Intelligent Affect: Rational Decision Making for Socially Aligned Agents

Nabiha Asghar and Jesse Hoey
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, CANADA
{nasghar,jhoey}@cs.uwaterloo.ca

Abstract

Affect Control Theory (ACT) is a mathematical model that makes accurate predictions about human behaviour across a wide range of settings. The predictions, which are derived from statistics about human actions and identities in real and laboratory environments, are shared *prescriptive* and *affective* behaviours that are believed to lead to solutions to everyday cooperative problems. A generalisation of ACT, called *BayesAct*, allows the principles of ACT to be used for human-interactive agents by combining a probabilistic version of the ACT dynamical model of affect with a utility function encoding external goals. Planning in *BayesAct*, which we address in this paper, then allows one to go beyond the affective prescription, and leads to the emergence of more complex interactions between "cognitive" and "affective" reasoning, such as deception leading to manipulation and altercasting. We use a continuous variant of a successful Monte-Carlo tree search planner (POMCP) that dynamically discretises the action and observation spaces while planning. We give demonstrations on two classic two-person social dilemmas.

1 INTRODUCTION

BayesAct [4, 20, 21, 22] is a partially-observable Markov decision process (POMDP) model of affective interactions between a human and an artificial agent. *BayesAct* is based upon a sociological theory called "Affect Control Theory" (ACT) [16], but generalises this theory by modeling affective states as probability distributions, and allowing decision-theoretic reasoning about affect. *BayesAct* posits that humans will strive to achieve consistency in shared affective cultural sentiments about events, and will seek to increase *alignment* (decrease *deflection*) with other agents (including artificial ones). Importantly, this need to align implicitly defines an affective heuristic (a *prescription*[1]) for making decisions quickly within interactions. Agents with sufficient resources can do further planning beyond this prescription, possibly allowing them to manipulate other agents to achieve individual profit in collaborative games.

BayesAct arises from the symbolic interactionist tradition in sociology and proposes that humans learn and maintain a set of *shared* cultural affective *sentiments* about people, objects, behaviours, and about the dynamics of interpersonal events. Humans use a simple affective mapping to appraise individuals, situations, and events as sentiments in a three dimensional vector space of evaluation (good vs. bad), potency (strong vs. weak) and activity (active vs. inactive). These mappings can be measured, and the culturally shared consistency has repeatedly been demonstrated to be extremely robust in large cross-cultural studies [17, 29]. Many believe this consistency "gestalt" is a keystone of human intelligence. Humans use it to make predictions about what others will do, and to guide their own behaviour. The shared sentiments, and the resulting *affective ecosystem* of vector mappings, encodes a set of social prescriptions that, if followed by all members of a group, results in an equilibrium or *social order* [14] which is optimal for the group as a whole, rather than for individual members. Humans living at the equilibrium "feel" good and want to stay there. The evolutionary consequences of this individual need are beneficial for the species.

Nevertheless, humans are also a curious, crafty and devious bunch, and often use their cortical processing power to go beyond these prescriptions, finding individually beneficial strategies that are still culturally acceptable, but that are not perfectly normative. This delicate balance is maintained by evolution, as it is beneficial for the species to avoid foundering within a rigid set of rules. In this paper, starting from the principles of *BayesAct*, we investigate how planning beyond cultural prescriptions can result in deceptive or manipulative strategies in two-player social dilemma games. To handle the continuous state, action and observation spaces in *BayesAct*, we use a Monte-Carlo tree

[1] We prefer *prescription*, but also use *norm*, although the latter must not be mis-interpreted as logical rules (see Section 5).

search (MCTS) algorithm that dynamically clusters observations and actions, and samples actions from the *BayesAct* prescriptions as a distribution over the action space.

This paper makes two contributions. First, it describes how to use MCTS planning in *BayesAct*, and gives arguments for why this is an appropriate method. This idea was only hinted at in [22]. Second, it shows how this planning can lead to realistic and manipulative behaviours in the *prisoner's dilemma* and *battle of the sexes* games.

2 BACKGROUND

2.1 Partially Observable Markov Decision Processes

A partially observable Markov decision process (POMDP) [1] is a stochastic control model that consists of a finite set $\mathcal{S}$ of states; a finite set $\mathcal{A}$ of actions; a stochastic transition model $\Pr : S \times A \to \Delta(S)$, with $\Pr(s'|s,a)$ denoting the probability of moving from state s to s' when action a is taken, and $\Delta(S)$ is a distribution over $\mathcal{S}$; a finite observation set $\mathbf{\Omega_s}$; a stochastic observation model, $\Pr(\omega_s|s)$, denoting the probability of making observation $\omega_s \in \mathbf{\Omega_s}$ while the system is in state s; and a reward assigning $R(a,s')$ to a transition to s' induced by action a. A *policy* maps *belief states* (i.e., distributions over $\mathcal{S}$) into actions, such that the expected discounted sum of rewards is (approximately) maximised. We use *factored* POMDPs in which the state is represented by the cross-product of a set of variables or features. POMDPs have been used as models for many human-interactive domains, including assistive technologies [19].

2.2 Affect Control Theory

Affect Control Theory (ACT) arises from work on the psychology and sociology of human social interaction [16]. ACT proposes that social perceptions, behaviours, and emotions are guided by a psychological need to minimize the differences between culturally shared fundamental affective sentiments about social situations and the transient impressions resulting from the interactions between elements within those situations. Fundamental sentiments, $\mathbf{f}$, are representations of social objects, such as interactants' identities and behaviours, as vectors in a 3D affective space, hypothesised to be a universal organising principle of human socio-emotional experience [29]. The basis vectors of affective space are called Evaluation/valence, Potency/control, and Activity/arousal (EPA). EPA profiles of concepts can be measured with the *semantic differential*, a survey technique where respondents rate affective meanings of concepts on numerical scales with opposing adjectives at each end (e.g., good, nice vs. bad, awful for E, weak, little vs. strong, big for P, and calm, passive vs. exciting, active for A). Affect control theorists have compiled lexicons of a few thousand words along with average EPA ratings obtained from survey participants who are knowledgeable about their culture [17]. For example, most English speakers agree that professors are about as nice as students (E), more powerful (P) and less active (A). The corresponding EPAs are $[1.7, 1.8, 0.5]$ for professor and $[1.8, 0.7, 1.2]$ for student[2]. In Japan, professor has the same P (1.8) but students are seen as less powerful (-0.21).

The three dimensions were found by Osgood to be extremely robust across time and cultures. More recently these three dimensions are also thought to be related directly to intrinsic reward [12]. That is, it seems that reward is assessed by humans along the same three dimensions: Evaluation roughly corresponds with expected value, Potency with risk (e.g. powerful things are more risky to deal with, because they do what they want and ignore you), and Activity corresponds roughly with uncertainty, increased risk, and decreased values (e.g. faster and more excited things are more risky and less likely to result in reward) [12]. Similarly, Scholl argues that the three dimensions are in correspondence with the major factors governing choice in social dilemmas [33]. Evaluation is a measure of affiliation or correspondence between outcomes: agents with similar goals will rate each other more positively. Potency is a measure of dependence: agents who can reach their goals independently of other agents are more powerful. Activity is a measure of the magnitude of dependence: agents with bigger payoffs will tend to be more active.

Social events can cause transient impressions, $\boldsymbol{\tau}$ (also three dimensional in EPA space) of identities and behaviours that may deviate from their corresponding fundamental sentiments, $\mathbf{f}$. ACT models this formation of impressions from events with a grammar of the form actor-behaviour-object. Consider for example a professor (actor) who yells (behaviour) at a student (object). Most would agree that this professor appears considerably less nice (E), a bit less potent (P), and certainly more aroused (A) than the cultural average of a professor. Such transient shifts in affective meaning caused by specific events are described with models of the form $\boldsymbol{\tau}' = \boldsymbol{M}\mathscr{G}(\mathbf{f}', \boldsymbol{\tau})$, where $\boldsymbol{M}$ is a matrix of statistically estimated prediction coefficients from empirical impression-formation studies and $\mathscr{G}$ is a vector of polynomial features in $\mathbf{f}'$ and $\boldsymbol{\tau}$. In ACT, the weighted sum of squared Euclidean distances between fundamental sentiments and transient impressions is called *deflection*, and is hypothesised to correspond to an aversive state of mind that humans seek to avoid. This *affect control principle* allows ACT to compute *prescriptive* actions for humans: those that minimize the deflection. Emotions in ACT are computed as a function of the difference between fundamentals and transients [16], and are thought to be communicative signals of vector deflection that help maintain alignment between cooperative agents. ACT has been shown to be highly accurate in explaining verbal behaviours of mock leaders in a computer-simulated business [34], and group dynamics [18], among others [27].

[2] All EPA labels and values in the paper are taken from the Indiana 2002-2004 ACT lexicon [17]. Values range by historical convention from -4.3 to $+4.3$.

2.3 Bayesian Affect Control Theory

Recently, ACT was generalised and formulated as a POMDP for human-interactive artificially intelligent systems [22]. This new model, called *BayesAct*, generalises the original theory in three ways. First, sentiments and impressions are viewed as probability distributions over latent variables (e.g., $\mathbf{f}$ and $\boldsymbol{\tau}$) rather than points in the EPA space, allowing for multimodal, uncertain and dynamic affective states to be modeled and learned. Second, affective interactions are augmented with *propositional* states and actions (e.g. the usual state and action space considered in AI applications). Third, an explicit reward function allows for goals that go beyond simple deflection minimization. We give a simplified description here; see [21, 22] for details.

A *BayesAct* POMDP models an interaction between two agents (human or machine) denoted *agent* and *client*. The state, $\boldsymbol{s}$, is the product of six 3-dimensional continuous random variables corresponding to fundamental and transient sentiments about the *agent*'s identity $(\mathbf{F}_a, \mathbf{T}_a)$, the current (*agent* or *client*) behaviour $(\mathbf{F}_b, \mathbf{T}_b)$ and the *client*'s identity $(\mathbf{F}_c, \mathbf{T}_c)$. We use $\mathbf{F} = \{\mathbf{F}_a, \mathbf{F}_b, \mathbf{F}_c\}$ and $\mathbf{T} = \{\mathbf{T}_a, \mathbf{T}_b, \mathbf{T}_c\}$. The state also contains an application-specific set of random variables $\mathbf{X}$ that are interpreted as *propositional* (i.e. not *affective*) elements of the domain (e.g. whose turn it is, game states - see Section 4), and we write $\boldsymbol{s} = \{\mathbf{f}, \boldsymbol{\tau}, \mathbf{x}\}$. Here the *turn* is deterministic (*agent* and *client* take turns), although this is not necessary in *BayesAct*. The *BayesAct* reward function is application-specific over $\mathbf{x}$. The state is not observable, but observations $\boldsymbol{\Omega_x}$ and $\boldsymbol{\Omega_f}$ are obtained for $\mathbf{X}$ and for the affective behaviour $\mathbf{F}_b$, and modeled with probabilistic observation functions $Pr(\boldsymbol{\omega_x}|\mathbf{x})$ and $Pr(\boldsymbol{\omega_f}|\mathbf{f}_b)$, respectively.

Actions in the *BayesAct* POMDP are factored in two parts: $\mathbf{b_a}$ and a, denoting the *affective* and *propositional* components, respectively. For example, if a tutor gives a hard exercise to do, the manner in which it is presented, and the difficulty of the exercise, combine to form an affective impression $\mathbf{b_a}$ that is communicated. The actual exercise (content, difficulty level, etc) is the *propositional* part, a.

The state dynamics factors into three terms as $Pr(\boldsymbol{s}'|\boldsymbol{s},\mathbf{b_a},a) = Pr(\boldsymbol{\tau}'|\boldsymbol{\tau},\mathbf{f}',\mathbf{x}) Pr(\mathbf{f}'|\mathbf{f},\boldsymbol{\tau},\mathbf{x},\mathbf{b_a}) Pr(\mathbf{x}'|\mathbf{x},\mathbf{f}',\boldsymbol{\tau}',a)$, and the fundamental behaviour, $\mathbf{F}_b$, denotes either observed *client* or taken *agent* affective action, depending on whose *turn* it is (see below). That is, when *agent* acts, there is a deterministic mapping from the affective component of his action $(\mathbf{b_a})$ to the *agent*'s behaviour $\mathbf{F}_b$. When *client* acts, *agent* observes $\boldsymbol{\Omega_f}$ (the affective action of the other agent). The third term in the factorization of the state dynamics is the *Social Coordination Bias*, and is described in Section 2.4. Now we focus on the first two terms.

The transient impressions, $\mathbf{T}$, evolve according to the impression-formation operator in ACT ($\boldsymbol{M\mathscr{G}}$), so that $Pr(\boldsymbol{\tau}'|...)$ is deterministic. Fundamental sentiments are expected to stay approximately constant over time, but are subject to random drift (with noise $\boldsymbol{\Sigma}_f$) and are expected to also remain close to the transient impressions because of the *affect control principle*. Thus, the dynamics of $\mathbf{F}$ is[3]:

$$Pr(\mathbf{f}'|\mathbf{f}, \boldsymbol{\tau}) \propto e^{-\psi(\mathbf{f}', \boldsymbol{\tau}) - \xi(\mathbf{f}', \mathbf{f})} \tag{1}$$

where $\psi \equiv (\mathbf{f}' - \boldsymbol{M\mathscr{G}}(\mathbf{f}', \boldsymbol{\tau}))^T \boldsymbol{\Sigma}^{-1} (\mathbf{f}' - \boldsymbol{M\mathscr{G}}(\mathbf{f}', \boldsymbol{\tau}))$ combines the *affect control principle* with the impression formation equations, assuming Gaussian noise with covariance $\boldsymbol{\Sigma}$. The inertia of fundamental sentiments is $\xi \equiv (\mathbf{f}' - \mathbf{f})^T \boldsymbol{\Sigma}_f^{-1} (\mathbf{f}' - \mathbf{f})$, where $\boldsymbol{\Sigma}_f$ is diagonal with elements $\beta_a, \beta_b, \beta_c$. The state dynamics are non-linear due to the features in $\mathscr{G}$. This means that the belief state will be non-Gaussian in general, and *BayesAct* uses a *bootstrap filter* [11] to compute belief updates.

The distribution in (1) gives the prescribed (if *agent* turn), or expected (if *client* turn), action as the component $\mathbf{f}'_b$ of $\mathbf{f}'$. Thus, by integrating over $\mathbf{f}'_a$ and $\mathbf{f}'_c$ and the previous state, we obtain a probability distribution, $\pi^\dagger$, over $\mathbf{f}'_b$ that acts as a *normative action bias*: it tells the agent what to expect from other agents, and what action is expected from it in belief state $b(\boldsymbol{s})$:

$$\pi^\dagger(\mathbf{f}'_b) = \int_{\mathbf{f}'_a, \mathbf{f}'_c} \int_{\boldsymbol{s}} Pr(\mathbf{f}'|\mathbf{f}, \boldsymbol{\tau}, \mathbf{x}) b(\boldsymbol{s}) \tag{2}$$

2.4 *BayesAct* Instances

As affective identities $(\mathbf{f}_a, \mathbf{f}_c)$ are latent (unobservable) variables, they are learned (as inference) in the POMDP. If behaving normatively (according to the *normative action bias*), an agent will perform affective actions $\mathbf{b_a} = \arg\max_{\mathbf{f}'_b} \pi^\dagger(\mathbf{f}'_b)$ that allow other agents to infer what his (true) identity is. The *normative action bias* (NAB) defines an affective signaling mechanism as a shared set of prescriptions for translating information about identity into messages. In *BayesAct*, the NAB is given by Equation (2).

The NAB is only prescriptive: all agents are free to select individually what they really send, allowing for deception (e.g. "faking" an identity by sending incorrect information in the affective dimension of communication). Possible outcomes are manipulation (the other agent responds correctly, as its own identity, to the "fake" identity), and altercasting (the other agent assumes a complementary identity to the faked identity, and responds accordingly), both possibly leading to gains for the deceptive agent.

The dynamics of $\mathbf{X}$ is given by $Pr(\mathbf{x}'|\mathbf{f}', \boldsymbol{\tau}', \mathbf{x}, a)$, that we refer to as the *social coordination bias* (SCB): it defines what agents are expected to do (how the state is expected to change, including other agents' propositional behaviours) in a situation $\mathbf{x}$ when action a was taken that resulted in sentiments $\mathbf{f}'$ and $\boldsymbol{\tau}'$. For example, we may expect faster student learning if deflection is low, as cognitive resources do not need to be spent dealing with mis-alignment.

The SCB is a set of shared rules about how agents, when acting normatively, will behave *propositionally* (action a,

[3] We leave out the dependence on $\mathbf{x}$ for clarity, and on $\mathbf{b_a}$ since this is replicated in $\mathbf{f}'_b$.

as opposed to affectively with action $\mathbf{b_a}$). Assuming identities are correctly inferred (as insured by the shared nature of the NAB), each agent can both recognize the type of the other agent and can thereby uncover an optimistic policy[4] that leads to the normative mean accumulated future reward (as defined by the social coordination bias). However, with sufficient resources, an agent can use this prescribed action as a heuristic only, searching for nearby actions that obtain higher individual reward. For example, a teacher who seems very powerful and ruthless at the start of a class, often may choose to do so (in a way that would be inappropriate in another setting, e.g., the home, but is appropriate within the classroom setting) in order to establish a longer-term relationship with her students. The teacher's actions feel slightly awkward if looked at in the context of the underlying social relationship with each student (e.g. as would be enacted according to normative *BayesAct*), but are leading to longer-term gains (e.g. the student passes).

Thus, the NAB (along with a communication mechanism) allows the relaying of information about identity, while the SCB allows agents to make predictions about other agents' future actions *given* the identities. This combination allows agents to assume cooperative roles in a joint task, and is used as an emotional "fast thinking" heuristic (Kahneman's "System 1" [23]). If agents are fully cooperative and aligned, then no further planning is required to ensure goal achievement. Agents do what is expected (which may involve planning over $\mathbf{X}$, but not $\mathbf{F}$ and $\mathbf{T}$), and expect others to as well. However, when alignment breaks down, or in non-cooperative situations, then slower, more deliberative ("System 2") thinking arises. The Monte-Carlo method in Section 3 naturally trades-off slow vs. fast thinking.

3 POMCP-C

POMCP [36] is a Monte-Carlo tree search algorithm for POMDPs that progressively builds a search tree consisting of nodes representing histories and branches representing actions or observations. It does this by generating samples from the belief state, and then propagating these samples forward using a blackbox simulator (the known POMDP dynamics). The nodes in the tree gather statistics on the number of visits, states visited, values obtained, and action choices during the simulation. Future simulations through the same node then use these statistics to choose an action according to the UCB1 formula, which adds an exploration bonus to the value estimate based on statistics of state visits (less well-visited states are made to look more salient or promising). Leaves of the tree are evaluated using a set of *rollouts*: forward simulations with random action selection. The key idea is that fast and rough rollouts blaze the trail for the building of the planning tree, which is more carefully explored using the UCB1 heuristic. POMCP uses a timeout (processor or clock time) providing an anytime solution.

[4]optimistic in the sense that it assumes all agents will also follow the same normative policy.

In our algorithm, POMCP-C, we make use of an *action bias*, π_{heur}: a probability distribution over the action space that guides action choices[5]. In *BayesAct*, we naturally have such a bias: the normative action bias (for $\mathbf{b_a}$) and the social coordination bias (for a). At each node encountered in a POMCP-C simulation (at history h), an action-observation pair is randomly sampled as follows. First, a random sample is drawn from the action bias, $\mathbf{a} \sim \pi_{heur}$. The action $\mathbf{a}$ is then compared to all existing branches at the current history, and a new branch is only created if it is significantly different, as measured by distance in the action space (Euclidean for $\mathbf{b_a}$, binary for a) and a threshold parameter δ_a ('action resolution'), from any of these existing branches. If a new branch is created, the history ha is added to the planning tree, and is evaluated with a rollout as usual. If a new branch is not created, then a random sample o is drawn from the observation distribution $Pr(o|h,a)$[6].

The continuous observation space raises two significant problems. First, the branching factor for the observations is infinite, and no two observations will be sampled twice. To counter this, we use a dynamic discretisation scheme for the observations, in which we maintain $\mathbf{o}(h)$, a set of sets of observations at each history (tree node). So $\mathbf{o}(h) = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_{N_o}\}$, where $N_o \in \mathbb{N}$. A new observation o is either added to an existing set $\mathbf{o}_j$ if it is close enough to the mean of that set (i.e. if $|o - \bar{\mathbf{o}}_\mathbf{j}| < \delta_o$ where δ_o is a constant, the 'observation resolution'), or, if not, it creates a new set $\mathbf{o}_{N_o+1} = \{o\}$. This simple scheme allows us to dynamically learn the observation discretisation.

The second problem raised by continuous observations stems from the fact that POMCP uses a black box simulator that should draw samples from the same distribution as the environment does. Thus, the simulated search tree replicates actual trajectories of belief, and can be re-used after each action and observation in the real world (after each pruning of the search tree). This works for discrete observations, but it may not work for continuous observations since the same observation will rarely be encountered twice. Here, we prune the tree according to the closest observation set $\mathbf{o}_j$ to the observation obtained (see also [4]).

4 EXPERIMENTS AND RESULTS

We present highlights of results on two social dilemmas. Full results and other experiments are in [4].

4.1 Prisoner's Dilemma (Repeated)

The prisoner's dilemma is a classic two-person game in which each person can either *defect* by taking \$1 from a (common) pile, or *cooperate* by giving \$10 from the same pile to the other person. There is one Nash equilibrium in which both players defect, but when humans play the game

[5]The idea of using a heuristic to guide action selection in POMCP was called *preferred actions* [36].

[6]POMCP-C also uses a cut-off N_A^{max} on the branching factor.

they often are able to achieve the optimal solution where both cooperate. A rational agent would first compute the strategy for the game as the Nash equilibrium (of "defect"), and then look up the affective meaning of such an action using e.g. a set of appraisal rules, and finally apply a set of coping rules. For example, such an agent might figure out that the goals of the other agent would be thwarted, and so that he should feel ashamed or sorry for the other agent. However, appraisal/coping theories do not specify the probabilities of emotions, do not take into account the affective identities of the agents, and do not give consistent accounts of how coping rules should be formulated.

Instead, a *BayesAct* agent (called a *pd-agent* for brevity here), computes what *affective* action is prescribed in the situation (given his estimates of his and the other's identities, and of the affective dynamics), and then seeks the best propositional action ($a \in \{cooperate, defect\}$) to take that is consistent with this prescribed affect. As the game is repeated, the *pd-agent* updates his estimates of identity (for self and other), and adjusts his play accordingly. For example, a player who defects will be seen as quite negative, and appropriate affective responses will be to defect, or to cooperate and give a nasty look.

The normative action bias (NAB) for *pd-agents* is the usual deflection minimizing affective $\mathbf{f}_b$ given distributions over identities of *agent* and *client* (Equation 2). Thus, if *agent* thought of himself as a *friend* (EPA:$\{2.75, 1.88, 1.38\}$) and knew the other agent to be a *friend*, the deflection minimizing action would likely be something good (high E). Indeed, a simulation shows that one would expect a behaviour with EPA= $\{1.98, 1.09, 0.96\}$, with closest labels such as *treat* or *toast*. Intuitively, cooperate seems like a more aligned propositional action than defect. This intuition is confirmed by the distances from the predicted (affectively aligned) behaviour to *collaborate with* (EPA:$\{1.44, 1.11, 0.61\}$) and *abandon* (EPA:$\{$-2.28, -0.48, -0.84$\}$) of 0.4 and 23.9, respectively. Table 1 shows all combinations if each agent could also be a *scrooge* (EPA:$\{$-2.15, -0.21, -0.54$\}$). We see that a *friend* would still collaborate with a *scrooge* (in an attempt to *reform* the scrooge), a *scrooge* would abandon a *friend* (*look away from* in shame), and two scrooges would defect.

The *agent* will predict the *client*'s behavior using the same principle: compute the deflection minimising affective action, then deduce the propositional action based on that. Thus, a *friend* would be able to predict that a *scrooge* would defect. If a *pd-agent* has sufficient resources, he could search for an affective action near to his optimal one, but that would still allow him to defect. To get a rough idea of this action, we find the point on the line between his optimal action $\{0.46, 1.14,$ -0.27$\}$ and *abandon* that is equidistant from *abandon* and *collaborate with*. This point, at which he would change from cooperation to defection, is $\{$-0.8, 0.6, -0.4$\}$ (*glare at*), which only has a slightly higher deflection than *reform* (6.0 vs 4.6). Importantly, he is *not trading off costs in the game with costs of disobeying the social prescriptions*: his resource bounds and action search strategy are preventing him from finding the more optimal (individual) strategy, implicitly favoring those actions that benefit the group and solve the social dilemma.

agent	client	optimal behaviour	closest labels	dist. from coll.	dist. from ab.
F	F	$1.98, 1.09, 0.96$	treat toast	0.4	23.9
F	S	$0.46, 1.14,$ -0.27	reform lend money to	1.7	10.5
S	F	-0.26, -0.81, -0.77	curry favor look away	8.5	4.2
S	S	-0.91, -0.80, -0.01	borrow money chastise	9.6	2.7

Table 1: Optimal (deflection minimising) behaviours for two *pd-agents* with fixed identities. F=friend, S=scrooge, coll.=collaborate with, ab.=abandon

PD-agents are dealing with a slightly more difficult situation, as they do not know the identity of the other agent. However, the same principle applies, and the social coordination bias (SCB) is that agents will take and predict the propositional action that is most consistent with the affective action. Agents have culturally shared sentiments about the propositional actions (defection and cooperation), and the distance of the deflection minimizing action (*agent*, $\mathbf{b_a}$) or behaviour (*client*, $\mathbf{f}_b$) to these sentiments is a measure of how likely each propositional action is to be chosen (*agent* turn), or predicted (*client* turn). That is, on *agent* turn, the affective actions $\mathbf{b_a}$ will be sampled and combined with a propositional action a sample drawn proportionally to the distance from $\mathbf{b_a}$ to the shared sentiments for each a. On *client* turn, affective behaviours $\mathbf{f}_b$ will be predicted and combined with a value for a variable representing *client* play in $\mathbf{X}$ drawn proportionally to the distance from $\mathbf{f}_b$.

We model *agent* and *client* as having two (simultaneous) identities: *friend* or *scrooge* with probabilities 0.8 and 0.2, respectively. Each *pd-agent* starts with a mixture of two Gaussians centered at these identities with weights $0.8/0.2$ and variances of 0.1. The SCB interprets cooperation as *collaborate with* (EPA:$\{1.44, 1.11, 0.61\}$) and defection as *abandon* (EPA:$\{$-2.28, -0.48, -0.84$\}$), and the probability of the propositional actions using a Gibbs measure over distance with a variance of 4.0. We use propositional state $\mathbf{X} = \{Turn, Ag\_play, Cl\_play\}$ denoting whose turn it is ($\in \{agent, client\}$) and *agent* and *client* state of play ($\in \{not\_played, cooperate, defect\}$). The agents' reward is only over the game (e.g. 10, 1, or 0), so there is no intrinsic reward for deflection minimization as in [22]. We use a two time-step game in which both *agent* and *client* choose their actions at the first time step, and then communicate this to each other on the second step. The agents also communicate affectively, so that each agent gets to see both what action the other agent took (cooperate or defect), and also *how* they took it (expressed in $\mathbf{f}_b$)[7]. If one were to imple-

[7]Agents may also relay emotions (see Sec. 2.2), but here we only use emotional labels for explanatory purposes.

ment this game in real life, then $\mathbf{f}_b$ would be relayed by e.g. a facial expression. We use a Gaussian observation function $Pr(\boldsymbol{\omega}_{\boldsymbol{f}}|\mathbf{f}_b)$ with mean at $\mathbf{f}_b$ and std. dev. of $\sigma_b = 0.1$. Our simulations consist of 10 trials of 20 games/trial, but agents use an infinite horizon with a discount γ.

We simulate one *pd-agent* (*pdA*) with a POMCP-C (processor time) timeout value of t_a, and the other (*pdC*) either: (1), a similar agent with the same timeout $t_c = t_a$, or with a timeout of $t_c = 1s$; or (2), a fixed strategy agent that plays one of: (co) always cooperate; (de) always defect; (tt) tit-for-tat; (to): two-out; (t2): tit-for-two-tat; (2t): two-tit-for-tat. Except for (de), these fixed strategy agents always cooperate on the first turn, and then: (tt) mirrors the other agent; (to) cooperates twice, then always defects; (t2): defects if the other agent defects twice in a row; (2t): cooperates if the other agent cooperates twice in a row[8]. Fixed strategy agents always relay *collaborate with* and *abandon* as $\mathbf{f}_b$ when playing cooperate and defect, respectively.

First, we consider agents that use the same timeout. In this case, if the discount factor is 0.99, both agents cooperate all the time, and end up feeling like *warm*, *earnest* or *introspective ladies*, *visitors* or *bridesmaids* (EPA$\sim \{2, 0.5, 1.0\}$). This occurs regardless of the amount of timeout given to both agents. Essentially, both agents are following the norm. If they don't have a long timeout, this is all they can evaluate. With longer timeouts, they figure out that there is no better option. However, if the discount is 0.9 (more discounting, so they will find short-term solutions), then again cooperation occurs if the timeout is short (less than 10s), but then one agent starts trying to defect after a small number of games, and this number gets smaller as the timeout gets longer (see Figure 1). With more discounting, more time buys more breadth of search (the *agent* gets to explore more short-term options), and finds more of them that look appealing (it can get away with a defection for a short while). With less discounting, more time buys more depth, and results in better long-term decisions.

Table 2 shows the first five games with a *client* playing two-out (to), who sends affective values of $\{1.44, 1.11, 0.61\}$ and cooperates on the first two moves. This affective action makes the *pd-agent* feel much less good (E) and powerful (P) than he normally would (as a *failure*), as he'd expect a more positive and powerful response (such as *flatter* EPA=$\{2.1, 1.45, 0.82\}$) if he was a *friend*, so this supports his *scrooge* identity more strongly[9]. He infers *client* is friendly (a *newlywed* is like a *girlfriend* in EPA space). He therefore cooperates on the second round, and feels somewhat better. Then, the *client* defects on the third round, to which the *agent* responds by re-evaluating the *client* as less good (an *immoral purchaser*). He still tries to cooperate, but gives up after two more rounds, after which he thinks of the *client* as nothing but a *selfish hussy*, and himself as a *disapproving divorcée*. The *agent* consistently defects after this point. Interactions with (tt), (2t) and (t2) generally follow a similar pattern, because any defection rapidly leads to both agents adopting long-term defection strategies. However, as shown in Table 3 (also see full results [4]), less discounting leads to better solutions against these strategies, as longer-term solutions are found.

γ	(tt)	(t2)	(2t)
0.9	1.64 ± 2.24	3.98 ± 2.48	1.72 ± 2.35
0.99	7.33 ± 1.17	7.28 ± 1.68	7.63 ± 0.91

Table 3: Results (avg. rewards) against the tit-for strategies

When playing against (co), *pd-agents* generally start by cooperating, then defect, resulting in a feeling of being a *self-conscious divorcée* (EPA:$\{$-0.23, -0.62, 0.32$\}$) playing against a *conscientious stepsister* (EPA:$\{0.12, -0.04, 0.35\}$). When playing against (de), *pd-agents* generally start by cooperating, but then defect, feeling like a *dependent klutz* (EPA:$\{$-0.76, -1.26, 0.37$\}$) playing against an *envious ex-boyfriend* (EPA:$\{$-1.30, -0.49, -0.13$\}$).

4.2 Affective Cooperative Robots (CoRobots)

CoRobots is a multi-agent cooperative robot game based on the classic "Battle of the Sexes" problem[10]. We are specifically interested in asymmetrical situations wherein one robot has more resources and can do planning in order to *manipulate* the other robot, taking advantage of the social coordination bias. We start with a simplified version in which the two robots maintain affective fundamental sentiments, but do not represent the transient impressions. The normative action bias is a simple average instead of as the result of more complex impression formation equations.

Concretely, two robots, Rob1 and Rob2, move in a 1D continuous state space. We denote their positions with variables X_1 and X_2. At each time step, Rob1, Rob2 take actions $a_1, a_2 \in \mathbb{R}$ respectively. This updates their respective positions $x_i, i \in \{1, 2\}$ according to $x_i \leftarrow x_i + a_i + \nu_i$ and $\nu_i \sim \mathcal{N}(0, \sigma)$. There are two fixed locations $L_1 \in \mathbb{R}^+$ and $L_2 \in \mathbb{R}^-$. For each robot, one of these locations is the major goal g (with associated high reward r) and the other is the minor goal $\bar{g}$ (with associated low reward $\bar{r}$). A robot is rewarded according to its distance from g and $\bar{g}$, but only if the other robot is nearby. The reward for Rob$i$ is:

$$R_i(x_1,x_2)=\mathbb{I}(|x_1-x_2|<\Delta_x)[r\cdot e^{-(x_i-g)^2/\sigma_r^2}+\bar{r}\cdot e^{-(x_i-\bar{g})^2/\sigma_r^2}], \quad (3)$$

where $\mathbb{I}(y) = 1$ if y is true, and 0 otherwise, and where σ_r is the reward variance, Δ_x is a threshold parameter governing how "close" the robots need to be, and $r, \bar{r} \in \mathbb{R}$, such that $r \gg \bar{r} > 0$. Both σ_r and Δ_x are fixed and known by both robots. Each robot only knows the location of its own major goal. Furthermore, at any time step, each robot

[8](t2) is more "generous", and (2t) is more "wary" than (tt).

[9]Examples of more positive affective actions in [4].

[10]A husband wants to go to a football game, and his wife wants to go shopping, but neither wants to go alone. There are two pure Nash equilibria, but the optimal strategy requires coordination.

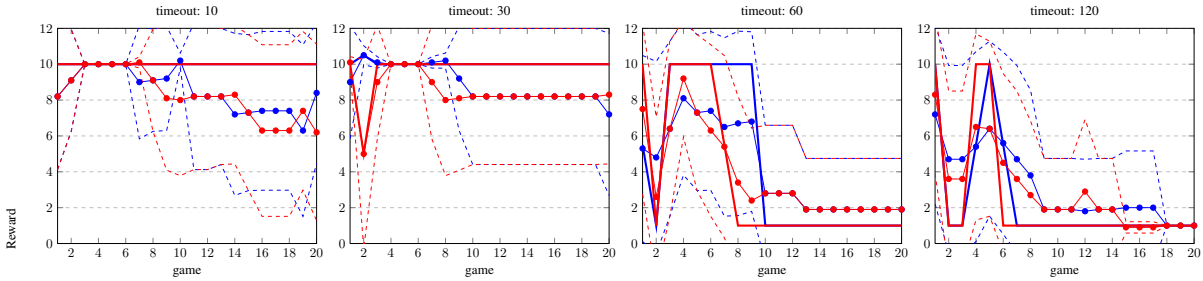

Figure 1: PD with client strategy: (same) and discount $\gamma = 0.9$. Red=client; Blue=agent; dashed=std.dev.; solid (thin, with markers): mean; solid (thick): median. As timeout increases, more defections give less reward for both agents.

game	post-play sentiments (*agent*)			defl-	identities		emotions		actions	
#	$\mathbf{f}_a$	$\mathbf{f}_c$	$\mathbf{f}_b$	ection	agent	client	agent	client	agent	client
1	-1.36,-0.01,-0.35	2.32,1.61,1.27	2.62,1.58,1.73	4.44	failure	newlywed	easygoing	idealistic	coop.	coop.
2	-0.66,0.04,-0.05	1.77,1.27,1.06	2.23,1.00,1.76	3.70	parolee	husband	easygoing	self-conscious	coop.	coop.
3	-0.23,-0.08,0.20	1.02,0.93,0.84	2.49,0.97,1.87	7.19	stepmother	purchaser	female	immoral	coop.	def.
4	-0.12,-0.33,0.33	0.27,0.62,0.62	2.37,0.48,1.34	4.99	stuffed_shirt	roommate	dependent	unfair	coop.	def.
5	-0.26,-0.47,0.32	-0.26,0.26,0.42	-0.59,0.41,-0.23	3.27	divorcée	gun_moll	dependent	selfish	def.	def.
6	-0.37,-0.66,0.26	-0.61,0.00,0.28	-0.10,-0.41,-0.27	2.29	divorcée	hussy	disapproving	selfish	def.	def.

Table 2: Example games with *client* playing (to). Identities and emotions are *agent* interpretations.

can move in any direction, receives observations of the locations of both robots, and has a belief over X_1 and X_2.

In order to coordinate their actions, the robots must relay their reward locations to each other, and must choose a *leader* according to some social coordination bias. The robots each have a 3D *identity* (as *BayesAct*), where the valence, $\mathbf{f}_{ae}$, describes their goal: if $\mathbf{f}_{ae} > 0$, then $g = L_1$. If $\mathbf{f}_{ae} < 0$, then $g = L_2$. The power and activity dimensions will be used for coordination (see below). Robots can move (propositional action a) at any time step, but must coordinate their communications. That is, only one robot can communicate at a time (with affective action $\mathbf{b_a}$ perceived by the other robot as $\boldsymbol{\omega_f}$), but this turn-taking behaviour is fixed. The normative action bias (NAB) in the first (simplified) CoRobots problem is the mean of the two identities:

$$\pi^\dagger \propto \mathcal{N}((\mathbf{f}_a + \mathbf{f}_c)/2, \Sigma_b). \tag{4}$$

In *BayesAct* Corobots, the NAB is given by Equation (2).

The social coordination bias (that the leader will lead) defines each robot's action bias for a_i, and action prediction function (for *client*'s x) through a 2D sigmoid $leader$ function, known to both agents. This sigmoid function is ≥ 0.5 if the *agent* estimates he is more powerful or more active than the *client* ($(\mathbf{f}_{ap} > \mathbf{f}_{cp}) \vee (\mathbf{f}_{aa} > \mathbf{f}_{ca})$) and is < 0.5 otherwise. If the *agent* is the leader, his action bias will be a Gaussian with mean at $+1.0$ in the direction of his major goal (as defined by $\mathbf{f}_{ae}$), and in the direction of the *client*'s major goal (as defined by his estimate of $\mathbf{f}_{ce}$) otherwise. *Agent*'s prediction of *client*'s motion in x is that the *client* will stay put if *client* is the leader, and will follow the *agent* otherwise, as given succinctly by:

$$Pr(x'_c|\mathbf{f}'_a,\mathbf{f}'_c) = \mathcal{N}(\mathbb{I}(leader(\mathbf{f}'_a,\mathbf{f}'_c) \geq 0.5)\lambda_a + x_c, \sigma_p) \tag{5}$$

where $\lambda_a = 1$ if $\mathbf{f}'_{ae} > 0$, and -1 otherwise and $\sigma_p = 1.0$.

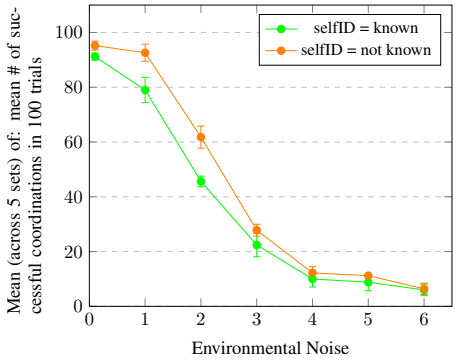

Figure 2: *BayesAct* Corobots cannot coordinate properly when the communication channel is bad or non-existent.

We first investigate whether corobots can coordinate when they have identities drawn from the set of 500 human (male) identities in the ACT lexicon (see footnote 2). In the first experiment, the two identities are selected at random on each trial. Each corobot knows his self-ID ($\mathcal{N}$(self-ID, 0.1)) but does not know the other's ID ($\mathcal{N}$([0.0, 0.0, 0.0], 2.0)). Furthermore, each corobot has a stable self-identity ($\beta_a = 0.1$), but it believes that the other is less stable ($\beta_c = 2.0$). Finally, both corobots have equal POMCP-C planning resources ($\Sigma_b = 0.5, N_A^{max} = 3, \delta_a = 2.0, \delta_o = 6.0$ and $Timeout = 2.0$ seconds). The other CoRobots game parameters are $r = 100, \bar{r} = 30, L_1 = 10, L_2 = -10, \sigma_r = 2.5, \Delta_x = 1.0$ and iterations = 30. We run 5 sets of 100 simulated trials of the CoRobots Game with varying *environmental noise*, i.e., we add a normally distributed value, with standard deviation corresponding to the noise level, to the computation and communication of $\boldsymbol{\Omega_x}$ and $\boldsymbol{\Omega_f}$ (observations of x and $\mathbf{f}$, resp.). Figure 2 (green line) shows the mean and standard error of mean number of successful coordinations by the corobots.

The percentage of successful coordination falls from 91% to 6% when the environmental noise is increased, and the average total reward per trial falls from 1403 to 19.4. We see that with no environmental noise, the corobots are able to easily learn the other's identity, and can coordinate based on the social coordination bias. As the environmental noise increases, corobots are unable to easily relay identities, and require a much longer time to find cooperative solutions.

Figure 2 (orange line) shows results where the self-ID is also unknown initially ($\mathcal{N}([0.0, 0.0, 0.0], 2.0)$), and is less stable ($\beta_a = 2.0$). We see that the general trend is the same; however, the corobots have a higher percentage of successful coordinations, and consequently gain a higher average total reward, for the three lowest noise values. It is surprising to see that the corobots perform better with unknown self-IDs. This is because corobots quickly assume contrasting identities (i.e. one assumes a less powerful identity than the other) in order to coordinate. With known self-IDs, however, the corobots show less flexibility and spend the initial few iterations trying to convince and pull the other corobot towards themselves. Due to this rigidity, these corobots suffer a lot when they have similar power; this does not happen when the self-ID is unknown.

Next, we investigate whether one agent can *manipulate* the other. A manipulation is said to occur when the weaker and less active agent deceives the client into believing that the agent is more powerful or active, thereby persuading the client to converge to the agent's major goal g (to within $\pm|0.2g|$). In order to demonstrate manipulative behaviour, we introduce asymmetry between the two agents by changing the parameters Σ_b, N_A^{max} and $Timeout$ for one agent (unbeknownst to the other). In addition, we allow this agent to start with a slightly better estimate of the other's identity. This agent will then sample actions that are farther from the norm than expected by the other agent, and will allow such an agent to "fake" his identity so as to manipulate the other agent. The agent's and client's self-identities are noisy ($\sigma = 0.1$) versions of $[2.0, -1.0, -1.0]$ and $[-2.0, 1.0, 1.0]$ respectively, $r = 100, \bar{r} = 30, L_1 = 5, L_2 = -5, \Delta_x = 1, \sigma_r = 2.5, \delta_a = 2.0, \delta_o = 6.0, N_A^{max} = 3, \Sigma_b = 0.5$ and $Timeout = 2.0$ for both robots. Each game is set to run for 40 iterations, and starts with the agent and client located at 0.0. Since $g_a = 5$, $g_c = -5$, both robots should converge to $g_c = -5$ (*client* is leader) if following normative actions.

When $N_A^{max} = 3$, $\Sigma_b = 0.5$, and $Timeout = 2.0$ for the agent, the agent displays manipulative behaviour in only $80/1000$ games, as expected (both follow normative behaviour). If we allow the *agent* to start with a better estimate of the *client*'s identity (*agent*'s initial belief about $\mathbf{f}_c$ is a Gaussian with mean $[-2.0, 1.0, 1.0]$ and variance 1.0), we see manipulative behaviour in almost twice as many games (150). However, it is not a significant proportion, because although it spends less time learning the other's identity, it cannot find much more than the normative behaviour.

Next, we also give the *agent* more planning resources by setting $N_A^{max} = 6$ and $\Sigma_b = 2$ for the agent, and we run

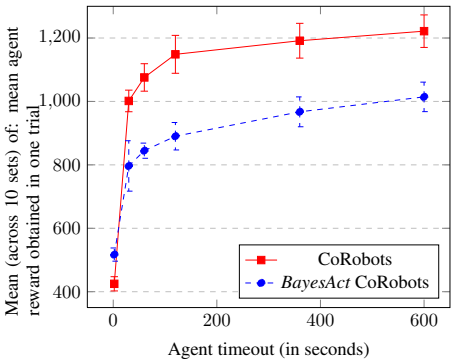

Figure 3: CoRobots: With higher N_A^{max}, Σ_b and $Timeout$, a weaker and less active agent becomes increasingly manipulative by 'faking' his identity, and accumulates higher rewards.

10 sets of 100 simulated trials for each of the following values of agent's $Timeout$: $2, 30, 60, 120, 360, 600$ seconds[11]. Figure 3 (solid red line) shows means and standard error of agent reward per trial (in each set of 100 trials). As the model incorporates noise in movements as well as observations, the robots spend about 20 initial iterations coordinating with each other to choose a leader, during which time they do not receive reward. Thus, a realistic upper bound on the *agent*'s reward is $20 \times 100 = 2000$. Figure 3 shows that at $Timeout = 600$, the reward is about 61% of this realistic maximum, which makes sense given the manipulation rate of about 48%. There is a diminishing rate of return as timeout increases in Figure 3 that is explained by the exponential growth of the MCTS search tree as $Timeout$ increases linearly. The results are relatively insensitive to the choice of parameters such as δ_a and δ_o.

Finally, we play the CoRobots Game with *BayesAct* Robots. This means that the normative behaviour is the deflection minimising action given by Affect Control Theory, instead of Equation (4), and the transient impressions are used to compute the deflection. The game trials are set up exactly as before, and the results are shown in Figure 3 (blue line). As expected, we see the same trends as those obtained previously, but with correspondingly lower values as the transient impressions are used and introduce further complexity to the planning problem (18D state space rather than 9D). Our results demonstrate that the POMCP-C algorithm is able to find and exploit manipulative affective actions within the *BayesAct* POMDP, and gives some insight into manipulative affective actions in *BayesAct*.

5 RELATED WORK

Damasio has convincingly argued, both from a functional and neurological standpoint, for emotions playing a key role in decision making and for human social action [7]. His *Somatic Marker Hypothesis* is contrasted against the

[11] We use a Python implementation that is unoptimized. An optimised version will result in realistic timeouts.

Platonic "high-reason" view of intelligence, in which pure rationality is used to make decisions. Damasio argues that, because of the limited capacity of working memory and attention, the Platonic view will not work. Instead, learned neural markers focus attention on actions that are likely to succeed, and act as a neural bias allowing humans to work with fewer alternatives. These *somatic markers* are "cultural prescriptions" for behaviours that are "rational relative to the social conventions and ethics" ([7], p200).

LeDoux [24] argues the same thing from an evolutionary standpoint. He theorises that the subjective feeling of emotion must take place at both unconscious and conscious levels in the brain, and that consciousness is the ability to relate stimuli to a sense of identity, among other things.

With remarkably similar conclusions coming from a more functional (economic) viewpoint, Kahneman has demonstrated that human emotional reasoning often overshadows, but is important as a guide for, cognitive deliberation [23]. Kahneman presents a two-level model of intelligence, with a fast/normative/reactive/affective mechanism being the "first on the scene", followed by a slow/cognitive/deliberative mechanism that operates if sufficient resources are available. Akerlof and Kranton attempt to formalise *fast thinking* by incorporating a general notion of identity into an economic model (utility function) [2]. Earlier work on *social identity theory* foreshadowed this economic model by noting that simply assigning group membership increases individual cooperation [38].

The idea that unites Kahneman, LeDoux, and Damasio (and others) is the tight connection between emotion and action. These authors, from very different fields, propose emotional reasoning as a "quick and dirty", yet absolutely necessary, guide for cognitive deliberation. ACT gives a functional account of the quick pathway as sentiment encoding prescriptive behaviour, while *BayesAct* shows how this account can be extended with a slow pathway that enables exploration and planning away from the prescription.

Our work fits well into a wide body of work on *affective computing* (AC) [30, 32], with a growing focus on sociocultural agents (e.g. [9]). In AC, emotions are usually framed following the rationalistic view proposed by Simon as "interrupts" to cognitive processing [37]. Emotions are typically inferred based on cognitive appraisals (e.g. a thwarted goal causes anger) that are used to guide action through a set of "coping" mechanisms. Gratch and Marsella [15] are possibly the first to propose a concrete computational mechanism for coping. They propose a five stage process wherein beliefs, desires, plans and intentions are first formulated, and upon which emotional appraisals are computed. Coping strategies then use a set of *ad hoc* rules by modifying elements of the model such as probabilities and utilities, or by modifying plans or intentions. Si *et al.* [35] compute emotional appraisals from utility measures (including beliefs about other agent's utilities, as in an I-POMDP [13]), but they leave to future work *"how emotion affects the agents decision-making and belief update processes"* ([35] section 8). Goal prioritization using emotional appraisals have been investigated [3, 25, 28], as have normative multi-agent systems (NorMAS) [5]. There has been recent work on facial expressions in PD games, showing that they can significantly affect the outcomes [8].

Most approaches to emotional action guidance only give broad action guides in extreme situations, leaving all else to the cognitive faculties. *BayesAct* specifies one simple coping mechanism: minimizing inconsistency in continuous-valued sentiment. This, when combined with mappings describing how sentiments are appraised from events and actions, can be used to prescribe actions that maximally reduce inconsistency. These prescriptions are then used as guides for higher-level cognitive (including rational) processing and deliberation. *BayesAct* therefore provides an important step in the direction of building models that integrate "cognitive" and "affective" reasoning.

BayesAct requires anytime techniques for solving large continuous POMDPs with non-Gaussian beliefs. There has been much recent effort in solving continuous POMDPs with Gaussian beliefs (e.g. [10]), but these are usually in robotics motion planning where such approximations are reasonable. Point-based methods (e.g. [31]) require the value function to be closed under the Bellman operator, which is not possible for *BayesAct*.

Monte-Carlo tree search (MCTS) methods have seen more scalability success [6], and are anytime. POMCP [36] uses MCTS to efficiently solve POMDPs with continuous state spaces. By design, POMCP is unable to handle models with continuous action spaces, such as *BayesAct*. POMCoP uses POMCP to guide a sidekick's actions during a cooperative video game [26]. While this game has many similarities to CoRobots, it does not have continuous actions and restricts agent types to a small and countable set. MCTS methods are more appealing for *BayesAct* than other solvers because: (1) MCTS does not require a computation of the value function over the continuous state space and non-linear dynamics; (2) MCTS provides an anytime "quick and dirty" solution that corresponds naturally to our interpretation of the "fast thinking" heuristic.

6 CONCLUSION

We have studied decision-theoretic planning in a class of POMDP models of affective interactions, *BayesAct*, in which culturally shared sentiments are used to provide normative action guidance. *BayesAct* is an exciting new development in artificial intelligence that combines affective computing, sociological theory, and probabilistic modeling. We use a Monte-Carlo Tree Search (MCTS) method to show how a simple and parsimonious model of human affect in decision making can yield solutions to two classic social dilemmas. We investigate how asymmetry between agent's resources can lead to manipulative or exploitative, yet socially aligned, strategies.

References

[1] K. J. Åström. Optimal control of Markov decision processes with incomplete state estimation. *J. Math. Anal. App.*, 10:174–205, 1965.

[2] George A. Akerlof and Rachel E. Kranton. Economics and identity. *Quar. J. Econ.*, CXV(3), August 2000.

[3] Dimitrios Antos and Avi Pfeffer. Using emotions to enhance decision-making. In *Proc. International Joint Conferences on Artificial Intelligence*, Barcelona, Spain, 2011.

[4] Nabiha Asghar and Jesse Hoey. Monte-Carlo planning for socially aligned agents using Bayesian affect control theory. TR CS-2014-21, Univ. of Waterloo Sch. of CS, 2014.

[5] Tina Balke, *et al.* Norms in MAS: Definitions and Related Concepts. In *Normative Multi-Agent Systems*, volume 4 of *Dagstuhl Follow-Ups*, Schloss Dagstuhl, 2013.

[6] C.B. Browne, *et al.* A survey of Monte Carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43, March 2012.

[7] Antonio R. Damasio. *Descartes' error: Emotion, reason, and the human brain*. Putnam's sons, 1994.

[8] Celso M. de Melo, *et al.* Bayesian model of the social effects of emotion in decision-making in multiagent systems. In Proc. *AAMAS*, Valencia, Spain, 2012.

[9] Nick Degens, *et al.* Creating a world for socio-cultural agents. In LNAI no. 8750. Springer, 2014.

[10] Marc Peter Deisenroth and Jan Peters. Solving nonlinear continuous state-action-observation POMDPs for mechanical systems with gaussian noise. In *Proceedings of the European Workshop on Reinforcement Learning (EWRL)*, 2012.

[11] Arnaud Doucet, Nando de Freitas, and Neil Gordon, editors. *Sequential Monte Carlo in Practice*. Springer-Verlag, 2001.

[12] John G. Fennell and Roland J. Baddeley. Reward is assessed in three dimensions that correspond to the semantic differential. *PLoS One*, 8(2: e55588), 2013.

[13] Piotr Gmytrasiewicz and Prashant Doshi. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79, 2005.

[14] Erving Goffman. *Behavior in Public Places*. The Free Press, New York, 1963.

[15] Jonathan Gratch and Stacy Marsella. A domain-independent framework for modeling emotion. *Cognitive Systems Research*, 5(4):269 – 306, 2004.

[16] David R. Heise. *Expressive Order: Confirming Sentiments in Social Actions*. Springer, 2007.

[17] David R. Heise. *Surveying Cultures: Discovering Shared Conceptions and Sentiments*. Wiley, 2010.

[18] David R. Heise. Modeling interactions in small groups. *Social Psychology Quarterly*, 76:52–72, 2013.

[19] Jesse Hoey, Craig Boutilier, Pascal Poupart, Patrick Olivier, Andrew Monk, and Alex Mihailidis. People, sensors, decisions: Customizable and adaptive technologies for assistance in healthcare. *ACM Trans. Interact. Intell. Syst.*, 2(4):20:1–20:36, January 2012.

[20] Jesse Hoey and Tobias Schröder. Bayesian affect control theory of self. In *Proc. AAAI*, 2015.

[21] Jesse Hoey, Tobias Schröder, and Areej Alhothali. Affect control processes: Intelligent affective interaction using a partially observable Markov decision process. http://arxiv.org/abs/1306.5279, 2013.

[22] Jesse Hoey, Tobias Schröder, and Areej Alhothali. Bayesian affect control theory. In *Proc. ACII*, 2013.

[23] Daniel Kahneman. *Thinking, Fast and Slow*. Doubleday, 2011.

[24] Joseph LeDoux. *The emotional brain: the mysterious underpinnings of emotional life*. Simon and Schuster, New York, 1996.

[25] Christine Laetitia Lisetti and Piotr Gmytrasiewicz. Can a rational agent afford to be affectless? a formal approach. *Applied Artificial Intelligence*, 16(7-8):577–609, 2002.

[26] Owen Macindoe, Leslie Pack Kaelbling, , and Tomás Lozano-Pérez. Pomcop: Belief space planning for sidekicks in cooperative games. In *AIIDE 2012*, 2012.

[27] Neil. J. MacKinnnon and Dawn T. Robinson. 25 years of research in affect control theory. *Advances in Group Processing*, 31, 2014.

[28] Robert P. Marinier III and John E. Laird. Emotion-driven reinforcement learning. In *Proc. Meeting of the Cognitive Science Society*, Washington, D.C., 2008.

[29] Charles E. Osgood, William H. May, and Murray S. Miron. *Cross-Cultural Universals of Affective Meaning*. University of Illinois Press, 1975.

[30] Rosalind W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, 1997.

[31] Josep M. Porta, Nikos Vlassis, Matthijs T.J. Spaan, and Pascal Poupart. Point-based value iteration for continuous POMDPs. *JMLR*, 7:2329–2367, 2006.

[32] Klaus R. Scherer, Tanja Banziger, and Etienne Roesch. *A Blueprint for Affective Computing*. Oxford University Press, 2010.

[33] Wolfgang Scholl. The socio-emotional basis of human interaction and communication: How we construct our social world. *Social Science Information*, 52:3 – 33, 2013.

[34] Tobias Schröder and Wolfgang Scholl. Affective dynamics of leadership: An experimental test of affect control theory. *Social Psychology Quarterly*, 72:180–197, 2009.

[35] Mei Si, Stacy C. Marsella, and David V. Pynadath. Modeling appraisal in theory of mind reasoning. *Autonomous Agents and Multi-Agent Systems*, 20(1):14–31, 2010.

[36] David Silver and Joel Veness. Monte-Carlo planning in large POMDPs. In *Proc. NIPS*, December 2010.

[37] Herbert A. Simon. Motivational and emotional controls of cognition. *Psychological Review*, 74:29–39, 1967.

[38] Henri Tajfel and John C. Turner. An integrative theory of intergroup conflict. In Stephen Worchel and William Austin, editors, *The social psychology of intergroup relations*. Brooks/Cole, Monterey, CA, 1979.