Bayesian Interaction Shaping: Learning to Influence Strategic Interactions in Mixed Robotic Domains

Aris Valtazanos
School of Informatics
University of Edinburgh
Edinburgh, EH8 9AB
a.valtazanos@sms.ed.ac.uk

Subramanian Ramamoorthy
School of Informatics
University of Edinburgh
Edinburgh, EH8 9AB
s.ramamoorthy@ed.ac.uk

ABSTRACT

Despite recent advances in getting autonomous robots to follow instructions from humans, strategically intelligent robot behaviours have received less attention. *Strategic* intelligence entails influence over the beliefs of other interacting agents, possibly adversarial. In this paper, we present a learning framework for strategic *interaction shaping* in physical robotic systems, where an autonomous robot must lead an unknown adversary to a desired *joint* state. Offline, we learn composable interaction templates, represented as *shaping regions* and *tactics*, from human demonstrations. Online, the agent empirically learns the adversary's responses to executed tactics, and the reachability of different regions. Interaction shaping is effected by selecting tactic sequences through Bayesian inference over the expected reachability of their traversed regions. We experimentally evaluate our approach in an adversarial soccer penalty task between NAO robots, by comparing an autonomous shaping robot with and against human-controlled agents. Results, based on 650 trials and a diverse group of 30 human subjects, demonstrate that the shaping robot performs comparably to the best human-controlled robots, in interactions with a heuristic autonomous adversary. The shaping robot is also shown to progressively improve its influence over a more challenging strategic adversary controlled by an expert human user.

Categories and Subject Descriptors

I.2.9 [**Robotics**]

Keywords

Interaction shaping, online strategic decision-making, learning from demonstration, adversarial robotic environments

1. INTRODUCTION

As the physical capabilities of autonomous robots improve, there is also a growing demand for multi-agent applications where robots can interact more seamlessly with other agents. In many such domains, and particularly in those featuring humans or human-controlled agents, robots are currently restricted to a passive role, which requires them to follow and execute instructions or perform tasks of limited behavioural complexity. By contrast, ***strategic*** interactions requiring active influence over the beliefs of an interacting agent remain a challenging problem. The development of such influencing behaviours would constitute an important step towards several practical human-robot interaction applications, where trust and implicit persuasion are relevant issues that need to be incorporated into task specifications.

In this paper, we consider the problem of strategic interaction shaping in adversarial mixed robotic domains. A ***mixed robotic domain*** features both autonomous and human-controlled robots, which have identical hardware but differ at the behavioural level. In this context, the ***interaction shaping*** problem deals with the ability of an autonomous robot to affect the state of a non-cooperative agent in a strategic interactive task (Figure 1) . However, we consider interactions that are only ***partially controllable***, in that the autonomous robot cannot directly force the adversary to the desired joint state. Thus, the robot must shape the interaction *indirectly*, by identifying actions that can cause the adversary to behave in accordance with its own goal. Moreover, a robust shaping robot must learn to influence a given adversary from its own experience, without the provision of additional information on the characteristics (e.g. human-controlled vs. autonomous) of that agent.

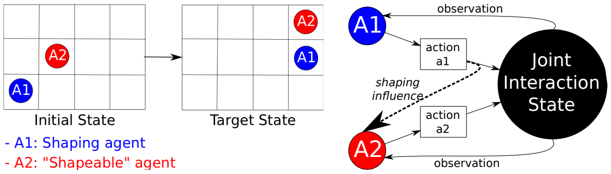


Figure 1: Conceptual structure of the interaction shaping problem. Agent A1 seeks to lead, *without directly forcing*, a *non-cooperative* agent A2 to a new *joint* target state. A1 must achieve this objective by *indirectly influencing*, through its own chosen actions, the actions selected by A2.

Our framework for strategic interaction shaping in adversarial mixed robotic environments (Figure 2) first learns offline a set of interaction templates from provided human demonstrations of the desired strategic behaviour. These templates are encoded as *shaping regions* and *tactics*, which represent salient interactive modes of the state and action spaces. Online, the shaping robot seeks to lead the interaction to a desired joint state by chaining sampled sequences of tactics as an interactive strategy. To achieve this, the

Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013)*, Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.

agent updates local empirical models of its given opponent's responses to individually executed tactics, and a distribution over the *reachability* of the various regions against this player. Tactic sequences are sampled through *iterated prediction* of the adversary's expected responses, and selected through *Bayesian inference* over the expected reachability of their traversed regions. Thus, the shaping robot learns, through repeated interaction, strategies that are likely to successfully shape an interaction with a given adversary.

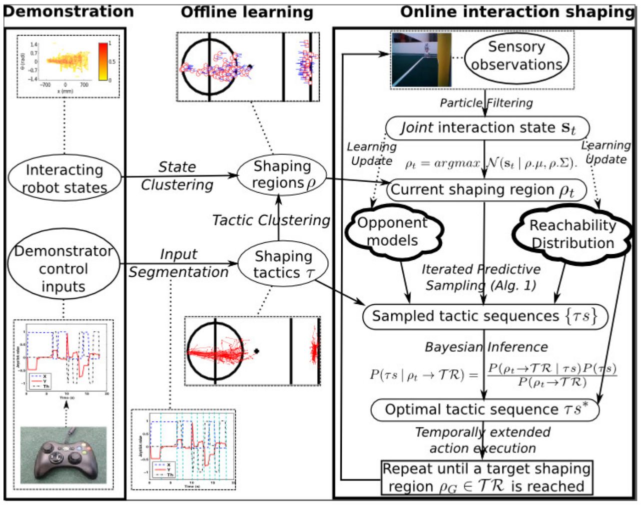

Demonstration
Offline learning
Online interaction shaping
θ (rad)
x (mm)
Interacting robot states
State Clustering
Shaping regions ρ
Tactic Clustering
Demonstrator control inputs
Input Segmentation
Shaping tactics τ
Sensory observations
Particle Filtering
Joint interaction state $\mathbf{s}_t$
Learning Update
$\rho_t = argmax \mathcal{N}(\mathbf{s}_t \mid \rho.\mu, \rho.\Sigma)$.
Learning Update
Current shaping region ρ_t
Opponent models
Reachability Distribution
Iterated Predictive Sampling (Alg. 1)
Sampled tactic sequences $\{\tau s\}$
Bayesian Inference
$P(\tau s \mid \rho_t \to \mathcal{TR}) = \frac{P(\rho_t \to \mathcal{TR} \mid \tau s)P(\tau s)}{P(\rho_t \to \mathcal{TR})}$
Optimal tactic sequence τs^*
Temporally extended action execution
Repeat until a target shaping region $\rho_G \in \mathcal{TR}$ is reached

Figure 2: Our approach to strategic interaction shaping. Human demonstrators provide traces of the desired behaviour, which are converted into shaping regions and tactics, and used as composable templates by an autonomous agent. Online, the shaping agent attempts to reach a desired joint state by sampling tactic sequences through iterated prediction of its adversary's expected responses, and selecting optimal sequences through Bayesian inference. Opponent models and expected region reachability are updated through repeated interaction.

In the following sections, we first situate our work with respect to related literature (Section 2). Then, we describe the interaction shaping method (Section 3), distinguishing between offline and online strategy learning. Our framework is evaluated in an adversarial robotic soccer interaction, for which we use the NAO robots (Section 4), and where we compare the resulting agent with and against human-controlled agents. Our results (Section 5) demonstrate an ability to improve shaping performance over time even when interacting with a challenging human-controlled adversary. We discuss possible extensions to our work in Section 6.

2. RELATED WORK

Reward shaping [11] is the process of affecting an agent's learning by providing additional rewards, as a means of inciting desirable behaviours. A related concept is active indirect elicitation [19], where an agent's reward function is inferred from supplied incentives. Behaviour shaping has also been considered in multi-agent interactions combining competition and collaboration, e.g. the lemonade stand game [18], where the reward experienced by an agent depends both on its own decisions and those of the other interacting agents.

The interaction shaping problem is similarly concerned with influence over the behaviour of an interacting agent. However, we consider purely *adversarial* interactions, where the learning agent is tasked with shaping the behaviour of *non-cooperative* agents, whose goals are conflicting with its own. This is a considerably harder problem that comes closer to most realistic human-robot interactions, where robots do not have explicit control over human actions.

Partially Observable Markov Decision Processes (POMDPs) [9] have been the standard decision model in partially observable domains. Interactive POMDPs (IPOM- DPs) [8] are defined over *interactive* states, which are a product of world states and possible models of the other agents' policies. Both processes are known to be intractable in problems with large state, action, and/or model spaces. Several approximation algorithms have been proposed to address this issue, such as point and sampling-based methods [16, 7, 10]. Furthermore, Decentralised POMDPs [5] consider models where transitions and reward functions are defined in terms of the states and actions of multiple agents.

Interaction shaping deals with similar representational and computational issues, arising from the need to find optimal actions against an unknown adversary. However, we focus on *online*, empirical learning on physical robots (and not simply *offline* optimisation), where equilibrium search is insufficient, so most approximations do not apply directly to our domain. Similarly, decentralised processes typically assume commonly aligned rewards between agents, so they are also incompatible with our setup. Instead, we learn salient modes of the state and action spaces from human demonstrations, and we use them to form a learning algorithm. Moreover, we predict adversarial behaviour by *iteratively sampling* from a set of empirical observations, thus recreating the nested reasoning effect [8] of IPOMDPs. Thus, we address the challenges of interaction shaping in physically grounded systems, where robots must interactively learn to influence non-cooperative agents from limited sensory data.

Semi-Markov Decision Processes (SMDPs) consider actions that can be extended over a time horizon, such as Markov *options* [14], which are generalisations of action selection policies with input and termination conditions. A related problem in the robot control community is the composition of actions into a global policy that leads to an overall goal. A classic example is the sequential composition of local controllers, represented as *funnels* with initial conditions and goal states, which can be chained and activated intermittently [6]. This method is extended to motion planning [15], where funnels represent linear quadratic regulators.

The formulation of shaping regions and tactics is motivated by the above concepts. Shaping tactics are analogous to options, in being temporally extended actions with specific preconditions. However, the input and target states of shaping tactics are interactive, so they account for both agents. This extends traditional SMDP formulations where there is no explicit reasoning about the adversary. Moreover, the synthesis of strategies as tactic sequences bears a similarity to funnel controllers in an interactive setting. In our approach, tactics are chained together through empirical distributions measuring the reachability of shaping regions. This leads to policies that are expected to maximise the probability of attaining a desired interactive target state.

Opponent modeling is often concerned with influence over the beliefs of adversarial agents, with the aim of attaining intelligent human behavioural traits, e.g. *bluffing* [13] and *deception* [17]. In interaction shaping, we seek to learn and reproduce similar behaviours on physical robots, in environments where the nature (human/autonomous) of the adversary is difficult to model analytically. To achieve this objective, our learning formulation combines techniques such as action sampling, iterated reasoning, and Bayesian inference. Thus, the resulting framework leads to interactive strategic learning in physically grounded multi-robot environments.

3. METHOD

3.1 Preliminaries and notation

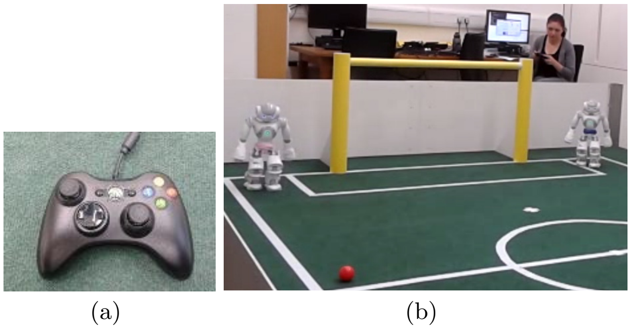
We consider a system of two robots, R and R', interacting in a planar world, where R is the *shaping* and R' is the *shapeable* agent. At time $t \in \Re^+$, the system is described by:

- The joint state of the two robots, $\mathbf{s}_t = \langle s_t, s'_t\rangle$, $s_t = [x_t, y_t, \theta_t]^T$, $s'_t = [x'_t, y'_t, \theta'_t]^T$, where $\{x_t, x'_t\} \in \Re$, $\{y_t, y'_t\} \in \Re$ are the positional coordinates, and $\{\theta_t, \theta'_t\} \in [-\pi, +\pi]$ are the orientations of the robots[1].
- The action vectors, $\vec{a}_t$, $\vec{a}'_t$ available to the robots – each vector may consist of both discrete and continuous actions. For example, in a task involving navigation and manipulation, one choice would be $[dx, dy, d\theta, grip]^T$, where $\{dx, dy, d\theta\} \in [-1.0, +1.0]$ are the requested translation and rotation as fractions of the maximum speed of the robot, and $grip \in \{open,\ close\}$ is a command for the robot's actuator.

The goal of R is to lead R' to one of several possible *target states* $\mathbf{z} \in \mathbf{Z}$, over a time horizon η, where each $\mathbf{z} = \langle s, s'\rangle$ represents a joint target configuration. In other words, R seeks to reach, at some time $t \leq \eta$, a joint state $\mathbf{s}_t \in \mathbf{Z}$. We use the superscript ^{H} for human-controlled robots; for example, s_t^H is the state of one such robot at time t.

3.2 Learning from human demonstration

In the offline learning phase, we extract basic behavioural templates from demonstrations by humans teleoperating R in the desired interaction (Figure 3). In this phase, R' can be either also human-controlled, or an autonomous robot executing a hard-coded behaviour. R is teleoperated through a game pad , which maps inputs to action vectors $\vec{a}^H$.



(a) (b)

Figure 3: Learning from demonstration. (a): The input device. (b): A demonstrator controlling a robot (right, blue waistband) against an autonomous robot (left, pink waistband) in an interactive task.

A demonstration $\mathbf{q}$ is a time-indexed sequence of recorded robot states and (demonstrator) actions, $\{\{t_i, s_{t_i}^H, s'_{t_i}, \vec{a}_{t_i}^H\} \mid i = 1...M\}$, where M is the number of recorded points. Each demonstration is manually annotated as a *successful* or *unsuccessful* example of shaping behaviour. We retain the set of successful demonstrations, $\mathbf{Q}^+ = \{\mathbf{q}_1^+, ..., \mathbf{q}_N^+\}$.

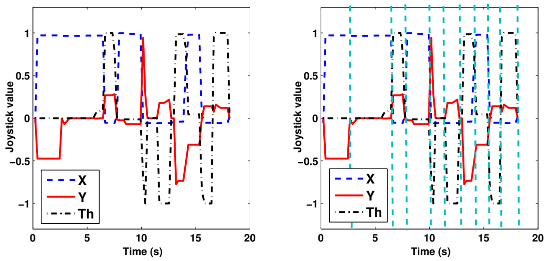
In the remainder of this section, we explain how the successful demonstrations are converted into shaping ***tactics*** and ***regions***, which serve as "building blocks" for the shaping agent in the online learning phase (Section 3.3).

3.2.1 Interaction shaping tactics

An interaction shaping tactic τ is a time-independent action continually evoked by R for a variable period of time:

$$\tau = \langle \mathbf{i\check{s}}, \mathbf{t\check{s}}, \check{a}, dt, \{\tilde{\mathbf{r}}\}\rangle, \tag{1}$$

where $\mathbf{i\check{s}}, \mathbf{t\check{s}}$, are the joint input and target states of the tactic, $\vec{\mathbf{a}}_{\mathbf{s}}$ is the action followed by R, dt is the duration of this action, and $\{\tilde{\mathbf{r}}\}$ is a set of normalised *expected responses* of R' to τ. A response $\tilde{m} = \langle dx, dy, d\theta\rangle \in \{\tilde{\mathbf{r}}\}$ is a possible move by R', normalised over a time interval $\bar{\mathbf{n}}$, in response to τ. For the remainder of the paper, we take $\bar{\mathbf{n}} = 1$ second.



(a) (b)

Figure 4: Segmentation of control inputs into tactics. (a): Raw translational and rotational motion inputs. (b): A new tactic begins when either of the inputs crosses the activity threshold of ± 0.4 (not shown). Vertical lines indicate tactic boundaries.

Tactics are extracted by segmenting the inputs of each successful demonstration, an example for which is given in Figure 4(a). To account for sensory error in input devices, we define an *activity threshold* ψ below which recorded values are treated as noise. The value of ψ depends on the hardware characteristics of the device. A new tactic begins when at least one input crosses ψ. A tactic is in progress while all inputs remain on their current side of ψ. Figure 4(b) illustrates the segmentation process using this heuristic.

For every extracted tactic τ, we record the start and end times at its boundaries, t_s and t_e. Then, the tactic time interval, input, and target states are $\tau.dt \leftarrow t_e - t_s$, $\tau.\mathbf{i\check{s}} \leftarrow \langle s_{t_s}^H, s'_{t_s}\rangle$, and $\tau.\mathbf{t\check{s}} \leftarrow \langle s_{t_e}^H, s'_{t_e}\rangle$. Similarly, the tactic action vector, $\tau.\check{a}$, is the mean of the inputs over the duration of τ.

Finally, the expected responses of the adversary, $\tau.\{\tilde{\mathbf{r}}\}$, are initialised by dividing each tactic interval into $\lceil \tau.dt/\bar{\mathbf{n}}\rceil$ fixed-length segments. Each segment yields a candidate response by R', which is the change of the state of R' between the segment endpoints, averaged over its duration. We also set a bound $\boldsymbol{m}$ on the size of each $\{\tilde{\mathbf{r}}\}$, so if the number of tactic segments, $\boldsymbol{n}$, exceeds this bound, $\boldsymbol{n}$–$\boldsymbol{m}$ randomly selected candidate responses are discarded. We let $\boldsymbol{m} = \lceil c \cdot \tau.dt/\bar{\mathbf{n}}\rceil$, where $c \geq 1$ is a small positive constant.

[1] Positive directions: forward, left, counter-clockwise.

3.2.2 *Interaction shaping regions*

A shaping region ρ is a normal distribution $\mathcal{N}$ over related states frequently visited by the robots during the interaction:

$$\rho = \langle \mu, \Sigma, \{\tau\} \rangle, \tag{2}$$

where μ is the mean joint state of ρ, Σ is the covariance matrix, and $\{\tau\}$ are the tactics that can be invoked from ρ.

Shaping regions are computed by clustering extracted tactics based on their input states. In particular, we form a set of tactic groups $TG = \{\tau g_1, ..., \tau g_M\}$ based on input state similarity, so that any two tactics τ_i, τ_j within a tactic group τg satisfy $d([\tau_i.\check{s}; \tau_i.\check{s}'], [\tau_j.\check{s}; \tau_j.\check{s}']) < \delta$, where $d(\cdot, \cdot)$ is the distance between two state pairs, and δ is a distance threshold. Each group $\tau g \in TG$ is converted to a new region ρ, whose tactic set is $\rho.\{\tau\} = \tau g$. The parameters $\rho.\mu$ and $\rho.\Sigma$ are the mean and covariance of the input states of all tactics in $\rho.\{\tau\}$. Thus, we obtain a set of shaping regions, $\mathcal{ISR}$.

Finally, the target states of the last tactics are similarly clustered to obtain the set of *target regions*, $\mathcal{TR}$, representing states the shaping agent eventually seeks to reach.

3.3 Bayesian interaction shaping

In the online learning phase, R searches for **tactic sequences**, $\{\tau_1, ..., \tau_N\}$, that are likely to lead R' to a desired state. We represent this as a two-stage tactic *sampling* and *selection* process, which is formulated as a Bayesian problem through an *empirical reachability likelihood* distribution. We expand on these concepts in the remainder of this section, and illustrate how beliefs are updated during the interaction.

3.3.1 *Empirical reachability likelihood*

Interaction shaping depends on the *compliance* of R' with selected tactics. To model this effect, we define the *empirical reachability likelihood* distribution, $\mathcal{RD}$, which expresses the probability of reaching a region with a given tactic:

$$\mathcal{RD}(\rho_1, \tau, \rho_2, \rho_3) \doteq P(\rho_3 \mid \rho_1, \tau, \rho_2). \tag{3}$$

Thus, $\mathcal{RD}(\rho_1, \tau, \rho_2, \rho_3)$ gives the probability of reaching ρ_3, given that τ was invoked from ρ_1 with the intention of reaching ρ_2. As explained in Section 3.3.2, the correlation between intended and actually reached regions is the main bias in selecting robust shaping tactics. We initialise $\mathcal{RD}$ assuming "perfect control" over tactics, so $P(\rho_3 \mid \rho_1, \tau, \rho_2) = 1$ if $\rho_2 = \rho_3$, 0 otherwise. However, these values are continually updated from the observations of R during the interaction.

3.3.2 *Tactic sampling and iterated prediction*

The complexity of the tactic sampling process is bounded by the maximum number of sequences, N_S, and the length of each sequence, L_S, sampled at every decision point. We set $N_S = \max_{\rho \in \mathcal{ISR}} |\rho.\{\tau\}|$, $L_S = (\max_{\mathbf{q} \in \mathbf{Q}^+} |\mathbf{q}|)$, as the sizes of the largest tactic set and longest demonstration, respectively.

The world state at time t, $\mathbf{s}_t$, is estimated through the sensory observations of R. Then, the current region, ρ_t, is

$$\rho_t = \arg \max_{\rho \in \mathcal{ISR}} \mathcal{N}(\mathbf{s}_t \mid \rho.\mu, \rho.\Sigma). \tag{4}$$

To generate a new sequence τs, we first select a tactic τ from $\rho_t.\{\tau\}$ with probability

$$P(\tau \sim \rho_t.\{\tau\}) = \frac{1}{|\mathcal{ISR}|} \sum_{\acute{\rho}} P(\acute{\rho} \mid \rho_t, \tau, \acute{\rho}), \tag{5}$$

so as to reflect the overall expected *successful* reachability of regions from ρ_t using τ (i.e. the overall accordance between expected and actually reached regions). Then, we iteratively predict how the interaction may evolve if R follows τ. The expected state of R, $\tilde{s}$, upon completion of τ, is the target state $\tau.\mathbf{t}\check{\mathbf{s}}$. For R', we sample $\lceil \tau.dt/\bar{\mathbf{n}} \rceil$ responses from $\tau.\{\tilde{\mathbf{r}}\}$. Starting from the current state of R', $\tilde{s}' = s'_t$, we iteratively apply each sampled response, $\tilde{m}$, to $\tilde{s}'$, i.e.

$$\tilde{s}' \leftarrow \tilde{s}' + \sum_{i=1}^{\lceil \tau.dt/\bar{\mathbf{n}} \rceil} \tilde{m}_i \tag{6}$$

This gives the expected state, $\tilde{\mathbf{s}} = \langle \tilde{s}, \tilde{s}' \rangle$, at the end of τ.

We then compute the most likely region of $\tilde{\mathbf{s}}$, $\tilde{\rho} = \arg \max_{\rho \in \mathcal{ISR}} \mathcal{N}(\tilde{\mathbf{s}} \mid \rho.\mu, \rho.\Sigma)$. We call $\tilde{\rho}$ the *expected next region* of τ, denoted as $\tau.\rho_{+1}$. If $\tilde{\rho} \in \mathcal{TR}$, we return the tactics sampled so far as a tactic sequence τs. Otherwise, we repeat the above iterated prediction process, until either a target region is found, or the maximum sequence length L_S is reached. We repeat the whole procedure to obtain N_S sequence samples. The overall sampling method is summarised in Algorithm 1.

Algorithm 1 Tactic Sequence Sampling

```
Input: Joint state $\mathbf{s}_t$, shaping/target regions $\mathcal{ISR}$/$\mathcal{TR}$, reachability distribution $\mathcal{RD}$, max. seq. length $L_S$ and sampling attempts $N_S$, response interval $\bar{\mathbf{n}}$
Output: Set of tactic sequences $\{\tau s\}$
$\{\tau s\} \leftarrow \{\{\}\}$
$\rho_t \leftarrow \arg \max_{\rho \in \mathcal{ISR}} \mathcal{N}(\mathbf{s}_t \mid \rho.\mu, \rho.\Sigma)$ {find current region}
for $i = 1 \rightarrow N_S$ do
    $\tau s \leftarrow \{\}$ {initialise new tactic sequence}
    $j \leftarrow 1,\ \tilde{\rho} \leftarrow \rho_t,\ (\tilde{\mathbf{s}} \equiv \langle \tilde{s}, \tilde{s}' \rangle) \leftarrow \mathbf{s}_t$
    while $\tilde{\rho} \notin \mathcal{TR}$ and $j \leq L_S$ do
        $\tilde{\tau} \sim \tilde{\rho}.\{\tau\}$ {sample tactic using $\mathcal{RD}$ as in Eq. 5}
        $\tilde{s} \leftarrow \tilde{\tau}.\mathbf{t}\check{\mathbf{s}}$ {own expected state $\leftarrow$ tactic target}
        for $j = 1 \rightarrow \lceil \tilde{\tau}.dt/\bar{\mathbf{n}} \rceil$ do
            $\tilde{m} \sim \tilde{\tau}.\{\tilde{\mathbf{r}}\}$ {sample from tactic responses}
            $\tilde{s}' \leftarrow \tilde{s}' + \tilde{m}$ {apply sample}
        end for
        $j \leftarrow j + 1,\ \tilde{\mathbf{s}} \leftarrow \langle \tilde{s}, \tilde{s}' \rangle$
        $\tilde{\rho} \leftarrow \arg \max_{\rho \in \mathcal{ISR}} \mathcal{N}(\tilde{\mathbf{s}} \mid \rho.\mu, \rho.\Sigma)$
        $\tilde{\tau}.\rho_{+1} \leftarrow \tilde{\rho}$ {assign $\tilde{\rho}$ as expected next region}
        $\tau s$.insert($\tilde{\tau}$)
    end while
    $\{\tau s\}$.insert($\tau s$)
end for
return $\{\tau s\}$
```

Algorithm 1 predicts the evolution of at most $N_S \cdot L_S$ tactics in the worst case. The ability to find sequences leading to a target region depends on the convergence of the sampled adversarial responses. If these samples are a good representation of the "true" behaviour of R', expected next regions of a tactic will tend to match the actual reached regions. In interactions with non-stationary adversaries, however, R may not be able to find sequences leading to a target region, owing to the discrepancy between expected and reached states. If no such sequence is found, R attempts to transition to a different region from which better sequences may be retrieved. Thus, given the interactive nature of our domain, our objective is not an exhaustive search over tactics, but instead an efficient sampling procedure yielding bounded-length sequences that are likely to impact the interaction.

3.3.3 *Tactic selection*

Tactic sequences are selected with the intention of reaching a target region $\rho_G \in \mathcal{TR}$. Assuming that all ρ_G are equally desirable for R, we obtain the *posterior probability* of selecting a τs, given that R wants to eventually reach one such $\rho_G \in \mathcal{TR}$ from the current region ρ_t:

$$P(\tau s \mid \rho_t \to \mathcal{TR}) = \frac{P(\rho_t \to \mathcal{TR} \mid \tau s) P(\tau s)}{P(\rho_t \to \mathcal{TR})}. \qquad (7)$$

The *prior probability* of selecting τs is defined in terms of the *inverse total time* of the sequence, $\mathbf{T}^{-1}(\tau s) = 1/(\sum_{\tau \in \tau s} \tau.dt)$:

$$P(\tau s) = \beta(\tau s) \cdot \frac{\mathbf{T}^{-1}(\tau s)}{\sum_{\tau s' \in \{\tau s\}} \mathbf{T}^{-1}(\tau s')}, \qquad (8)$$

where $\beta(\tau s)$ penalises sequences whose last tactic, τ_N, is not expected to reach a target region, i.e. $\beta(\tau s) = 1$ if $\tau_N.\rho_{+1} \in \mathcal{TR}$, $0 < \beta(\tau s) < 1$ otherwise. Thus, short sequences leading to a target are a priori more preferable.

The *likelihood* of reaching a target region, given τs, $P(\rho_t \to \mathcal{TR} | \tau s)$, is computed as the discounted sum of the empirical reachability likelihoods of the constituent tactics of τs,

$$P(\rho_t \to \mathcal{TR} | \tau s) = \frac{\beta(\tau s)}{|\tau s|} \sum_{i=1}^{|\tau s|} \gamma^{i-1} \cdot P(\rho_{+1} | \rho_{-1}, \tau_i, \rho_{+1}) \quad (9)$$

where $\beta(\tau s)$ is defined as above, $0 < \gamma \leq 1$ is a discount factor for future tactics, τ_i is the i-th tactic of τs, ρ_{+1} is the expected next region of τ_i, and ρ_{-1} is the previous region,

$$\rho_{-1} = \begin{cases} \rho_t, & i = 1 \\ \tau_{i-1}.\rho_{+1}, & i > 1 \end{cases}. \qquad (10)$$

The likelihood provides a measure of the expected discounted *compliance* of the adversary with a tactic sequence. Finally, the *normalisation term*, $P(\rho_t \to \mathcal{TR})$, is given by

$$P(\rho_t \to \mathcal{TR}) = \sum_{\tau s' \in \{\tau s\}} P(\rho_t \to \mathcal{TR} \mid \tau s') P(\tau s'). \qquad (11)$$

We select the optimal tactic sequence τs^* as

$$\tau s^* = \arg \max_{\tau s \in \{\tau s\}} P(\tau s \mid \rho_t \to \mathcal{TR}). \qquad (12)$$

3.3.4 *Belief updates*

The shaping robot R learns to influence an adversary R' by updating the expected responses and region reachability distribution. Through these updates, the sequence sampling and selection procedures are biased to favour samples that more closely account for the observed behaviour of R'.

-*Learning adversary responses:* When executing a tactic τ, R observes the responses of R' and uses them to update the set $\tau.\{\tilde{\mathbf{r}}\}$. The tactic time interval, $\tau.dt$, is divided into $\lceil \tau.dt/\bar{\mathbf{n}} \rceil$ segments, and the observed state change of R' is recorded for each segment. If t_1, t_2 are the times at the endpoints of a segment σ, the candidate response $\tilde{m}$ for σ is

$$\tilde{m} = s'_{t_2} - s'_{t_1}. \qquad (13)$$

Given the bound on the maximum number of expected responses per tactic, $\boldsymbol{m}$, if $\tau.\{\tilde{\mathbf{r}}\}$ already has $\boldsymbol{m}$ samples, the oldest sample is replaced by $\tilde{m}$. Otherwise, $\tilde{m}$ is simply appended to the set. Through this procedure, $\tau.\{\tilde{\mathbf{r}}\}$ is biased to reflect the most recently observed reactions of R' to τ.

Adversarial responses model the *local* reactive behaviour of R', without making explicit assumptions about the long-term reasoning or strategic intentionality of that agent. These effects are implicitly addressed by the iterated predictions and expectations of shaping regions, which model the compliance of R' with a temporally extended sequence of actions.

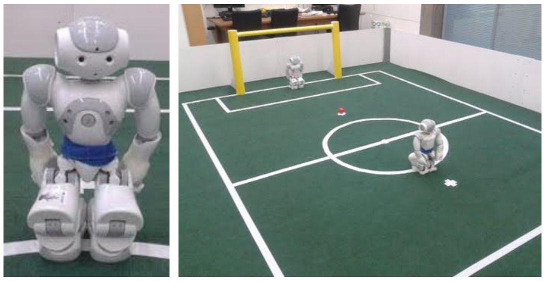
-*Learning region reachability:* Upon completion of a tactic τ, R updates $\mathcal{RD}$ based on the resulting region ρ_c. If τ was invoked from region ρ_i with the intention of reaching ρ', we update the probability $P(\rho | \rho_i, \tau, \rho')$ based on the rule

$$P(\rho | \rho_i, \tau, \rho') = \begin{cases} P(\rho | \rho_i, \tau, \rho') + w, & \rho = \rho_c \\ P(\rho | \rho_i, \tau, \rho') - \frac{w}{|\mathcal{ISR}|-1}, & \forall \rho \neq \rho_c \end{cases} \quad (14)$$

where $0 < w < 1$ is the update weight. Probabilities are also normalised after each weight update. Thus, the distribution progressively assigns higher weight to region-tactic-region transitions that are empirically found to be reachable.

-*Tactic sequence update frequency:* The *update interval*, $\mathbf{u}$, is the number of tactics after which a new sequence should be selected, based on the updated beliefs. For $\mathbf{u} = 1$, a new τs will be selected upon completion of every tactic.

4. EXPERIMENTAL SETUP



(a) (b)

Figure 5: Penalty game setup. (a): The NAO humanoid robot. (b): Soccer field and initial poses of the striker (near side) and the goalkeeper (far side).

4.1 Robot, field, and pose estimation

We use the NAO robot [3] (Figure 5(a)), the official robot of the RoboCup Standard Platform League (SPL) [1]. Our software is based on the B-Human code release [12], which provides modules for fast walking, vision, and self-localisation.

The soccer field (Figure 5(b)) is a scaled version of the official SPL field [4]. The goals are 1.40m wide and are painted yellow, and the ball is also colour-coded orange. Robots are not provided with any external sensory information (e.g. positions from overhead cameras). Instead, they use on-board cameras to identify relevant features (e.g. field lines) and compute their absolute pose through a particle filter.

Egocentric estimation of the pose of other robots is challenging for NAO robots due to visual processing constraints. To overcome this problem, robots wirelessly exchange their pose estimates, which serves as a proxy for good visual sensing. A drawback of this method is that network delays may yield outdated information on the state of the adversary.

Teleoperated robots are controlled through an Xbox pad (Figure 3(a)). There are commands for controlling the translational and rotational motion of the robot, kicking the ball (striker only), and "diving" to save the ball (goalkeeper only).

4.2 Strategic interaction: penalty shootout

Our strategic interaction is a soccer penalty shootout (Figure 5(b)) between a striker (shaping robot) and a goalkeeper (shapeable agent). The game loosely follows the rules of the SPL penalty shootout [4]. The striker has one minute to score and is allowed one kick per trial. The goalkeeper is not allowed to move outside the penalty box. Strikers have a single, straight kick they may execute; thus, to shoot towards goal edges, they must adjust their orientation accordingly.

The challenge for the striker is to deceive the goalkeeper into moving to a different side of the goal than the one it is going to shoot towards, thus maximising its chances of scoring. Thus, in the context of interaction shaping, the striker must select appropriate tactic sequences that can lead the goalkeeper to regions from which a shot cannot be blocked.

5. EXPERIMENTAL RESULTS

5.1 Shaping region and tactic computation

The shaping templates were learned from demonstration by subjects with prior experience of the NAO robots. Demonstrators provided traces of the desired behaviour by controlling the striker against a heuristic autonomous goalkeeper (HAG). The HAG algorithm is summarised as follows: given the striker's current orientation, θ, the expected ball trajectory is a straight line segment starting at the ball position, and following this angle. Then, the HAG moves to the point where this segment intersects the goal line, as the expected best blocking position. The HAG may also dive to block the ball when it detects it to move towards the goal line.

For each trace, we recorded the demonstrator inputs (dx, $dy, d\theta$) motion and kick commands) and the poses of the robots. In total, 29 successful trials were retained. These traces (Figure 6(a)) are characterised by a high intensity of motion around the penalty mark, which indicates an attempt to adjust the striker's pose and deceive the HAG.

The collected data yielded a total of 134 shaping regions and 320 tactics, computed as in Section 3.2. Figure 6(b) shows the means of the computed regions, whereas Figure 6(c) indicates how tactics can be chained to form a sequence.

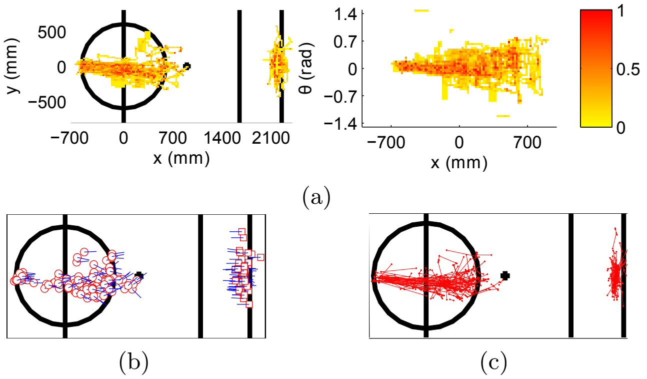
Target regions represent joint states from which the striker is likely to score. The striker (R) should be within the kicking distance of 190mm from the penalty mark, $pm = [px, py]$, where the ball is placed, and the goalkeeper (R') should be on the goal side opposite the striker's orientation, so that $\theta \cdot y' < 0$. Thus, R seeks to reach (one of) the regions

$$\mathcal{TR} = \{\rho \mid d(\rho.\mu.[x : y], pm) < 190, \rho.\mu.\theta \cdot \rho.\mu.y' < 0\} \quad (15)$$

5.2 Shaping agent evaluation

The evaluation of the autonomous interaction shaping striker (ISS) is twofold. First, we compare this agent ***with*** several human-controlled strikers (HCSs), in interactions with the HAG. Our aim is to compare the performance of these two agent types when they compete against the same adversary. Then, we evaluate the ISS ***against*** a more challenging human-controlled adversarial goalkeeper. Here, we assess how the interaction shaping performance is affected when the adversary is a truly strategic, human-controlled adversary, whose exact behavioural model is not known a priori.

To this end, we conduct three different experiments. First, we evaluate the performance of 30 human subjects, in 5 trials each, acting as strikers against the HAG. Our sample was varied, consisting of male and female subjects, children and adults, with subjects also having varying prior experience of robots. Second, we pit the ISS against the same adversary (HAG), in 10 independent sets of 25 trials. Third, we repeat the procedure of the second experiment, but we now pit the ISS against an expert human-controlled goalkeeper (EHCG). This robot is teleoperated by an experienced member of our research group, who is aware of the aims of the experiment.



(a)

(b)

(c)

Figure 6: Learning shaping templates. (a): Heat map representation of successful demonstrations. Colour indicates percentage of trials in which a point was recorded. *Left:* (x) - (y) motion, both players. *Right:* (x) - (θ) motion, striker. (b): Means of computed regions. Circles: striker states. Squares: goalkeeper states. Lines: orientations. A region mean comprises *both* a striker and a goalkeeper state. (c): Tactic graph – edges represent desired transitions between input and target regions.

The EHCG is a considerably harder adversary for two reasons. First, the human operator has *full visibility* of the environment through his own eyes, as opposed to autonomous robots that rely on their noisy camera feed. Second, the operator can learn to anticipate adversarial moves over time, in contrast to the HAG which has a fixed, non-adaptive behaviour. Thus, against the EHCG, the ISS must learn to shape interactions with another learning adversarial agent.

In the last two experiments, the ISS updates adversarial responses and region reachabilities using the parameters $N_S = 20$, $L_S = 10$, $\beta = 0.1, \gamma = 0.7, w = 0.1, \mathbf{u} = 1$.

Interaction (Striker vs Goalkeeper)	**HCSs vs HAG**	**ISS vs HAG**	**ISS vs EHCG**
Total goals scored	61/150	138/250	92/250
Mean striker success rate	40.67%	55.20%	36.80%
Standard deviation	± 20.60%	±5.72%	±6.67%

Table 1: Overall results. HCSs: Human-Controlled Strikers. ISS: Autonomous Interaction Shaping Striker. HAG: Heuristic Autonomous Goalkeeper. EHCG: Expert Human-Controlled Goalkeeper.

The overall results are shown in Table 1. When competing against the HAG, the ISS performs considerably better than the *mean* HCS. Furthermore, when the standard deviation is taken into account, the success rate of the ISS is found to be comparable to the *best* instances of HCS (around 60%). This suggests that the shaping template formulation and learning procedure can successfully generate strategic behaviours that match the sophistication of experienced human users. By contrast, the shaping ability of the ISS drops considerably against the more challenging EHCG, as indicated by the reduced success rate, which is however still comparable to the mean rates achieved by HCSs against the HAG.

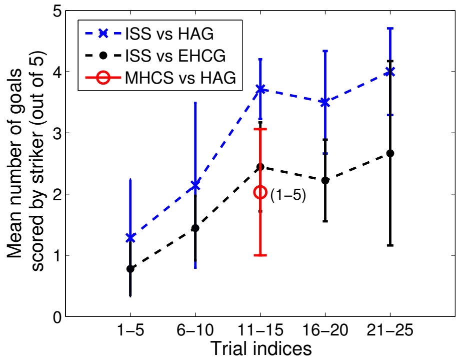


Figure 7: Inter-trial performance of the ISS. Each experimental run of 25 trials is split into blocks of 5, with results averaged over all 10 runs. The mean HCS success rate (MHCS), as averaged over the 5 trials taken by each of the 30 subjects, is also given.

To better understand how the ISS learns to shape interactions over time, we divided the sets of 25 trials of the second and third experiments into blocks of 5, and we measured the mean number of goals scored in each block. Thus, we seek to assess how the performance of the ISS varies across these blocks. The resulting scores are shown in Figure 7. Despite the discrepancy in the number of goals scored against the two adversaries, we observe that the overall progression rate is similar. In both cases, the ISS begins with a low success rate, which improves as interaction progresses. This is an important result demonstrating that the learning rate of our algorithm is not affected by the strategic sophistication of the adversary. Thus, even when the ISS is pitted against an adversary controlled by an expert human operator, it can empirically learn strategies that improve its success rate.

Despite their relation to overall performance, goal-scoring rates do not reflect the strategies used by the shaping robots. To address this, we measured the *distance of the goalkeeper from the optimal blocking position*, d^* (Figure 8(a)). Through this metric, we model how well goalkeepers were influenced into moving to a suboptimal position. A good shaping strategy should succeed in maximising d^* at the end of a trial.

As shown in Figures 8(b)-8(d), the ISS was more successful at maximising d^* than most HCSs, thus more explicitly trying to shape interactions. Moreover, in both ISS experiments, the dominant pattern is that d^* is initially small, reaching its maximum value around the midpoint of the trial and then dropping again. However, when competing against the EHCG, d^* drops more sharply towards the end. This indicates that the expert user is more adept at recovering from deceptive moves by the striker than the HAG, thus preventing the interaction from being shaped at his expense.

Furthermore, Figure 9 shows snapshots from two trials of the ISS against the HAG. In both cases, the ISS first turns to face the far side of the goal, before turning to the near side and shooting. However, in the successful attempt, the ISS waits longer during the first turn, in order to make the HAG move closer to the far side and reduce its subsequent recovery time. Thus, d^* is greater at the end of the trial, and the ISS manages to shape the interaction more effectively.

Further examples of ISS trials, including attempts against the EHCG, are available in the supporting video [2].

6. CONCLUSIONS

We present a framework for strategic interaction shaping in mixed robotic environments. Our approach combines offline learning of shaping regions and tactics from human demonstrations, and online learning and synthesis of these templates through Bayesian inference over the adversary's expected behaviour. Experimental results demonstrate that the shaping agent can shape interactions with a given heuristic adversary comparably to the best human subjects, as identified from a diverse group of 30 individuals. Moreover, the shaping agent can successfully learn, through repeated interaction, to improve its performance against a challenging, human-controlled adversary, who is empirically shown to be less susceptible to deceptive behaviours. Thus, our work constitutes a novel, practical approach to online strategic learning in physical robotic systems, in interactions with unknown, potentially human-controlled, adversarial agents.

We are currently extending our approach towards applications involving more than two robots, and potentially featuring both cooperative and competitive elements. Our aim is to develop autonomous shaping agents who can collaborate with other agents, possibly human-controlled, to influence interactions with strategic adversaries. We view this as an important step towards the realisation of practical, physically grounded, mixed-initiative robotic systems.

7. ACKNOWLEDGMENTS

This work has taken place in the Robust Autonomy and Decisions (RAD) group, School of Informatics, University of Edinburgh. The RAD Group is supported in part by grants from the UK Engineering and Physical Sciences Research Council (EP/H012338/1), the European Commission (TOMSY Grant 270436, FP7-ICT-2009.2.1 Call 6) and a Royal Academy of Engineering *Ingenious* grant. Aris Valtazanos has been supported by a doctoral studentship from the Research Consortium in Speckled Computing, funded by the Scottish Funding Council (Strategic Research Development Programme R32329) and EPSRC (Basic Technology Programme C523881).

8. REFERENCES

[1] RoboCup Standard Platform League. `http://www.tzi.de/spl`.

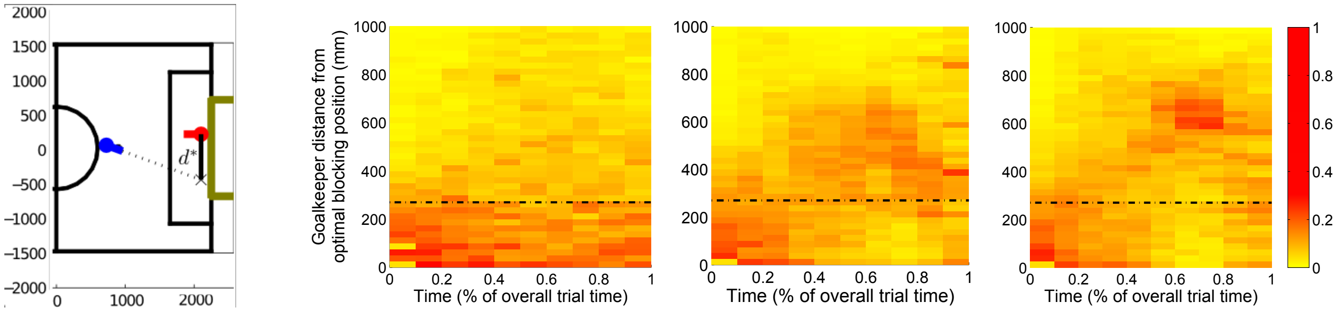
[2] Supporting video. `http://www.youtube.com/watch?v=5rYVhHZzHQQ`.

[3] NAO robot. `http://www.aldebaran-robotics.com/`.

[4] Standard Platform League Rule Book. `http://www.tzi.de/spl/pub/Website/Downloads/Rules2012.pdf`.

[5] D. S. Bernstein, S. Zilberstein, and N. Immerman. The complexity of decentralized control of Markov decision processes. In *UAI*, 2000.

[6] R. R. Burridge, A. A. Rizzi, and D. E. Koditschek. Sequential composition of dynamically dexterous robot behaviors. *I.J. Robotics Research*, 18(6):534–555, 1999.



(a) (b) HCSs vs HAG (c) ISS vs HAG (d) ISS vs EHCG

Figure 8: Goalkeeper distances from optimal blocking position, d^*. (a): Explanation of metric – optimal position for goalkeeper is the intersection of line formed by striker's orientation, and goalkeeper's line of motion. (b)-(c)-(d): Time-indexed heat maps of distances - colour indicates percentage of trials in which a particular time-distance pair was recorded. The black dotted line (d = 270mm) shows the expected minimum distance required to score – this is the length covered by the goalkeeper's leg after a dive to save the ball.

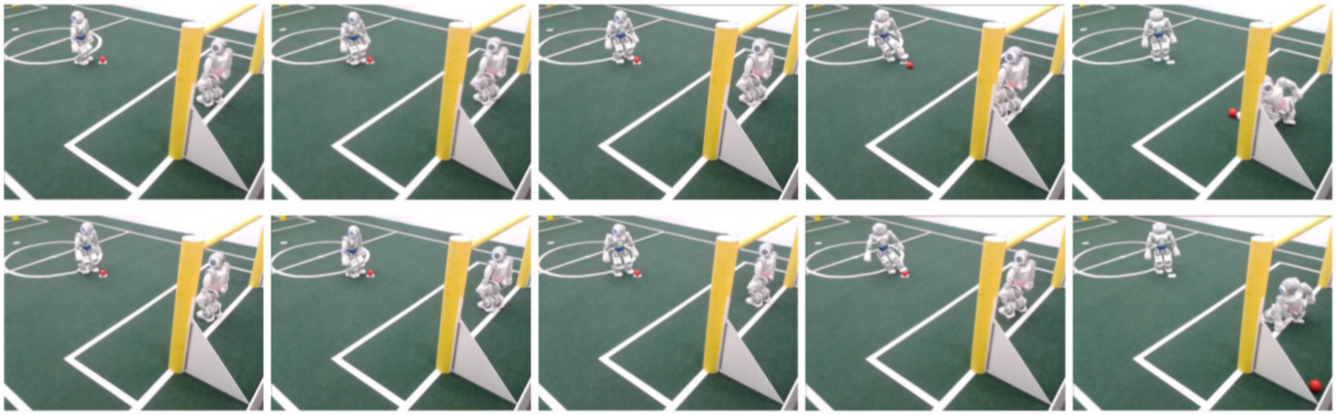


Figure 9: Snapshots from two trials, ISS vs HAG. *Top*: Unsuccessful attempt. *Bottom*: Successful attempt. The two strategies are similar, but in the second trial, the ISS waits longer for the HAG to move towards the far side of the goal (2nd-3rd snapshots), before turning to shoot towards the near side (4th-5th snapshots). Thus, the HAG is deceived into having less time to respond, and the interaction is shaped more effectively.

[7] P. Doshi and P. J. Gmytrasiewicz. Monte carlo sampling methods for approximating interactive POMDPs. *Journal of Artificial Intelligence Research*, 34:297–337, 2009.

[8] P. J. Gmytrasiewicz and P. Doshi. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:24–49, 2005.

[9] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

[10] H. Kurniawati, D. Hsu, and W. S. Lee. SARSOP: Efficient point-based POMDP planning by approximating optimally reachable belief spaces. In *RSS*, 2008.

[11] A. Y. Ng, D. Harada, and S. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, 1999.

[12] T. Röfer, T. Laue, J. Müller, A. Fabisch, F. Feldpausch, K. Gillmann, C. Graf, T. J. de Haas, A. Härtl, A. Humann, D. Honsel, P. Kastner, T. Kastner, C. Könemann, B. Markowsky, O. J. L. Riemann, and F. Wenk. B-human team report and code release 2011, 2011. `http://www.b-human.de/downloads/bhuman11_coderelease.pdf`.

[13] F. Southey, M. Bowling, B. Larson, C. Piccione, N. Burch, D. Billings, and C. Rayner. Bayes' bluff: Opponent modelling in poker. In *UAI*, 2005.

[14] R. S. Sutton, D. Precup, and S. P. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1-2):181–211, 1999.

[15] R. Tedrake. LQR-trees: Feedback motion planning on sparse randomized trees. In *RSS*, 2009.

[16] S. Thrun. Monte carlo POMDPs. In *NIPS*, 2000.

[17] A. R. Wagner and R. C. Arkin. Acting deceptively: Providing robots with the capacity for deception. *I. J. Social Robotics*, 3(1):5–26, 2011.

[18] M. Wunder, M. Kaisers, J. R. Yaros, and M. Littman. Using iterated reasoning to predict opponent strategies. In *AAMAS*, 2011.

[19] H. Zhang and D. Parkes. Value-Based Policy Teaching with Active Indirect Elicitation. In *AAAI*, 2008.