On the Trustworthy Fulfillment of Commitments

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

Abstract

An agent that adopts a commitment to another agent should act so as to bring about a state of the world meeting the specifications of the commitment. Thus, by faithfully pursuing a commitment, an agent can be trusted to make sequential decisions that it believes can cause an intended state to arise. In general, though, an agent's actions will have uncertain outcomes, and thus reaching an intended state cannot be guaranteed. For such sequential decision settings with uncertainty, therefore, commitments can only be probabilistic. We propose a semantics for the trustworthy fulfillment of a probabilistic commitment that hinges on whether the agent followed a policy that would be expected to achieve an intended state with sufficient likelihood, rather than on whether the intended state was actually reached. We have developed and evaluated algorithms that provably operationalize this semantics, with different tradeoffs between responsiveness and computational overhead. We also discuss opportunities and challenges in extending our proposed semantics to richer forms of uncertainty, and to other agent architectures besides the decision-theoretic agents that have been our initial focus of study. Finally, we consider the implications of our semantics on how trust might be established and confirmed in open agent systems.

1 Motivation

In open systems occupied by multiple (human and/or artificial) agents, where in order to best achieve their goals the agents need to rely on each other, issues of trust come to the fore. The many important challenges regarding trust include how to build trust, how to maintain trust, how to utilize trust, how to detect when trust is misplaced, how to propagate reputations about trustworthiness, and how to incentivize trustworthiness. In this paper, we consider the case where agents have incentive to be trustworthy, and where they will do everything in their power to merit the trust of other agents, but where uncertainty inherent in the environment, and the agents' uncertainty about the environment, mean that an agent sometimes cannot, or even should not, achieve outcomes that other agents might be relying upon.

Specifically, we ask the question of whether an agent that has made a (social) commitment to another agent, and has acted in good faith with respect to the commitment, can be said to have fulfilled the commitment even

In: J. Zhang, R. Cohen, and M. Sensoy (eds.): Proceedings of the 18th International Workshop on Trust in Agent Societies, Singapore, 09-MAY-2016, published at http://ceur-ws.org

if the outcomes of its actions fail to reach a state that the commitment was intended to bring about. By acting in good faith, has the agent earned trust despite not delivering the intended outcome? For instance, consider a physician who is treating a patient for some condition: There is a (perhaps implicit) commitment by the doctor to improve that condition, but the treatments administered might not be effective for the patient, or the doctor might even abandon treatment of the condition to address more consequential conditions instead. Has the doctor violated the trust of the patient and failed to meet the commitment?

To answer such questions about what it means for a computational agent to be trustworthy in its fulfillment of social commitments, we have been developing a decision-theoretic formulation for framing and studying the semantics of commitments in settings where the agents' environment, and/or what the agent knows about the environment, can be uncertain. To jump to the punchline of this paper, our investigations so far have led us to advocate a commitment semantics that focuses on whether the agent's choices of actions were consistent with the trustworthy fulfillment of its commitments (e.g., whether the physician followed proper standards of patient care), rather than on whether the state of the world reached by the agent's actions, coupled with the other partially-known ongoing processes and factors in the environment, had the intended outcome (e.g., whether the patient's condition was cured).

For reasons of publication restrictions we cannot here present a deeper technical treatment of our work, so our goal in this paper is to summarize the background, context, justification, and implications of adopting our suggested semantics for the trustworthy fulfillment of commitments, for consideration by the community. In what follows, we first briefly examine some of the relevant past literature on computational models of commitment, with a particular focus on commitments in uncertain, probabilistic worlds.[1] We then (Section 3) present a decision-theoretic formulation of the problem, highlighting the representation of the agent's uncertainty about rewards, and how the fact that pursuing commitments is a sequential process means that the agent might want to change its course of actions sometime between when the commitment is made and when the conditions it is trying to achieve could come about. This formulation then allows us to more formally state the semantics for trustworthy commitment achievement in the face of uncertain rewards that we advocate, and to summarize computational strategies for realizing those semantics (Section 4) for reward uncertainty and beyond. In Section 5, we speculate on how the semantics might apply to other, non-decision-theoretic agent architectures, and in Section 6 we consider the broader implications of our semantics on the problem of establishing trust.

2 Computational Models of Commitment

Munindar Singh provides a comprehensive overview of computational research into characterizing commitments using formal (modal and temporal) logic [Sin12], drawing on a broad literature (e.g., [CL90, Cas95, Sin99, MH03, CMMT13, ASBSEM14]). 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., [Jen93, XS01, Win06]). 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 coordination. 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 preferred goal?

To tighten predictability, commitments can be paired with conditions under which they are sure to hold [Raf82, Sin12, VKP09, AGJ07]. 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 [SL01] noted difficulties

[1] The material presented in Sections 2 and 3 has appeared in similar form in an unpublished symposium paper [DS15], and summaries of the algorithms presented in Section 4 appeared in an extended abstract at AAMAS12 [WCDS12].

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 [XL99, BLG10, WD09]. Xuan and Lesser [XL99] 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 [MSB$^+$08], 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, our treatment of commitments in uncertain settings is not only to provide predictive information to the recipient of a commitment about what might happen, but also to impose prescriptive semantics on the provider of a commitment to guide its behavior into a good faith effort in making those predictions come true.

3 Problem Formulation

Our initial 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 later we briefly 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[2] 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 built-in 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 notions of commitment and trust come 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 $U(\xi')$. Depending on the degree of cooperativeness of the agent, solving the agent's sequential decision problem might require taking into account these enablement-utility (U) values to the user of candidate enablement-commitments. If the agent adopts a commitment to the user, the user becomes a "trustor" of the agent and the agent a "trustee" of the user.

[2] Because the model is assumed accurate, the agent can be assumed to only formulate policies (and thus commitments) that it is capable of executing. Permitting inaccurate models (where an agent might make a commitment it is inherently incapable of fulfilling) is outside the scope of the focus of this paper on trustworthy fulfillment of commitments.

Some special cases of this formulation help motivate our commitment semantics:

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 standard 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[3] 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 [PVHR06].

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 uncertain 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 meets the commitment with a probability at least as high as the probability associated with 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 a self-interested setting, the agent's reward could predominate (the user is helped only as a side-effect of the agent's preferred policy), or in a setting where the agent is subordinate the user's utility could be preeminent.

Commitment in the face of Uncertain Rewards. This special case has been the main focus of our work, 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 on 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? For example, perhaps the agent's modified beliefs about rewards would tempt it to change its behavior in a way that actually improves the chances of achieving the intended conditions of the commitment, in a "win-win" way. But would changing its policy even then violate the trust of the user?

4 Commitment Semantics

We argue that a semantics for commitments in sequential-decision settings with stochastic transitions, 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.* Remember that "in expectation" in this context means that the probability of meeting the commitment is at least as high as the probability specified by the probabilistic commitment. So, if the commitment was to reach an intended state with probability 0.9, the agent would need to follow a policy that in expectation would reach the state at least 90% of the time, while if the commitment probability was only 0.1 the agent could follow a policy that in expectation would reach the state only 10% of the time. Thus, "in expectation" does not mean "as likely has possible." Nor does it mean "more likely than not" (better than a 50% chance). Instead, it means "at least as likely as promised by the commitment."

This semantics sounds straightforward enough, though as the sections that follow will show it is not always trivial to operationalize. Before considering algorithms for implementing the semantics, however, we first briefly consider how this semantics departs from prior semantics for computational commitments.

[3] Recall that a policy is defined over *all* (belief) states, and so covers every possible contingency that could arise during execution. We refer to a particular sequence of states and (policy-dictated) actions that might be experienced as a *trajectory.* Note that a policy thus differs from a *plan*, which is typically defined in terms of a specific (nominal) trajectory. Hence, a plan can fail (stimulating plan repair or replanning) when unintended action outcomes or external events cause a deviation from the plan's nominal trajectory. In contrast, a policy never "fails" because it specifies actions for every state (and thus for every possible trajectory).

4.1 Relationship to Other Commitment Semantics

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 environments with stochastic transitions 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 focuses 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, our commitment semantics shares similarities with work on joint policies in cooperative sequential decision 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. This is exactly the kind of flexibility that we seek to exploit when an agent is acting sequentially under reward uncertainty.

Our commitment semantics also hearkens back to some of the earliest work on agent commitments, which focused not on (social) commitments between agents, but rather on commitments an agent makes to its internal behavior as part of a meta-control strategy. The work by Kinny and Georgeff [KG91] considered the degree to which an agent should question continuing to pursue a plan in an uncertain world, where they explored strategies by which an agent might be "cautious" (reconsidering what plan it should follow every step of the way) or "bold" (pursuing its current plan until it is either finished, or is impossible to continue). Like that work, our semantics for commitment concentrates on commitments to action policies rather than outcomes, but unlike that work we view a (in our case social) commitment as a constraint on possible physical action policy choices rather than as a meta-construct for controlling reasoning effort.

4.2 Semantics-Respecting Algorithms

Of the algorithms we now summarize, the first can arguably be seen as being "bold" because it presses on with a policy without being responsive to changing circumstances, and thus avoids the overhead of questioning whether and how to respond to circumstances every step of the way. The second is "cautious" because it preplans for every foreseeable change to circumstances. As a result, it is extremely responsive, but incurs high reasoning costs. The third is a compromise between the first two, striving to be responsive enough without incurring excessive overhead.

4.2.1 Mean Reward (MR)

This algorithm most simply and directly implements our commitment semantics so that an agent can be trustworthy in fulfilling the commitment. Given a commitment and a distribution over the true reward function, the agent finds its Bayes-optimal policy that meets the probabilistic commitment. Specifically, at the outset, the agent formulates a commitment-constrained policy that is optimal for its initial reward belief-state, which equates [RA07] to an optimal policy for the distribution's Mean Reward (MR). The agent then pursues this policy without deviation, ignoring temptations and opportunities that arise during execution as it improves its understanding of the true reward function in its environment. Thus, the MR algorithm implements a "bold" strategy for commitment attainment using our semantics: The agent adheres to a policy that meets the commitment, and never reconsiders it. This satisfies our commitment semantics, and incurs computational cost only for computing an optimal policy given a single (mean) reward function; its downside is that it will not take advantage of serendipity, when new information about rewards would have allowed it to achieve higher reward while still meeting (or even exceeding) the probabilistic expectations of the commitment.

4.2.2 Extended Belief State (EBS)

This algorithm implements the most "cautious" of strategies by modeling all possible ways in which the agent's beliefs about the reward might change over its sequential actions, and developing a policy that accounts for every single one of them (while still meeting or exceeding the probabilistic commitment). The Extended Belief State approach adds directly into the state model a representation of the agent's beliefs about the reward function. Thus, as the agent models possible trajectories, it considers not only its choices of actions and their stochastic outcomes on the physical state, but also the possible reward observations it might make and the consequent posterior beliefs about the reward function it might have. The branching associated with action choices, action outcome stochasticity, and uncertainty over reward observations exponentially enlarges the number of trajectories the agent needs to explore, incurring high computational overhead. However, once the EBS policy has been derived, it is guaranteed to be optimal, not only being responsive to all possible evolving models of the environment's rewards, but even leading the agent to acting in ways that anticipate and exploit expected future reward observations.

4.2.3 Commitment-Constrained Iterative Mean Reward (CCIMR)

This algorithm is a compromise between the previous algorithms, seeking to gain some of the computational benefits of MR while permitting some degree of responsiveness like EBS, all while being trustworthy in adhering to the commitment semantics. Conceptually, the algorithm works with a space of policies Π_ξ that all satisfy the commitment ξ. If this space is empty, then the agent cannot make a commitment, but otherwise our commitment semantics allow the agent to commit to following *any* of these policies, as they are all equally satisfactory given commitment ξ. The crux of our Commitment-Constrained Iterative Mean Reward (CCIMR) algorithm, then, is to use a reward observation reactively (unlike EBS that proactively anticipates them) to compute a new posterior mean reward. It then selects from the current Π_ξ the policy that optimizes expected reward under the new reward beliefs and pursues that one. Note, though, that Π_ξ will shrink over time: as a particular policy has been followed to the current time, policies that would have chosen different actions in the states up until that time must be removed from Π_ξ. (A policy that appends the first half of one element of Π_ξ with the second half of another element might not itself be an element of Π_ξ, and so mustn't be allowed.)

While this gives the conceptual basis of our CCIMR algorithm, a key aspect of this algorithm is that it does *not* explicitly enumerate and manipulate the commitment-constrained set of policies Π_ξ, as this set can be exceedingly large. Instead, we have developed a linear programming approach that explicitly captures constraints, including not just those associated with the commitment(s) but also those associated with the policy pursued so far, so that the agent can construct, at any given time, only the optimal element of the (possibly shrunken) space of policies Π_ξ. This means CCIMR is solving a number of optimal policy calculations that is linear in time (one for each new mean reward, which at most happens once per time step), whereas MR only performs one optimal policy calculation (the initial MR policy), and EBS computes one optimal policy but for an exponentially-larger (belief) state space. CCIMR thus represents a compromise in terms of computation. Meanwhile, because it is responsive to changing reward beliefs, it is guaranteed to achieve rewards no worse than MR, while achieving rewards no better than EBS (because EBS is not only responsive but proactive).

4.2.4 Evaluation

Again, the technical details of the algorithms just described, including formal proofs about CCIMR being lower-bounded by MR and upper-bounded by EBS, and a proof that CCIMR conforms to our commitment semantics, are excluded from this paper due to publication restrictions. Similarly, we cannot present detailed empirical results for these algorithms. In brief, though, our experiments have shown that CCIMR can often allow an agent to achieve rewards close to what EBS permits (and significantly better than MR), and that scaling a problem to more states and longer time horizons can cause the time needs of EBS to explode while CCIMR's time needs increase more manageably.

4.3 Semantics with 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 the probabilities of action outcomes and thus to formulate good policies. 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 might 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. The question of what constitutes trustworthy pursuit of a commitment by a decision-theoretic agent, in the face of transition uncertainty, is an open problem that we are starting to tackle.

5 Implications for Non-Decision-Theoretic Agents

By framing the question of trustworthy pursuit of commitments despite uncertainty in a decision-theoretic setting, we have been able to exploit useful aspects of a decision-theoretic framework, such as explicit models of rewards and transitions, and of uncertainty over these. However, decision-theoretic approaches have weaknesses too, and in particular the power of these approaches generally comes at a steep computational cost. As a result, these techniques are often applied to only small problems, or (for particular classes of problems) approximation techniques are used to find good, but not necessarily optimal, solutions. Other agent architectures, such as those based on Belief-Desire-Intention concepts, can often be fruitfully applied to problems where the richness of the decision-theoretic approach is not needed and where the computational costs cannot be afforded. The question then arises as to whether our proposed semantics generalizes to other architectures based on more classical notions of plans and goals. While at this point we can only speculate about this, our initial thinking is that they can, and we now delve briefly into initial thoughts as to how.

In essence, mapping our commitment semantics to an agent using classical plans and goals would mean that such an agent is committing not to achieving particular goals (conditions in the world) but rather it is committing to executing one out of a set of plans. Therefore, before making a commitment, the agent would need to have identified a non-empty set of plans to commit to. Compared to the decision-theoretic setting, where we could summarize the space of commitment-satisfying policies based on a small number of parameters, it is less clear how (without incorporating new annotations to elements of a plan library) we could do the same in a more classical setting, but assuming such a set is specified (through enumeration if in no other way) then an agent can commit to executing one of the plans in that space, where each plan is executable when particular preconditions it depends upon are met.

The operational semantics for pursuing other goals that could conflict (depending on how they are pursued) with the goal(s) of the committed-to plans could then be patterned after the semantics we've outlined. For example, the semantics could dictate that an agent is not allowed pursue a plan for another goal if that plan would violate the preconditions for the element of the space of committed-to plans that is currently being followed. This would be like MR: the agent commits to a particular plan for the commitment, and it must follow that plan (in that it cannot take actions that preclude following the plan). The exception, however, is whereas the MR policy would associate an action for all eventualities that could arise, a typical plan will have a single nominal trajectory. If the world state deviates from this trajectory then the plan fails, triggering possible repairs or adoption of alternative plans. It is possible, however, that no recourse is possible. Again, based on our semantics, the agent still met its commitment because it followed the plan as promised.

It could be that, during pursuit of its plan, the agent acquires new or different goals, whose satisfaction would be incompatible with a plan it had committed to. The challenge then is determining whether it is possible to change its commitment plan so that it can pursue its new goal(s), while still being trustworthy regarding its commitment. Like CCIMR, the agent could instead commit to pursuing one plan out of a set of plans, and as its goals change could dynamically select from among the subset whose preconditions are still met. If the different plans are differentially susceptible to failure, then a commitment to the set of plans is only as strong as the most failure-prone of the set. This in turn suggests the need for some sort of "reliability" ordering over plans. Strategies for determining such orderings could be based on probabilistic knowledge (gravitating back towards decision-theoretic ideas), or could be based on a form of adversarial analysis to find the worst-case outcome of a plan given a purposely adversarial environment.

6 Conclusions

In this paper, we argue in favor of an operational semantics for commitments based on what an agent can control—its own actions. 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 an intended state. In this semantics, by "acting in good faith" an agent fulfills its commitment even if the intended state is not reached. We have summarized 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. 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, and we have speculated as to the transferability of these notions to agents other than decision-theoretic agents.

As advertised at the beginning of this paper, our emphasis has been on how an agent that is abiding by its commitments should constrain its behaviors (the space of action policies it considers) to act in good faith on the commitment. If we connect back to a human example of this, where a doctor is trustworthy if she follows standards of care even if a particular patient does not do well, then interesting questions arise as to how external parties, rather than the agent itself, can actually confirm trustworthiness. In the US, questions of medical trust often rely on an unbiased expert stepping through the decision making of a doctor to assess that appropriate sequential decision were being made. In open systems where trust needs to be earned, interesting questions arise as to how easy it would be to know whether to bestow trust on an agent, where the outcomes of the decisions are observable but the decisions, and the processes used to reach the decisions, are not. Perhaps notions of certification play a more significant role for this form of trust, where a certifying body has evaluated and approved the decision-making processes of an agent. These are potentially interesting topics for future consideration.

6.0.1 Acknowledgments

Inn-Tung (Anna) Chen, Stefan Witwicki, Alexander Gutierrez, and Qi Zhang contributed to various aspects of the ideas and algorithms described in this paper. We also benefited from detailed discussions about commitments with Munindar Singh and Scott Gerard, and from the questions and suggestions of the anonymous reviewers. This work was supported in part by the Air Force Office of Scientific Research under grant FA9550-15-1-0039.

References

[AGJ07] Thomas Agotnes, Valentin Goranko, and Wojciech Jamroga. Strategic commitment and release in logics for multi-agent systems (extended abstract). Technical Report IfI-08-01, Clausthal University, 2007.

[ASBSEM14] Faisal Al-Saqqar, Jamal Bentahar, Khalid Sultan, and Mohamed El-Menshawy. On the interaction between knowledge and social commitments in multi-agent systems. *Applied Intelligence*, 41(1):235–259, 2014.

[BLG10] Hadi Bannazadeh and Alberto Leon-Garcia. A distributed probabilistic commitment control algorithm for service-oriented systems. *IEEE Transactions on Network and Service Management*, 7(4):204–217, 2010.

[Cas95] Cristiano Castelfranchi. Commitments: From individual intentions to groups and organizations. In *Proceedings of the International Conference on Multiagent Systems*, pages 41–48, 1995.

[CL90] Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(2-3):213–261, 1990.

[CMMT13] Federico Chesani, Paola Mello, Marco Montali, and Paolo Torroni. Representing and monitoring social commitments using the event calculus. *Autonomous Agents and Multi-Agent Systems*, 27(1):85–130, 2013.

[DS15] Edmund H. Durfee and Satinder Singh. Commitment semantics for sequential decision making under reward uncertainty. In *Papers from the AAAI Fall Symposium on Sequential Decision Making for Intelligent Agents (AAAI Tech Report FS-15-06)*, 2015.

[Jen93] N. R. Jennings. Commitments and conventions: The foundation of coordination in multi-agent systems. *The Knowledge Engineering Review*, 8(3):223–250, 1993.

[KG91] David N. Kinny and Michael P. Georgeff. Commitment and effectiveness of situated agents. In *Proceedings of the 1991 International Joint Conference on Artificial Intelligence (IJCAI-91)*, pages 84–88, 1991.

[MH03] Ashok U. Mallya and Michael N. Huhns. Commitments among agents. *IEEE Internet Computing*, 7(4):90–93, 2003.

[MSB$^+$08] Rajiv Maheswaran, Pedro Szekely, Marcel Becker, Stephen Fitzpatrick, Gergely Gati, Jing Jin, Robert Neches, Nader Noori, Craig Rogers, Romeo Sanchez, Kevin Smyth, and Chris Van Buskirk. Look where you can see: Predictability & criticality metrics for coordination in complex environments. In *AAMAS*, 2008.

[PVHR06] Pascal Poupart, Nikos Vlassis, Jesse Hoey, and Kevin Regan. An analytic solution to discrete Bayesian reinforcement learning. In *Proceedings of ICML '06*, pages 697–704, 2006.

[RA07] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *IJCAI*, pages 2586–2591, 2007.

[Raf82] H. Raffia. *The Art and Science of Negotiation.* Harvard University Press, 79 Garden St. (Belknap Press), 1982.

[Sin99] Munindar P. Singh. An ontology for commitments in multiagent systems. *Artificial Intelligence in the Law*, 7(1):97–113, 1999.

[Sin12] Munindar P. Singh. Commitments in multiagent systems: Some history, some confusions, some controversies, some prospects. *The Goals of Cognition. Essays in Hon. of C. Castelfranchi*, pages 1–29, 2012.

[SL01] Tuomas Sandholm and Victor R. Lesser. Leveled commitment contracts and strategic breach. *Games and Economic Behavior*, 35:212–270, 2001.

[VKP09] Jirí Vokrínek, Antonín Komenda, and Michal Pechoucek. Decommitting in multi-agent execution in non-deterministic environment: experimental approach. In *8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, pages 977–984, 2009.

[WCDS12] Stefan Witwicki, Inn-Tung Chen, Edmund Durfee, and Satinder Singh. Planning and evaluating multiagent influences under reward uncertainty (extended abstract). In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1277–1278, 2012.

[WD09] Stefan Witwicki and Edmund Durfee. Commitment-based service coordination. *Int. Jour. of Agent-Oriented Software Engineering*, 3(1):59–87, 2009.

[Win06] Michael Winikoff. Implementing flexible and robust agent interactions using distributed commitment machines. *Multiagent and Grid Systems*, 2(4):365–381, 2006.

[XL99] Ping Xuan and Victor Lesser. Incorporating Uncertainty in Agent Commitments. *International Workshop on Agent Theories, Architectures, and Languages (ATAL-99)*, pages 57–70, January 1999.

[XS01] Jie Xing and Munindar P. Singh. Formalization of commitment-based agent interaction. In *Proceedings of the 2001 ACM Symposium on Applied Computing (SAC)*, pages 115–120, 2001.