OPT_h agrees with OPT_m, 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. In this case, the Monte Carlo planning in the first episodes attempts to execute the blue trajectory and gets overridden. Once the belief is updated with the override, the following episodes produce the purple trajectory. Figure 3 shows the machine's undiscounted return (i.e. $\sum_t \bar{r}_m^{(t)}$) for the three cases when γ is high. The dashed lines show the returns of 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 planner learns that the purple is. Thus, the machine improves the human's return 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.
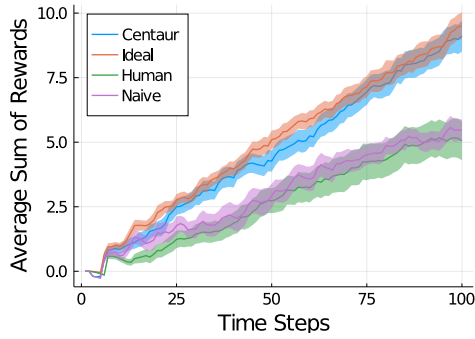
5.2 Overestimation of Probabilities

Motivation and setup. When using probabilities for making decisions, humans sometimes overestimate the probability of negative events, which may lead to less than ideal decisions, or underestimate, and take risks they cannot afford to. The former may result from a fear and an overestimation of the risk. When overestimation causes the human to unnecessarily avoid certain actions, it is called maladaptive avoidance behaviour [20].
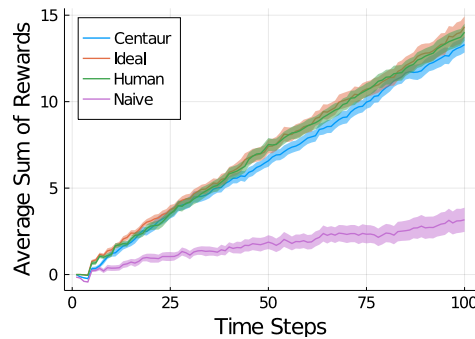
We will model the case of overestimation as follows: Given an MDP M, the STMs are $STM_i = (M_i, OPT), i \in \{h, m\}$ where the human and the machine disagree in transition probabilities. The T_h is an ϵ_{T_h}-approximation of T with $KL(T(. \mid s, a) \mid\mid T_h(. \mid s, a)) \leq \epsilon_{T_h}$ for all (s, a), whereas, $T_m = T$ which means $STM_m = OTM$. The latter can be relaxed to an ϵ_{T_m}-approximation as long as $\epsilon_{T_h} \geq \epsilon_{T_m}$. A similar setting with a fully-known human model has been investigated by Dimitrakakis et al. [5] and we use the same Food Shelter environment from their paper in our experiment, shown in Figure 2. Here, food re-appears uniformly randomly and the shelter collapses with non-zero probability. All actions have the same noise where the execution of an action fails with probability 0.1, transitioning the agent to a neighbour state uniformly random. The human believes that the diagonal action noise is $0.1 + 2\epsilon$ and the rest are $0.1 + \epsilon$.

Results. Dimitrakakis et al. [5] assume that both the T_h (i.e. ϵ) and c_h are fully known to the machine, but in our experiments they are unknown and must be learned from the interaction.

The worst human performance in their experiment is for $\epsilon = 0.45$, so we chose this setting and set a low cost of effort $c_h = 0.05$. The machine's uniform belief on $\epsilon \in [0.0, 0.45]$ and $c_h \in [0.0, 0.5]$ are discretized covering different behavioural classes. In this case, the initial set of particles had good coverage and re-invigoration was not necessary. We have run the experiment for a fixed horizon of 100 steps, and noticed that the centaur can infer the true human model STM_h quickly in about the first 10 time steps. Figure 4a shows the machine's undiscounted return (i.e. $\sum_t \bar{r}_m^{(t)}$) evaluated in the real environment (i.e. OTM) for three cases: *centaur* is our method, *naive* is when the machine tries to execute its optimal solution to STM_m, and *ideal* is when the machine knows STM_h and solves the BRM_m exactly. Additionally, the *human* shows the human's optimal solution to STM_h evaluated in the real environment. The centaur's

(a) The machine's model is correct and the human's is wrong. 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.

(b) The machine's model is wrong and the human's is correct. The *human* is the best case policy. The *naive* performs very badly since the machine is overridden many times. *Ideal* knows human's model, and the *centaur* can quickly infer the human's model and avoid getting overridden by learning to perform the better policy.

Figure 4: Machine's return in the Food Shelter environment under overrides of a human, with $(\epsilon = 0.45, c_h = 0.05, c_m = 0.2)$.

performance is consistently above the naive and the human, since in certain states it can nudge the human to allow diagonal actions.

Figure 4b shows the results for the case when the STMs of the two agents are swapped: the human is correct and the machine is wrong. The centaur learns to perform similarly to the human by inferring the human's model and using it to avoid getting overridden. This way, the human effectively teaches the machine to perform a better policy via overrides.
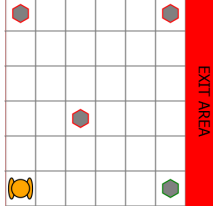
6 AN ANALYSIS OF THE PARTIALLY OBSERVABLE CASES

Even though we have defined the HuMaCe and our algorithm in general for partially-observable settings, the *OTM* and the *STMs* have been fully observable in our empirical results in order to make sure the qualitative aspects of the interaction between the human and the machine are apparent. When both STM_h and STM_m are POMDPs, differences between observation models and transition

models can still be converted to differences in transition dynamics in their corresponding fully-observable belief MDPs. However, this does not reduce to the same setting described in Section 5.2. In the fully observable case of Section 5.2, the machine and the human fully agree on the state they are in, but in the case of partial observability the two agents may disagree on this. Specifically, at any given time t, the agents' beliefs $(b_{h,t}, b_{m,t})$ may not be the same even if they have the exact same action-observation history. This can happen not only due to differences in transition or observation models, but also due to differences in how the beliefs are maintained. If $b_{h,t} \neq b_{m,t}$, then even when their value functions are the same, the agents may still disagree on what action to take. This is an additional challenge brought by partial observability, and we will call it the *belief alignment problem*.

Here, we present an intuitive example followed by an analysis of the belief alignment problem, and discuss its relation to the task structure. Our analysis reveals a novel trade-off for centaurs: to be able to nudge the human's decisions, the machine must make sure their beliefs are sufficiently aligned. However, aligning beliefs might be costly, and in some cases this trade-off can make it infeasible for the machine to augment the human. This result indicates the importance of studying the partial observability for centaurs in greater detail in future work.

An example for the belief alignment problem. Consider the Rock-Sample environment in Figure 5. Here, a rover is rewarded for picking good rock samples and penalized for picking bad ones. The rover knows its position on the map and the position of the rocks, but does not know the quality of the rocks. We denote the fully and partially observed parts of the state variable with s^k (known) and s^u (unknown). The rover can measure any rock from any grid, but the distance to the measured rock increases observation noise exponentially with a rate determined by the sensor efficiency parameter. When the rover is on top of the rock, the distance is zero, thus there is no noise. Now, let the subjective models of the human and the machine be STM_h and STM_m where the only difference between the two is in the observation models $O_h \neq O_m$: machine thinks the sensor efficiency is quite high whereas the human thinks it is extremely low. Let us also assume that the machine is correct about the true sensor efficiency, therefore $O_m = O$. In this case, since the human thinks observations are extremely noisy, their optimal policy is to visit each rock, measure them on top for a noise-free observation, pick the good ones and then exit. However, the rocks are quite far apart and this wastes time. The machine knows that the true sensor efficiency is much higher, and its optimal policy is to measure all rocks from the starting corner and simply pick up the good ones. Unfortunately, the human does not allow this and overrides. Qualitatively, a similar result to our previous experiments would be if the machine learns a policy that, instead of measuring from the start location, gets closer to the rocks before measuring. This way it can avoid getting overridden, pick up the good rocks, and still save some time compared to the human's policy. Unfortunately, this may not be possible. Since the measurements were not made on top of the rocks, the human considers them very noisy. Therefore, after receiving a "good" observation from a good rock, b_h changes ever slightly while b_m becomes confident that the rock is good, which means the human may still not allow the machine to pick it up. The

Figure 5: The RockSample environment. Red rocks are bad quality and greens are good. The rover must pick good rocks and avoid picking bad rocks. The quality of the rocks are partially observed through noisy measurement actions.

machine can try to convince the human by measuring the same rock multiple times to get b_h closer to b_m, however this can take even longer than simply traversing all the rocks. We note that this issue arises due to the properties of the RockSample environment and not our model. Next, we examine what these properties are, and derive theoretical results intended to guide further research.

Analysis of the RockSample task. In the rest of this section, we will analyse how the structure of transition and observation dynamics affect the belief alignment problem by using the RockSample as a running example. Let us define two operators $\mathbb{T}_i^a(b)$ and $\mathbb{O}_i^{o,a}(b)$ where $(\mathbb{T}_i^a b)(s) = \sum_{s'} b(s') T_i(s \mid s', a)$ and $(\mathbb{O}_i^{o,a} b)(s) = \frac{b(s) O_i(o \mid s)}{\sum_{s'} b(s') O_i(o \mid s')}$. The former propagates a belief through the dynamics model and the latter conditions it on the observation. Therefore, once an action a is taken and observation o is received, rational POMDP agents update their beliefs with $b' = \mathbb{O}_i^{o,a}(\mathbb{T}_i^a(b))$. For the rest of this section, we will assume the human and the machine both have the true dynamics model $T_h = T_m = T$. The dynamics of RockSample is fully deterministic, and the transitions do not change the unknown factor s^u (quality of rocks) at all, which means $\mathbb{T}^a(b) = b$ for all a. Moreover, only the measurement actions provide belief changing observations, so $\mathbb{O}_i^{o,a}(b) = b$ for any action other than measurement. The following results specify how the sparsity of belief changing observations and the fact that the transitions do not alter beliefs severely limit the possibility of the two agents aligning their beliefs.

Definition 6.1 (Value of observation [10]). Given a hidden Markov model with an observation distribution O, let $\mathbf{M}^O$ denote the observation matrix with entries $O(o \mid s)$. HMM's value of observation $\gamma(O)$ is defined as $\inf_{x:||x||_1=1} || \mathbf{M}^O x ||_1$. If $\gamma(O) = 0$ the process is unobservable, and if it is 1, fully observed.

For a fixed centaur policy π_c, the subjective POMDPs become hidden Markov models, and the value of observation for each agent depends on the π_c. In RockSample, if π_c does not execute any measurement action, the value of observation is 0 for both agents and their beliefs do not change. The more measurement actions π_c executes, the higher the value of observations gets.

Next, we focus on how the structure of transition dynamics affect agents' belief updates.

Definition 6.2 (Minimal mixing rate [4]). For a Markovian process with a stochastic transition dynamics T, minimal mixing rate $\alpha(T)$ is defined as $\min_{s_1,s_2} \sum_{s'} \min[T(s' \mid s_1), T(s' \mid s_2)]$.

The main result of Boyen and Koller [4] is that $KL(\mathbb{T}^a(b_m) \mid\mid \mathbb{T}^a(b_h)) \leq (1 - \alpha(T)) KL(b_m \mid\mid b_h)$. When the transition dynamics are deterministic, the $\alpha(T) = 0$, and the contraction result becomes uninformative. In RockSample, it is obvious that since $\mathbb{T}^a(b) = b$, the equality occurs. However, for deterministic dynamics in general, the bound does not tell us whether the contraction happens or not. The following lemma provides a necessary condition for the inequality to be strict in the case of deterministic transitions. Its proof is in Supplementary Section 1.2.

LEMMA 6.3. *Let $\mathbf{T}^a$ be the deterministic transition matrix with entries (i, j) as $T(s_i \mid s_j, a)$ where $\mathbb{T}^a(b) = \mathbf{T}^a b$. If $rank(\mathbf{T}^a) = \mid S \mid$ for all $a \in \mathcal{A}$, then $KL(\mathbb{T}^a(b_m) \mid\mid \mathbb{T}^a(b_h)) = KL(b_m \mid\mid b_h)$ for all $a \in \mathcal{A}$ and $b_h, b_m \in \Delta(S)$. Therefore, $rank(\mathbf{T}^a) < \mid S \mid$ is a necessary condition for the strict inequality.*

If $rank(\mathbf{T}^a) < \mid S \mid$ for some a but it is $\mid S \mid$ for others, whether the beliefs come closer or not will depend on π_c. In general, for $\alpha(T) = 0$ (deterministic dynamics) and $\gamma(O) = 0$ (unobservable), full-rank transition matrices mean beliefs cannot contract. The condition of Lemma 6.3 is indeed satisfied by the RockSample dynamics.

Next, we combine the value of observation with the minimal mixing rate to derive a bound that describes under what conditions the beliefs of two agents will come closer. The proof of the theorem combines a lemma from Even-Dar et al. [10] with the theorem of Boyen and Koller [4], and is given in Supplementary Section 1.2.

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;*

$$\mathbb{E}_{o \sim O_m(.|b_m, a_c)}[KL(b_m' \mid\mid b_h')] \leq \left(1 - \alpha(\mathbf{T}^{a_c})\right) KL(b_m \mid\mid b_h)$$
$$+ \gamma(O_m) 3\sqrt{\epsilon_O} - \left(\frac{\gamma(O_m) KL(b_m \mid\mid b_h)}{\sqrt{2}\log\frac{1}{\mu}}\right)^2,$$

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.

Theorem 6.4 shows that if the dynamics are deterministic and the O_m is unobservable, the beliefs will not expand. When the O_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 machine can do to align beliefs is to increase $\gamma(O_m)$. In RockSample, this means choosing a π_m that gets more measurement actions executed. However, the measurement actions cost time, and this trade-off can make it infeasible for the machine to align the human's beliefs with its beliefs sufficiently.

7 RELATED WORKS

Models and problems related to HuMaCe. Hybrid intelligence has emerged as a paradigm for designing AI systems that amplify human intelligence and decision-making [2]. Here, we see two main directions of research: (I) Autonomous AI agents that try

to influence the collective behaviour of humans positively and (II) Semi-autonomous AI tries to influence its own user via advice, negotiation, or other means. In the former, previous works have shown that self-driving cars can leverage their influence on other human drivers to improve the traffic conditions for everyone [26, 36, 38, 39, 41]. This line of work is complementary to ours, which follows the latter direction. Our machine focuses on improving the outcome for its human user by influencing the decisions, similar to behaviour change support systems [30]. The closest model to our HuMaCe is the multi-view MDPs and c-intervention games proposed by Dimitrakakis et al. [5], where the human and the machine's MDPs disagree only by transition probabilities and the human may override the machine's action. We generalize this to partially observable environments and other types of disagreements between models. Their interaction protocol is a Stackelberg competition where the machine commits to a policy for the whole horizon, and the human observes the machine's policy and best-responds. In many cases, it is more realistic to assume the human will be able to observe the machine's action, but not know its entire policy, and it may not be possible to make the human act as a follower. Thus, we designed our interaction protocol as a sequential game where the human has the advantage of observing the machine's action and then respond. Their work assumes the full knowledge of the human's STM_h and the cost of effort c_h. We generalize this to learning the human's STM_h and c_h from interaction. A similar yet less related model is the Helper Assistant MDP from Fern et al. [12], which only considers the case when the human's goals are unknown and the machine can execute its actions without permission.

Empirical evidence for our models of interactive human behaviour. Altshuler and Sarne [3] introduced a sub-optimal decision-making design for when the humans continuously control the autonomy of an assistant system. Their human studies indicate that humans may misjudge the system's performance because their subjective view of the world does not account for high stochasticity. Elmalech et al. [7] have demonstrated that in systems that advise humans, sometimes dispersing sub-optimal advice can increase the performance compared to always dispersing optimal advice, because the optimal advice may be too difficult for the human to heed. In the experiments of Grgić-Hlača et al. [18] for legal decision-making, the machine advice was more accurate than humans' final decisions, and humans were more likely to listen to the advice if they are weakly incentivized. Hyperbolic discounting has been proposed as a good model for time-inconsistent behaviour, consistent with the empirical evidence from human studies [1, 8, 9, 17, 23]. Previous work in behavioural economics investigated the structure of planning problems where human time-inconsistency may be harmful [24, 25]. The solutions proposed in these works forbid some options of the human by removing them, and therefore not suitable as actions for an AI.

8 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. 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.

9 DISCUSSIONS AND FUTURE WORK

The human in a centaur is modelled as a decision-maker who *may* be imperfect due to their bounded-rationality. Of course, all models of humans are wrong, but some are useful. We need not claim that our method can identify the *true model* of human decisions, just that given a good model space it can infer a useful model for assistance. In our interaction protocol, the AI cannot prohibit any trajectory. For instance, completely ignoring the AI and committing to always overriding is an admissible strategy for the human in the sequential game. We consider the human as an independent, self-interested agent, and the c_h comes from their own incentives. This 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) [33]. The STMs allow us to capture a rich set of differences between humans and AI systems. Since the STMs are surrogates for an objective task model, the subjective views of agents are not arbitrarily different. This is both realistic and computationally useful, since it reduces the size of the model space of humans.

Limitations and future work. BRMs can capture a wide range of behaviours, including bounded-rationality, and our experiments indicate that even infinite model spaces may lead to finite behavioural equivalence classes. Our framework is not limited to machine-optimistic humans, and future work can consider more advanced models for the human such as learning agents with bounded adaptation capabilities to the machine's behaviour. A richer action space can allow the machine to explain its reasoning by presenting its predictions to the human. An adaptive human can also learn a better O_h through interaction, which opens further theoretical questions such as how the learning rate of the human and the machine affect the belief contraction. In the most general case, the human would use a BRM and model the machine just as the machine models them, which leads to an infinite recursion of beliefs. Previous work indicates that humans on average do not recurse too deep when reasoning about others [6], and a fixed-depth recursion is still tractable within our framework thanks to Zettlemoyer et al. [42].

REFERENCES

[1] George Ainslie. 2001. *Breakdown of Will*. Cambridge University Press.

[2] Z. Akata, D. Balliet, M. de Rijke, F. Dignum, V. Dignum, G. Eiben, A. Fokkens, D. Grossi, K. Hindriks, H. Hoos, H. Hung, C. Jonker, C. Monz, M. Neerincx, F. Oliehoek, H. Prakken, S. Schlobach, L. van der Gaag, F. van Harmelen, H. van Hoof, B. van Riemsdijk, A. van Wynsberghe, R. Verbrugge, B. Verheij, P. Vossen, and M. Welling. 2020. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer* 53, 8 (2020), 18–28.

[3] Nadav Kiril Altshuler and David Sarne. 2018. Modeling Assistant's Autonomy Constraints as a Means for Improving Autonomous Assistant-Agent Design. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm, Sweden) (*AAMAS '18*). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1468–1476.

[4] Xavier Boyen and Daphne Koller. 1998. Tractable Inference for Complex Stochastic Processes. In *UAI '98: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, University of Wisconsin Business School, Madison, Wisconsin, USA, July 24-26, 1998*, Gregory F. Cooper and Serafín Moral (Eds.). Morgan Kaufmann, 33–42.

[5] Christos Dimitrakakis, David C. Parkes, Goran Radanovic, and Paul Tylkin. 2017. Multi-View Decision Processes: The Helper-AI Problem. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5443–5452.

[6] Prashant Doshi, Xia Qu, Adam Goodie, and Diana L. Young. 2010. Modeling recursive reasoning by humans using empirically informed interactive POMDPs. In *9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010), Toronto, Canada, May 10-14, 2010, Volume 1-3*, Wiebe van der Hoek, Gal A. Kaminka, Yves Lespérance, Michael Luck, and Sandip Sen (Eds.). IFAAMAS, 1223–1230.

[7] Avshalom Elmalech, David Sarne, Avi Rosenfeld, and Eden Shalom Erez. 2015. When Suboptimal Rules. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, Blai Bonet and Sven Koenig (Eds.). AAAI Press, 1313–1319.

[8] Owain Evans, Andreas Stuhlmüller, and Noah D. Goodman. 2016. Learning the Preferences of Ignorant, Inconsistent Agents. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, Dale Schuurmans and Michael P. Wellman (Eds.). AAAI Press, 323–329.

[9] Owain Evans, Andreas Stuhlmüller, John Salvatier, and Daniel Filan. 2017. Modeling Agents with Probabilistic Programs. http://agentmodels.org. Accessed: 2021-5-17.

[10] Eyal Even-Dar, Sham M. Kakade, and Yishay Mansour. 2007. The Value of Observation for Monitoring Dynamic Systems. In *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, Manuela M. Veloso (Ed.). 2474–2479.

[11] William Fedus, Carles Gelada, Yoshua Bengio, Marc G. Bellemare, and Hugo Larochelle. 2019. Hyperbolic Discounting and Learning over Multiple Horizons. *CoRR* abs/1902.06865 (2019). arXiv:1902.06865

[12] Alan Fern, Sriraam Natarajan, Kshitij Judah, and Prasad Tadepalli. 2014. A decision-theoretic model of assistance. *Journal of Artificial Intelligence Research* 50 (2014), 71–104.

[13] Shane Frederick, George Loewenstein, and Ted O'Donoghue. 2002. Time discounting and time preference: A critical review. *Journal of Economic Literature* 40, 2 (2002), 351–401.

[14] Samuel J Gershman, Eric J Horvitz, and Joshua B Tenenbaum. 2015. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* 349, 6245 (2015), 273–278.

[15] P. Gmytrasiewicz and E. Durfee. 2004. Rational Coordination in Multi-Agent Environments. *Autonomous Agents and Multi-Agent Systems* 3 (2004), 319–350.

[16] Piotr J. Gmytrasiewicz and Prashant Doshi. 2005. A Framework for Sequential Planning in Multi-Agent Settings. *J. Artif. Intell. Res.* 24 (2005), 49–79.

[17] Leonard Green, Nathanael Fristoe, and Joel Myerson. 1994. Temporal discounting and preference reversals in choice between delayed outcomes. *Psychonomic Bulletin & Review* 1, 3 (1994), 383–389.

[18] Nina Grgić-Hlača, Christoph Engel, and Krishna P. Gummadi. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 178 (Nov. 2019), 25 pages.

[19] A. Heifetz and Y. Spiegel. 2000. On the Evolutionary Emergence of Optimism. *Microeconomic Theory eJournal* (2000).

[20] Kristina M. Hengen and Georg W. Alpers. 2019. What's the Risk? Fearful Individuals Generally Overestimate Negative Outcomes and They Dread Outcomes of Specific Events. *Frontiers in Psychology* 10 (2019), 1676.

[21] Mark K. Ho and Thomas L. Griffiths. 2021. Cognitive science as a source of forward and inverse models of human decisions for robotics and control. *CoRR*

abs/2109.00127 (2021). arXiv:2109.00127

[22] Sammie Katt, Frans A. Oliehoek, and Christopher Amato. 2017. Learning in POMDPs with Monte Carlo Tree Search. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 1819–1827.

[23] Kris N Kirby and Richard J Herrnstein. 1995. Preference reversals due to myopic discounting of delayed reward. *Psychological Science* 6, 2 (1995), 83–89.

[24] Jon Kleinberg, Sigal Oren, and Manish Raghavan. 2016. Planning Problems for Sophisticated Agents with Present Bias. In *Proceedings of the 2016 ACM Conference on Economics and Computation* (Maastricht, The Netherlands) (*EC '16*). Association for Computing Machinery, New York, NY, USA, 343–360.

[25] Jon M. Kleinberg and Sigal Oren. 2014. Time-inconsistent planning: a computational problem in behavioral economics. In *ACM Conference on Economics and Computation, EC '14, Stanford , CA, USA, June 8-12, 2014*, Moshe Babaioff, Vincent Conitzer, and David A. Easley (Eds.). ACM, 547–564.

[26] Abdul Rahman Kreidieh, Cathy Wu, and Alexandre M. Bayen. 2018. Dissipating stop-and-go waves in closed and open networks via deep reinforcement learning. In *21st International Conference on Intelligent Transportation Systems, ITSC 2018, Maui, HI, USA, November 4-7, 2018*, Wei-Bin Zhang, Alexandre M. Bayen, Javier J. Sánchez Medina, and Matthew J. Barth (Eds.). IEEE, 1475–1480.

[27] Heather C. Lench and Shane W. Bench. 2012. Automatic Optimism: Why People Assume Their Futures will be Bright. *Social and Personality Psychology Compass* 6, 4 (2012), 347–360. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-9004.2012.00430.x

[28] Richard L. Lewis, Andrew Howes, and Satinder Singh. 2014. Computational Rationality: Linking Mechanism and Behavior Through Bounded Utility Maximization. *Topics in Cognitive Science* 6, 2 (2014), 279–311.

[29] George Loewenstein and Drazen Prelec. 1992. Anomalies in intertemporal choice: Evidence and an interpretation. *The Quarterly Journal of Economics* 107, 2 (1992), 573–597.

[30] Harri Oinas-Kukkonen. 2012. A foundation for the study of behavior change support systems. *Personal and Ubiquitous Computing* 17 (2012), 1223–1235.

[31] Frans A. Oliehoek and Christopher Amato. 2014. Best Response Bayesian Reinforcement Learning for Multiagent Systems with State Uncertainty. In *Proceedings of the Ninth AAMAS Workshop on Multi-Agent Sequential Decision Making in Uncertain Domains (MSDM)*.

[32] Frans A. Oliehoek and Christopher Amato. 2016. *A Concise Introduction to Decentralized POMDPs*. Springer.

[33] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[34] Todd Rogers, Don A. Moore, and Michael I. Norton. 2017. The Belief in a Favorable Future. *Psychological Science* 28, 9 (2017), 1290–1301. arXiv:https://doi.org/10.1177/0956797617706706

[35] Stéphane Ross, Brahim Chaib-draa, and Joelle Pineau. 2007. Bayes-Adaptive POMDPs. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis (Eds.). Curran Associates, Inc., 1225–1232.

[36] Dorsa Sadigh, Shankar Sastry, Sanjit A. Seshia, and Anca D. Dragan. 2016. Planning for Autonomous Cars that Leverage Effects on Human Actions. In *Robotics: Science and Systems XII, University of Michigan, Ann Arbor, Michigan, USA, June 18 - June 22, 2016*, David Hsu, Nancy M. Amato, Spring Berman, and Sam Ade Jacobs (Eds.).

[37] Tali Sharot. 2011. The optimism bias. *Current Biology* 21, 23 (2011), R941–R945.

[38] Eugene Vinitsky, Aboudy Kreidieh, Luc Le Flem, Nishant Kheterpal, Kathy Jang, Cathy Wu, Fangyu Wu, Richard Liaw, Eric Liang, and Alexandre M. Bayen. 2018. Benchmarks for reinforcement learning in mixed-autonomy traffic. In *Proceedings of The 2nd Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 87)*, Aude Billard, Anca Dragan, Jan Peters, and Jun Morimoto (Eds.). PMLR, 399–409.

[39] Eugene Vinitsky, Kanaad Parvate, Aboudy Kreidieh, Cathy Wu, and Alexandre M. Bayen. 2018. Lagrangian Control through Deep-RL: Applications to Bottleneck Decongestion. In *21st International Conference on Intelligent Transportation Systems, ITSC 2018, Maui, HI, USA, November 4-7, 2018*, Wei-Bin Zhang, Alexandre M. Bayen, Javier J. Sánchez Medina, and Matthew J. Barth (Eds.). IEEE, 759–765.

[40] Louise Whiteley and Maneesh Sahani. 2012. Attention in a Bayesian Framework. *Frontiers in Human Neuroscience* 6 (2012), 100.

[41] Cathy Wu, Aboudy Kreidieh, Eugene Vinitsky, and Alexandre M. Bayen. 2017. Emergent Behaviors in Mixed-Autonomy Traffic. In *Proceedings of the 1st Annual Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 78)*, Sergey Levine, Vincent Vanhoucke, and Ken Goldberg (Eds.). PMLR, 398–407.

[42] Luke Zettlemoyer, Brian Milch, and L. Kaelbling. 2008. Multi-Agent Filtering with Infinitely Nested Beliefs. In *NIPS*.