Sequential Decision Making for Intelligent Agents
Papers from the AAAI 2015 Fall Symposium

Commitment Semantics for Sequential Decision Making Under Reward Uncertainty

Edmund H. Durfee and **Satinder Singh**
Computer Science and Engineering
University of Michigan, Ann Arbor, MI 48109
durfee@umich.edu, baveja@umich.edu

Abstract

A commitment represents an agent's intention to attempt to bring about some state of the world that is desired by some agent (possibly itself) in the future. Thus, by making a commitment, an agent is agreeing to make sequential decisions that it believes can cause the desired state to arise. In general, though, an agent's actions will have uncertain outcomes, and thus reaching the desired state cannot be guaranteed. For such sequential decision settings with uncertainty, therefore, commitments can only be probabilistic. We argue that standard notions of commitment are insufficient for probabilistic commitments, and propose a new semantics that judges commitment fulfillment not in terms of whether the agent achieved the desired state, but rather in terms of whether the agent made sequential decisions that in expectation would have achieved the desired state with (at least) the promised probability. We have devised various algorithms that operationalize our semantics, to capture problem contexts with probabilistic commitments arising because action outcomes are uncertain, as well as arising because an agent might realize over time that it does not want to fulfill the commitment.

Our focus in this paper is on what it means for an agent to pursue a commitment it has made to another agent when: the agents operate in a sequential decision setting; the agent pursuing the commitment has uncertainty about the model of the environment (and not just about the current state of the environment); and the agent, while sequentially executing decisions, can make model-informative observations—observations that change its beliefs about the correct model of the environment. In particular, we largely focus on *reward uncertainty*, where as it experiences the world the agent better learns what rewards to associate with reaching different states of the world.

To concentrate our exposition on the question of commitment semantics in the face of such model uncertainty, we restrict our attention in this paper to the two agent case, and without loss of generality will refer to the agent to whom a commitment is made as the *User* and the agent making the commitment to the user simply as the *agent*. Intuitively, the agent is acting in its environment in part to try to enable the user to satisfy her objectives. Note that in a single-agent commitment setting, the "user" and "agent" are the same entity, where the entity is acting in a particular role at a given time and can be committed to actions that support itself when acting in a different role and/or at a later time.

The contributions we offer in this paper, corresponding to the sequence of sections below, are as follows. We begin with a brief summary of computational models of commitments to highlight the limitations of past work with respect to providing clear semantics for how commitments should impact the sequential decisions of an agent that is attempting to fulfill the commitment. This in turn leads to our contributing an initial general characterization of the commitment semantics problem in a decision-theoretic formulation, along with specializations of interest. We then stake out our position on what the semantics *should be* for probabilistic commitments in a sequential-decision-theoretic setting, and how those semantics depart from prior stances. The remainder of the paper then explores some possible algorithms for operationalizing the semantics for: reward uncertainty and stochastic actions; reward uncertainty and competing objectives; and other kinds of model uncertainty.

Computational Models of Commitment

Munindar Singh (unrelated to co-author Satinder Singh) provides a comprehensive overview of computational research into characterizing commitments using formal (modal and temporal) logic (Singh 2012), drawing on a broad literature (e.g., (Cohen and Levesque 1990; Castelfranchi 1995; Singh 1999; Mallya and Huhns 2003; Chesani et al. 2013; Al-Saqqar et al. 2014)). In brief, these formulations support important objectives such as provable pursuit of mutually agreed-upon goals, and verification of communication protocols associated with managing commitments. When commitments are uncertain to be attained, they can have associated conventions and protocols for managing such uncertainty (e.g., (Jennings 1993; Xing and Singh 2001; Winikoff 2006)). For example, by convention an agent unable to keep a commitment must inform dependent agents.

Dropping commitments too readily, however, obviates their predictive value for cooperation. The logical formulations above explicitly enumerate the conditions under which an agent is permitted to drop a local component of a mutual goal, where these conditions usually amount to either (1)

when the agent believes its local component is unachievable; (2) when the agent believes that the mutual goal is not worth pursuing any longer; or (3) when the agent believes some other agents have dropped their components of the mutual goal. However, while logically reasonable, these conditions do not impose a commitment semantics on an agent's local decisions. For example, to avoid the first condition, should an agent never take an action that would risk rendering is local component unachievable? What if every action it can take has some chance of rendering the local component unachievable? For the second condition, should it really be allowed to unilaterally abandon the mutual goal and renege on other agents just because it has recognized it can achieve a slightly more desirable goal?

To tighten predictability, commitments can be paired with conditions under which they are sure to hold (Raffia 1982; Singh 2012; Vokrínek, Komenda, and Pechoucek 2009; Agotnes, Goranko, and Jamroga 2007). For example, an agent could commit to providing a good or service conditioned on first receiving payment. Of course, this representation also admits to weakening commitments to the point where they are worthless, such as committing to achieving a local component of a mutual goal under the condition that no better local goal arises in the meantime! Sandholm and Lesser (Sandholm and Lesser 2001) noted difficulties in enumerating such conditions, and verifying they hold in decentralized settings. Their leveled-commitment contracting framework associates a decommitment penalty with each commitment to accommodate uncertainty but discourage frivolous decommitment. The recipient of a commitment, however, will generally be unable to know the likelihood that the commitment will be fulfilled, because it will lack knowledge of the internals of the agent making the commitment, including how likely it is that uncertain action outcomes or evolving local goals will make paying the decommitment penalty the only/better choice.

An alternative means to quantify uncertainty is to explicitly make probabilistic commitments, where an agent provides a probability distribution over possible outcomes of the commitment, including how well it will be fulfilled (if at all) and when (Xuan and Lesser 1999; Bannazadeh and Leon-Garcia 2010; Witwicki and Durfee 2009). Xuan and Lesser (1999) explain how probabilistic commitments can improve joint planning by allowing agents to suitably hedge their plans to anticipate possible contingencies, including anticipating even unlikely outcomes and planning for consequent changes to probabilities of reaching commitment outcomes. A more myopic (hence more tractable) variation on this approach was developed for the DARPA Coordinators program (Maheswaran et al. 2008), where only as circumstances unfolded would the agents update probabilistic predictions about future outcomes, and then exchange updates and reactively compute new plans. These prior approaches however treat commitment probabilities fundamentally as predictions about how whatever plan an agent has chosen to follow will affect other agents. In contrast, this paper emphasizes probabilistic commitments that provide both predictive information about what might happen and prescriptive semantics for making those predictions come true.

Problem Formulation

Our strategy for capturing intuitive, everyday notions of commitment semantics that account for and respond to model uncertainty is to map these notions into a principled, decision-theoretic framework for agent use. Here, we present a reward-uncertainty-centered formulation that we use most in this paper, though at the paper's end we generalize this to other forms of model uncertainty. In our initial formulation, we restrict our attention to the class of problems with the following properties. 1) A single intelligent agent interacts with a single human user (operator). 2) The agent's actions influence what is possible for the user to achieve but not vice-versa (though, because the user also derives reward from the agent's actions, the user's preferences might influence what the agent *should* do). 3) The agent has an accurate controlled Markov process model of its environment dynamics defined by a multidimensional state space, an action space, and a transition probability function. The state space $\Phi = \Phi_1 \times \Phi_1 \times \cdots \times \Phi_n$ is the cross product of n discrete-valued state variables. The transition probability $T(\phi'|\phi, a)$ is the probability of the next state being ϕ' given the agent took action a in state ϕ. 4) The agent has uncertainty over its reward function expressed via a prior distribution μ_0^b over possible reward functions $R_1^b, R_2^b, \ldots, R_n^b$, where each R_i^b maps $\Phi \rightarrow \mathbb{R}$. Each reward function R_i^b captures both the designed-rewards for the agent (e.g., a large negative reward for exceeding power or memory constraints), and the uncertain rewards that can arise over time in the environment. From the perspective of the single human-user in this problem, these multiple sources of reward are "built-in" and the uncertainty over them is summarized into the distribution over $\{R_i^b\}$. The agent obtains samples of the true built-in reward-function as it acts in the world and thus can update its distribution over $\{R_i^b\}$ during execution.

Finally, 5) the user has her own goals and acts in the world, and the agent's actions may **enable** the user to obtain higher reward than she would without the agent's help. This is where the notion of commitment from the agent comes into play. Consider an agent that could make either of two commitments to an operator: commitment ξ, where it commits to producing an analysis within 2 minutes with probability at least 0.95, and commitment ξ' where it commits to producing the analysis in 1 minute but with probability only 0.5 (e.g., its faster analysis tool works in fewer cases). Commitment ξ enables the operator's optimal policy to prepare for the analysis output with associated enablement-utility $U(\xi)$, while commitment ξ' induces an optimal policy where the operator begins doing the analysis herself (as a backup in case the agent fails) with lower utility $U(\xi')$. Solving the agent's planning problem requires taking into account these enablement-utility (U) values to the user of candidate enablement-commitments.

Some special cases of this formulation highlight aspects of our approach:

Bayes-MDP. In this special case, the agent is not enabling user actions (no U's and hence no need for commitments), but the agent is uncertain about which of the built-in rewards $\{R_i^b\}$ applies. The agent thus faces a stan-

dard Bayesian-MDP problem (a particular kind of partially-observable MDP, or POMDP, where partial observability is only with respect to the true reward function in $\{R_i^b\}$). One can define an extended belief-state MDP in which the belief-state of the agent at time t is the joint pair (ϕ_t, μ_t^b) where μ_t^b is the posterior belief of the agent over $\{R_i^b\}$ after the first $t-1$ observations about reward as it acts in the world. The Bayes-optimal policy is a mapping from belief-states to actions that maximizes the expected cumulative reward for the agent. Exact algorithms (applicable only to small problems) and approximate algorithms (with increased applicability) exist to solve the belief-state MDP for (near-Bayes-optimal) policies and we exploit them as one component in our research (Poupart et al. 2006).

Commitment-Only. In this case, there are enablement-actions but the built-in reward function is known to be R^b. Because of stochastic transitions, the agent could find itself in unlikely states from which it cannot enable the user, and thus commitments are in general only probabilistic. Because the agent can only control its actions, and not their outcomes, we assert that, in stochastic worlds, *the decision-theoretic semantics of what it means for an agent to faithfully pursue a probabilistic commitment is that it adheres to a policy that in expectation meets the commitment.* Given that its rewards are fixed (in this special case) the agent will at the outset commit to a policy that maximizes some function of its expected reward and the user's enablement utility, and follow that policy unswervingly. In a cooperative setting (including when a single agent is making a commitment to another facet of itself), the function could simply sum these. In other settings, the agent's reward could predominate (the user is helped only as a side-effect of the agent's preferred policy) or the user's utility could be preeminent.

Commitment in the face of Uncertain Rewards. This special case is the main focus of this paper, where there is uncertainty over the agent's rewards ($\{R_i^b\}$), and there is the possibility of enablement (U). The departure from the previous commitment-only case is that now the agent learns about its built-in reward function as it acts in the world. As in the previous case, in general commitments are only probabilistic because transitions are stochastic, so the agent has limitations in its ability to help the user attain the enablement utility U despite its best efforts. Compounding this problem, the evolving model of the reward function might also tempt the agent toward redirecting its efforts away from the enablement. What can we expect of an agent in terms of making sequential decisions that live up to a commitment when it is faced with such limitations and temptations?

The sections that follow stake out our position regarding our proposed answer to this question. In the next section, we posit and defend a commitment semantics for such settings. Then we turn to questions of operationalizing this semantics. We first consider the case where failure to achieve the enablement happens due to "bad luck" in action outcomes (corresponding to the first condition for an agent to drop its local component of a mutual goal—belief the goal is unachievable), and subsequently we examine the case where an agent might purposely drop the commitment as being "not worth it" (corresponding to the second condition for dropping a local component of a mutual goal). (Note that the final condition for dropping a local component—because at least one other agent has dropped its local component—implies that some other agent dropped its local component for one of the first two reasons.)

Commitment Semantics

Our position is that the semantics for commitments in stochastic, sequential-decision settings, as was mentioned in the previous section, should be as follows: *The semantics of what it means for an agent to faithfully pursue a probabilistic commitment is that it adheres to a policy that in expectation meets the commitment.* This sounds straightforward enough, though as the sections that follow will show it is not always trivial to operationalize. Before looking at algorithms for implementing the semantics, however, we first briefly consider how this semantics departs from prior semantics for computational commitments.

Probably the most thorough and precise computational semantics for commitments is that of Munindar Singh and his colleagues. In that vein of work, commitments are expressed in terms of expressions over state variables, describing what state(s) the agent(s) making the commitment promises to bring about, possibly conditioned on other agents achieving other aspects of the state. However, as we have discussed, in stochastic environments agents cannot commit to assuredly achieving particular states because outcomes of actions are not fully under their control. Agents however do have control over the actions they take, and hence our semantics focus not on states of the world *but rather on the actions agents have control over.* Agents commit to acting in ways that, with sufficiently high probability, will lead to outcomes that other agents care about.

In this regard, then, at some level our commitment semantics is more similar to work on joint policies in cooperative planning frameworks like Decentralized (Partially-Observable) Markov Decision Processes. In Dec-(PO)MDP solutions, agents' joint policies dictate a particular policy for each agent to follow, where the policy of each agent is (approximately) optimized with respect to the policies to be followed by the others. Thus, optimal joint behavior is achieved when agents precisely execute their assigned policies. Our commitment semantics similarly restrict agents' policy choices, but differ from Dec-POMDPs in that our semantics are agnostic about cooperation (we treat the reason why agents adopt commitments as orthogonal to what the commitments that have been adopted mean) and only require that an agent pursue a policy that in expectation will achieve the commitment: If there are multiple such policies, then the agent is free to select from among them. As we will see next, this is exactly the kind of flexibility that we seek to exploit when an agent is sequentially acting under reward uncertainty.

Managing Commitments Given Uncertain Rewards and Stochastic Actions

We here provide an algorithmic treatment of, and some empirical illustrations for, operationalizing our commitment se-

mantics in settings where commitments are probabilistic because an agent might have "bad luck." Our main result[1] here is an algorithm for finding a good commitment and then behaving consistently with respect to the commitment as the agent learns about its true reward function.

First, consider a *stylized* illustrative example in which a robotic agent and a human user occupy two different regions (see Figure 1(upper)). In two different locations (marked with "switch" icons) in its region, the robot can activate an enabling action that remotely opens a gate, thus *enabling* the user's direct path from (0,1) to (0,2). If the gate is open the user can take a higher-reward path to her goal destination (0,2). The rewards for the robot and user are shown in each location in their respective regions. The robot faces the decision of whether or not to make its way to the initially safe location (1,2) to open the gate and risk encountering a gathering crowd. The size of the crowd is uncertain to the robot, creating a distribution over the robot's possible built-in reward functions $\{R_i^b\}$ where, at each time step, there is a 10% chance that the rewards in the shaded locations will all decrease (by 3 in switch location (1,2) and .01 in the other shaded places). Moreover, the robot must decide how long to linger in location (1,2), flicking the switch again if the gate does not open on earlier tries (a switch works with probability 0.7 each try). Analogously, depending on the robot's commitment, the user must decide whether to wait for the robot to open the gate or to follow a more costly detour to reach her positive reward. Note that, because a switch only works probabilistically, the robot cannot promise the user that the gate will ever be open with certainty: any commitment is inherently probabilistic.

In this stylized example, the only state variable mutually modeled by both human and robot is the status of the gate, and thus any commitment can be restricted to be a promise "to open the gate by time step X with probability at least p". For each commitment ξ there is a set of *commitment-constrained* policies, Π_ξ, that achieve that commitment. Recall the robot's policies are mappings from its belief-states to actions in the belief-state MDP as defined in the above section on the Bayes-MDP problem. If the set Π_ξ is empty, the associated commitment is not feasible and we needn't consider infeasible commitments. For each feasible commitment ξ, the user can compute an optimal policy (whether to wait for the robot to open the gate or follow the detour) and the associated optimal value for the user's start state then defines a scalar measure, $U^h(\xi)$ that captures the user's enablement utility for commitment ξ; this is all the robot needs to know in considering the effect of its commitment on the user. So what commitment should the robot make? We will define this in two steps. The optimal policy for the robot given a feasible commitment ξ is

$$\pi_\xi^* = \arg\max_{\pi \in \Pi_\xi} V^{\pi,b} \tag{1}$$

[1]The algorithms and comparisons in this section have appeared before in a MSDM paper (Chen et al. 2012) and an AAMAS extended abstract (Witwicki et al. 2012), but have not previously been related to the overarching commitment semantics that are the focus of this paper.

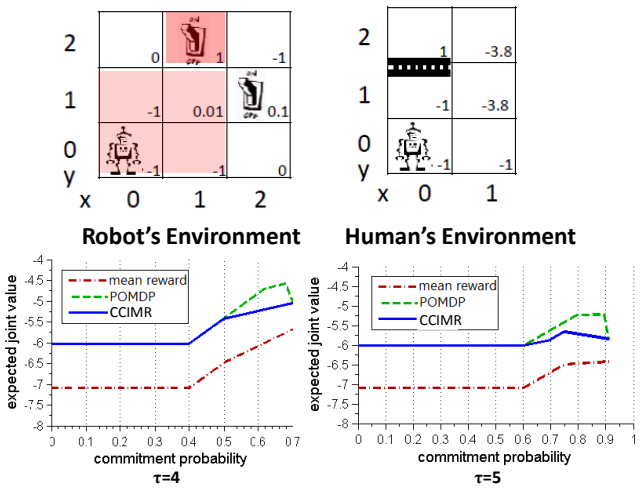

Figure 1: Stylized Gate-Control Problem Depiction (upper) and Experimental Results (lower).

where $V^{\pi,b}$ is the expected value for the start state obtained from the built-in reward function when the robot behaves according to policy π. By exploiting influence-abstraction techniques (Witwicki and Durfee 2010) for this problem's special case of Decentralized POMDP (Bernstein et al. 2002; Becker, Zilberstein, and Lesser 2004; Goldman and Zilberstein 2004), we actually never have to compute or represent the constrained-policy sets explicitly but instead we can incorporate the commitment-constraints directly in a linear programming approach to find π_ξ^*. Then, the optimal commitment from some set of feasible commitments Ξ is a function of $U^h(\xi)$ and $V^{\pi_\xi^*,b}$, such as the sum of these when agents are cooperative:

$$\xi^* = \arg\max_{\xi \in \Xi} \left\{ U^h(\xi) + V^{\pi_\xi^*,b} \right\} \tag{2}$$

Of course, there is a lot of structure to be exploited. For example, if ξ and ξ' are commitments that are the same except that ξ promises higher probability of enablement, then $\Pi_\xi \subseteq \Pi_{\xi'}$ and $U^h(\xi) \geq U^h(\xi')$. The same holds if they differ only in ξ promising earlier enablement. We can exploit such structure to significantly reduce the computational burden of finding optimal commitments and associated policies.

Next we turn to some methods the robot might use to handle belief-revision as it acts in the world.

Expanded-Belief-State (EBS) Algorithm. Computing value functions in the belief-state-MDP for Equations 1 and 2 accounts for every possible stochastic outcome of the robot's actions as well as every way that reward observations could change its belief state about rewards. This *Expanded Belief-State* (EBS) algorithm computes optimal commitments and policies because it builds into a policy (and hence the commitment to the user) an optimal response to all possible reward updates, so the robot will never benefit by deviating from its commitment to executing this comprehensive policy. *Hence, our commitment semantics map directly to the case of reward uncertainty, in this case requiring the robot to execute a commitment-constrained pol-*

icy over the expanded belief state. However, the belief-state-MDP can be prohibitively large for all but the smallest of problems (the policy tree branches on all possible belief updates as well as action outcomes), and so we explore alternate solutions.

Mean-Reward (MR) Algorithm. A suboptimal algorithm, but far more tractable than EBS, that nevertheless maintains the semantics of commitments can be derived from the following observation. Uncertainty only about rewards does not impact a robot's ability to meet its commitments: the underlying dynamics of the world (how action choices lead stochastically to outcomes) are unaffected by changing beliefs about rewards. So, if the robot has found a commitment-constrained local policy, following this policy regardless of beliefs about rewards, like in the specialized Commitment-Only case, satisfies its commitments. Obviously, a robot that is locked into its initial commitment-constrained policy should pick a policy that is optimal in expectation for its initial reward belief-state, which equates (Ramachandran and Amir 2007) to an optimal policy for the distribution's Mean Reward (MR). A robot using the resulting MR algorithm gains no benefit from modeling changing reward belief states, dramatically reducing computation relative to EBS. However, because it is insensitive to changing beliefs about the distribution over reward functions, the MR algorithm is, in general, only an approximation to the optimal EBS algorithm.

Commitment-Constrained Iterative Mean Reward (CCIMR) Algorithm. Our CCIMR algorithm is a compromise between the extremes of EBS and MR. We use the MR ideas for computational advantage but don't lock the robot into its initial policy. To meet our semantics for commitment, however, the robot's alternative choices for policies must be carefully circumscribed.

We begin by considering how MR could be used to respond to changing belief states about rewards *in the absence of commitments*. This is the iterative mean-reward approach for Bayesian-MDPs (Poupart et al. 2006), which reapplies MR after each update to beliefs about the true reward function. Since the posterior distribution over reward functions can change, so can the mean reward, and hence adopting the policy optimal for the updated mean reward may outperform the policy adopted at the previous iteration.

The shortcomings of letting the robot simply use iterative mean-reward are obvious: The robot changes its policy given its evolving belief about rewards, and the commitment goes out the window. A simple but flawed extension of iterative mean-reward to the commitment context is for the robot to iteratively compute a new commitment-constrained optimal policy to follow from the point where its reward belief-state changes. Unfortunately, this is untenable, because the robot's stochastic action outcomes could have put it into a state where *no* policy from this state forward can achieve the commitments with sufficient probability. Indeed, note that even adhering only to its initial policy, the robot could reach states from which that policy cannot *from this state forward* achieve the desired probabilistic influence!

This observation helps us further refine the semantics of a probabilistic commitment: **The robot fulfills a commitment if it follows a policy that, *at the time the commitment was made*, achieves the promised commitment, in expectation.** If the policy that the robot actually follows during execution is among the commitment-constrained policies that it *could* have selected at the outset (Π_ξ), then by our definition the robot acted faithfully with respect to its commitments. Clearly, the EBS and the MR algorithms satisfy this condition, because they follow the same commitment-constrained policies throughout. We now look at how an iterative algorithm can satisfy this condition as well. First, the robot uses the MR algorithm to find its initial mean-reward optimal commitment-constrained policy (Equation 1). After taking the first action, say a_1, its stochastic action outcome and reward observations put it in a new belief state. Unlike MR which would continue executing the initially-adopted policy, CCIMR recomputes an optimal policy as follows: it uses MR to find the policy *from the initial state* that is optimal given the updated mean reward for the new belief state, but where the policy (a) must prescribe action a_1 for the initial state (because it cannot undo the past) and (b) must satisfy the probabilistic commitments from the initial state. This is repeated at successive times, where actions chosen so far shrink the number of allowable (consistent) choices left in Π_ξ, which is assured to be non-empty because the policy adopted at the previous timestep will always be one (possibly the only remaining) option at the next timestep. By construction, this process satisfies our commitment semantics.

We have proven that the expected value achieved by CCIMR is lower-bounded by that of MR and upper-bounded by that of EBS. This is empirically shown, for the illustrative problem of Figure 1(upper), in Figure 1(lower) which show the expected joint reward for each method over all feasible commitment probabilities for opening the gate by time $\tau = 4$ (one try with a switch) and $\tau = 5$ (2 tries), respectively. Note EBS and CCIMR outperform MR since they can change policies in response to updated beliefs about rewards to redirect the robot to the switch at (2,1). As expected, EBS also outperforms CCIMR. Not shown, however, is that, depending on the planning time horizon, CCIMR can run orders of magnitude faster than EBS, and thus provides a useful cost/quality compromise. Finally, for this problem the CCIMR optimal commitment is to open the gate at $\tau = 4$ with probability about 0.7 (slightly less for EBS). But for $\tau = 5$ notice that CCIMR peaks at a significantly lower commitment probability (.75) than the others (about .9)—by making a weaker commitment, this "hedging" intentionally expands the space of commitment-satisfying policies Π_ξ to reserve more alternatives for adapting to evolving beliefs about rewards!

Managing Commitments Given Uncertain Rewards and Competing Objectives

In the previous section, a commitment was probabilistic because of uncertainty about actions' effects, such as whether the switch would work; the agent could drop the enablement goal of opening the gate because, due to bad luck rather than its own choices, it reaches a state where the goal is not

achievable. Now we turn to the other condition that others have identified for when an agent should unilaterally drop a goal to work with others: when it believes the goal is not worth pursuing.

To illustrate this kind of situation, consider a simple variation on the gate-control problem used in the previous section. In this variation, there is only one switch, the one in risky location (1,2). In addition, to keep things simple, let's assume that the switch is 100% reliable (no transition stochasticity), that the rewards in the shaded regions only probabilistically decrease after the first (rather than every) time step, but now with a 50% chance, and that the robot's time horizon is 4. In this setting, with no action stochasticity, the agent could choose to commit to opening the gate with certainty at time 4, with expected cumulative value of -3 (-1.005 is the mean reward for state (1,0), plus .005 for state (1,1), plus twice -1 for state (1,2)). It could also choose to make no commitment to opening the gate, in which case it goes to safe state (2,1) with an expected cumulative value of -0.8. Let's say that the expected utility to the user of opening the gate is 3 more than if the gate is unopenned. Then the optimal commitment from equation 2 would be to open the gate with certainty.

This is the choice that CCIMR would make under that assumption. In the 50% of cases where the agent discovers that the rewards have been lowered, CCIMR would search for a better alternative policy that would still open the gate with certainty, and fail to find one. To adhere to our commitment semantics, the robot would bite the bullet and follow through with the commitment despite its local cumulative reward of -5.01 in this case, making this choice suboptimal by equation 2. (Note that had we left the other switch in, CCIMR would still not alter the policy to go to it unless that switch was also 100% reliable.)

EBS, on the other hand, can exploit its look-ahead model of possible outcomes of reward-informative observations to do better. Specifically, the EBS branching policy tree includes branching at timestep 1 for each of the possible reward observations: if when reaching state (1,0) the built-in reward received is -1 (expected to happen 50% of the time), then proceeding to the switch is optimal with local reward of 1.01, but if the observed reward is -1.01 then moving through (1,1) to the safe haven of (2,1) results in a local reward of -0.81, compared to -5.01 had it proceeded to the switch. Hence, unlike a CCIMR agent that would only consider commitments with probabilities 1 and 0, an EBS-based agent can also consider a commitment to open the gate with probability 0.5. Notice that, unlike the previous section, this probability is not reflecting the chances that its actions will turn out unlucky (e.g., the switch will fail), but rather the chances that it will later decide that the commitment is not worth keeping. Using equation 2, EBS with these additional commitment choices is assured of finding a commitment no worse, and (depending on relative rewards of competing objectives) often substantially better, than CCIMR.

Unfortunately, the computational costs of EBS are still prohibitive; if new reward observations could be made every timestep like in the original version of the problem, the branching factor can quickly overwhelm the agent's computational resources. To combat these costs, we are developing a CCIMR algorithm extension that limits branching due to reward observations. At the time that this paper is being written, we lack sufficient empirical results to justify claims about the algorithm, so we restrict the remainder of this section to describing and illustrating the algorithm under development.

As its name implies, our new CCIMR with reward-observation branching (CCIMR-ROB) algorithm introduces reward-observation branching to CCIMR, but unlike EBS does so only *retrospectively*. That is, unlike EBS which looks ahead prospectively to all possible action, outcome, and reward-observation trajectories, CCIMR-ROB only includes reward-observation branching points for observation events (not just the specific observations) that it has experienced *so far*. CCIMR-ROB then applies the CCIMR techniques of computing commitment-constrained policies, but using the mean-reward induced on each of the branches (meaning that it extends the belief-state representation) and allowing stochastic policies.

To illustrate the algorithm, consider the problem variation introduced at the beginning of this section. For the moment, assume that the agent using CCIMR-ROB has agreed to the (EBS-optimal) commitment probability of 0.5. Before taking any action, it uses CCIMR to find the optimal commitment-constrained mean-reward policy. As mentioned before, CCIMR can find policies that achieve the commitment with probability 1 or probability 0. If the commitment with probability 1 is the better choice given mean rewards (as in the example above where the enablement gain is 3), then CCIMR-ROB will follow it for the first timestep. (Recall, the semantics stipulate that the agent needs to satisfy the commitment with *at least* the specified probability, so overshooting is fine.) If, on the other hand, the probability 0 commitment is optimal (e.g., had the enablement gain been only 1), there is competition between the agent's local rewards and the commitment. Unless the commitment can be renegotiated (which is beyond the scope of this paper), the agent must achieve it, so in this case CCIMR-ROB can adopt a stochastic policy (which is not something that CCIMR could do), which in this case means that in state (1,1) the agent will move North with probability 0.5, and East with probability 0.5, which meets the commitment and (under this assumption) is better than the pure policy of always moving North.

Now the agent takes the first step in the policy, which in the example brings it to state (1,0) (all the policies take this same first action). Now the agent makes a reward observation, and knows whether state (1,2) has reward of 1 or -2. At this iteration, CCIMR-ROB does the following. Given that it has experienced a reward observation, it recreates the state trajectory with branching associated with making a reward observation in that state, but (using the same model information EBS does to create reward-observation branches) includes branches for each of the observations it *could* have made, rather than just the one it *did* make. It annotates the root state of each of these branches with what the reward-belief state would be had the corresponding observation been made. And then it computes the commitment-

constrained optimal policy for the new representation, where the mean reward function can differ in each of the branches. (Recall that the policy up to this point is unalterable, so the branching factor up to this point does not change.) In the example problem, CCIMR-ROB introduces branches to indicate the 50-50 chances of the two different reward functions. It then computes a better commitment-constrained policy: in the branch with the higher mean reward, in state (1,1) it moves North with certainty, while in the other branch it moves East with certainty. The commitment is still fulfilled based on our semantics: given the agent's model, it is following a policy that, had it selected it at the outset, would indeed have had a 50% chance of opening the gate.

CCIMR-ROB worked perfectly in this illustrative example, but not surprisingly it suffers from being myopic. Specifically, it will run into trouble in cases where retrospectively adding reward-observation branches comes too late—after decisions have already been taken that prevent it from exploiting later improvements to its model. In the example problem, the agent benefited from the fact that all of the policies specified the same actions (in this case just one action) up to the point where the crucial reward observation arrived. In contrast, consider a variation where a wall blocks movement between (1,0) and (1,1), state (1,0) has a reward of -.99, and the enablement utility is negligible. The optimal mean-reward policy would lead the agent to state (2,1) through (1,0). After the first action, it observes the true reward function, and would realize that, if it could start again, it would have been better first going to (0,1), but at this point it is too late.

CCIMR-ROB thus represents another point in the spectrum between the full but expensive optimality of EBS and the unresponsively-suboptimal but cheap MR algorithm. In our limited experiments with it so far, our expectations have been confirmed that its performance, and costs, are between those of CCIMR and EBS. Among our ongoing research threads is examining whether heuristic techniques can decrease CCIMR-ROB's myopic limitations by, for example, picking a policy not based only on optimizing expected value given the commitment (equation 1), but also on the degree to which it shares a prefix (first step(s)) with other policies. That is, to favor policies that require less commitment in exactly how the agent will pursue its commitment!

Handling Other Kinds of Uncertainty

The work we've done so far has emphasized the need to account for and respond to an agent's uncertain rewards. However, uncertainty can arise in other decision model components too. For example, an agent can apply machine learning techniques to resolve uncertainty about its transition model: by maintaining statistics about the effects of actions in various states, it improves its ability to predict action outcomes and thus to plan. Making commitments in the face of transition uncertainty unfortunately appears to be qualitatively different from the reward uncertainty case. A key observation is that, when uncertainty is only over rewards, then the agent can *always* faithfully pursue its commitment by, in the worst case, turning a blind eye to what it learns about rewards and simply following its initial commitment-fulfilling policy throughout. That is, what it learns about rewards has no effect on what states of the world it can probabilistically reach, but just in how happy it is to reach them. In contrast, an agent with transition uncertainty can learn, during execution, that states it thought likely to be reached when it made its commitment are in fact unlikely, and *vice versa*. Hence, in contrast to reward uncertainty where a committed agent was obligated to pursue one of the initial commitment-constrained policies (limiting its later choices), with transition uncertainty it could be argued that a faithful agent should be *required* to shift to a policy outside this initial set under some changes to its updated beliefs. If unchecked, this latitude renders commitment semantics meaningless. Yet, for reasons briefly mentioned earlier, requiring the agent to adopt a commitment-constrained policy from its current state given its new transition model is untenable (there might exist no such policy). This is an open problem that we are starting to tackle.

Conclusion

This paper establishes a position on what the semantics of a commitment should be when agents making and pursuing commitments are uncertain not only about how their actions will affect the world but also about how their preferences about the rewards associated with states of the world might change. In a nutshell, our argument is that the emphasis of most prior work that viewed commitments in terms of intended states of the world is somewhat misguided. Instead, we advocate for a semantics where an agent's commitments are to what it can control—its own actions—and thus fulfilling a commitment corresponds to pursuing an action policy, beginning at the time the commitment was made, that has sufficient likelihood of coercing the world into a desirable state. In this semantics, by "acting in good faith" an agent fulfills its commitment even if the desirable state is not reached. We have described algorithms, based on these semantics, that operationalize foundational concepts about when an agent is permitted to drop a committed-to goal, and more importantly that guide agents' decisions to act in good faith until such a goal is met or dropped. Our CCIMR algorithm focuses on failures due to the world reaching a state where the committed-to goal cannot be achieved, and bounds the degree to which an agent can change its plans in a self-interested way (as its goals evolve) without introducing additional risk of such failure as a side effect. Our CCIMR-ROB algorithm confronts the second case, where an agent might in some circumstances want to act on its local goals that perforce cause the committed-to goal to fail; in this case, the probabilistic commitment can explicitly account for the chances that the agent will decide that it doesn't want to achieve the goal. These algorithms represent potential starting points in a broader exploration of the semantics and utilization of commitments to coordinate sequential decision-making agents in highly-uncertain environments.

A number of future directions have already been identified in this paper since the work reported here is work in progress. These include a more careful characterization and empirical understanding of the CCIMR and CCIMR-ROB algorithms, and the extension of probabilistic commitment

semantics to cases involving other forms of model uncertainty besides reward uncertainty. In addition, the work described in this paper has been largely agnostic as to the source of a commitment, and instead has emphasized what it means for an agent to "act in good faith" to fulfill a probabilistic commitment that it has made. We have pointed out situations where, depending on the algorithm (MR, CCIMR, EBS) being used, different commitment probabilities and/or timings might optimally balance the commitment's utility to the external agent with the flexibility retained for improving rewards while fulfilling it. Important questions thus arise as to why an agent would agree to a commitment in the first place (such as whether to optimize some collective performance or to improve an outcome based on its subjective view (Doshi 2012)), and how it would find the "best" commitment for such purposes in more efficient ways than the enumerative process we used to show performance profiles in our experiments.

Acknowledgments.

Inn-Tung (Anna) Chen, Stefan Witwicki, and Alexander Gutierrez contributed to the ideas, as well as the formulation and experimentation of the algorithms, described in this paper. This work was supported in part by the Air Force Office of Scientific Research under grant FA9550-15-1-0039.

References

Agotnes, T.; Goranko, V.; and Jamroga, W. 2007. Strategic commitment and release in logics for multi-agent systems (extended abstract). Technical Report IfI-08-01, Clausthal University.

Al-Saqqar, F.; Bentahar, J.; Sultan, K.; and El-Menshawy, M. 2014. On the interaction between knowledge and social commitments in multi-agent systems. *Applied Intelligence* 41(1):235–259.

Bannazadeh, H., and Leon-Garcia, A. 2010. A distributed probabilistic commitment control algorithm for service-oriented systems. *IEEE Transactions on Network and Service Management* 7(4):204–217.

Becker, R.; Zilberstein, S.; and Lesser, V. R. 2004. Decentralized markov decision processes with event-driven interactions. In *3rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2004)*, 302–309.

Bernstein, D. S.; Givan, R.; Immerman, N.; and Zilberstein, S. 2002. The complexity of decentralized control of markov decision processes. *Math. Oper. Res.* 27(4):819–840.

Castelfranchi, C. 1995. Commitments: From individual intentions to groups and organizations. In *Proceedings of the International Conference on Multiagent Systems*, 41–48.

Chen, I.-T.; Durfee, E.; Singh, S.; and Witwicki, S. 2012. Influence-based multiagent planning under reward uncertainty. In *MSDM (AAMAS Workshop)*.

Chesani, F.; Mello, P.; Montali, M.; and Torroni, P. 2013. Representing and monitoring social commitments using the event calculus. *Autonomous Agents and Multi-Agent Systems* 27(1):85–130.

Cohen, P. R., and Levesque, H. J. 1990. Intention is choice with commitment. *Artificial Intelligence* 42(2-3):213–261.

Doshi, P. 2012. Decision making in complex multiagent contexts: A tale of two frameworks.

Goldman, C. V., and Zilberstein, S. 2004. Decentralized control of cooperative systems: Categorization and complexity analysis. *J. Artif. Intell. Res. (JAIR)* 22:143–174.

Jennings, N. R. 1993. Commitments and conventions: The foundation of coordination in multi-agent systems. *The Knowledge Engineering Review* 8(3):223–250.

Maheswaran, R.; Szekely, P.; Becker, M.; Fitzpatrick, S.; Gati, G.; Jin, J.; Neches, R.; Noori, N.; Rogers, C.; Sanchez, R.; Smyth, K.; and Buskirk, C. V. 2008. Look where you can see: Predictability & criticality metrics for coordination in complex environments. In *AAMAS*.

Mallya, A. U., and Huhns, M. N. 2003. Commitments among agents. *IEEE Internet Computing* 7(4):90–93.

Poupart, P.; Vlassis, N.; Hoey, J.; and Regan, K. 2006. An analytic solution to discrete Bayesian reinforcement learning. In *Proceedings of ICML '06*, 697–704.

Raffia, H. 1982. *The Art and Science of Negotiation*. Harvard University Press, 79 Garden St. (Belknap Press).

Ramachandran, D., and Amir, E. 2007. Bayesian inverse reinforcement learning. In *IJCAI*, 2586–2591.

Sandholm, T., and Lesser, V. R. 2001. Leveled commitment contracts and strategic breach. *Games and Economic Behavior* 35:212–270.

Singh, M. P. 1999. An ontology for commitments in multiagent systems. *Artificial Intelligence in the Law* 7(1):97–113.

Singh, M. P. 2012. Commitments in multiagent systems: Some history, some confusions, some controversies, some prospects. *The Goals of Cognition. Essays in Hon. of C. Castelfranchi* 1–29.

Vokrínek, J.; Komenda, A.; and Pechoucek, M. 2009. Decommitting in multi-agent execution in non-deterministic environment: experimental approach. In *8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, 977–984.

Winikoff, M. 2006. Implementing flexible and robust agent interactions using distributed commitment machines. *Multiagent and Grid Systems* 2(4):365–381.

Witwicki, S., and Durfee, E. 2009. Commitment-based service coordination. *Int. Jour. of Agent-Oriented Software Engineering* 3(1):59–87.

Witwicki, S. J., and Durfee, E. H. 2010. Influence-based policy abstraction for weakly-coupled Dec-POMDPs. In *Proceedings of the 20th International Conference on Automated Planning and Scheduling (ICAPS)*, 185–192.

Witwicki, S.; Chen, I.-T.; Durfee, E.; and Singh, S. 2012. Planning and evaluating multiagent influences under reward uncertainty (extended abstract). In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 1277–1278.

Xing, J., and Singh, M. P. 2001. Formalization of commitment-based agent interaction. In *Proceedings of the 2001 ACM Symposium on Applied Computing (SAC)*, 115–120.

Xuan, P., and Lesser, V. 1999. Incorporating Uncertainty in Agent Commitments. *International Workshop on Agent Theories, Architectures, and Languages (ATAL-99)* 57–70.