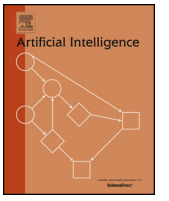
Artificial Intelligence 317 (2023) 103864

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint



Risk-aware analysis for interpretations of probabilistic achievement and maintenance commitments ☆

Qi Zhang [a,*], Edmund H. Durfee [b], Satinder Singh [b]

[a] University of South Carolina, United States of America
[b] University of Michigan, United States of America

ARTICLE INFO

Article history:
Received 15 January 2022
Received in revised form 16 January 2023
Accepted 21 January 2023
Available online 26 January 2023

Keywords:
Probabilistic commitment
Cooperative multiagent planning
Planning under uncertainty
Achievement commitment
Maintenance commitment

ABSTRACT

Probabilistic commitments provide a computational framework for multi-agent coordination, where one autonomous agent (the commitment provider), commits to a future course of action that probabilistically influences the local state of another agent (the commitment recipient) in ways that the recipient desires. Conventionally, a probabilistic commitment is specified abstractly so as to give the provider latitude at run time about how to achieve it. Unfortunately, as we analyze in this article, this abstraction incurs a risk of suboptimal performance for the recipient. For (achievement) commitments by the provider to achieve conditions that the recipient prefers but that do not initially hold, we prove that the recipient can make modeling choices that bound its risk of suboptimality. Somewhat surprisingly, however, for (maintenance) commitments by the provider to maintain conditions whose initial values are already ones the recipient prefers, we prove that no such bounds on suboptimality risk are possible. We study the two types of commitments empirically to measure the suboptimality they incur under different conditions, and based on our theoretical and empirical results suggest that adding selective details when specifying probabilistic maintenance commitments can be beneficial.

1. Introduction

1.1. Motivation

A cooperative artificial intelligence (AI) system involves multiple autonomous agents collectively accomplishing shared tasks, with examples including drone delivery systems, household robots, and smart manufacturing lines. Successful cooperation generally requires coordination among agents, because an individual agent's actions might not yield desired outcomes unless others also act in concert. While the most efficient coordination can be best achieved by a centralized decision maker that directly controls all agents, in many interesting settings the agents are physically distributed, sensing and communication are limited, and centralization introduces vulnerabilities and computational bottlenecks. Thus, we focus on multiagent systems where each individual agent needs to make its own decisions based on its local information and on judiciously-chosen information communicated by others.

☆ This paper is part of the Special Issue: "Risk-aware Autonomous Systems: Theory and Practice".

* Corresponding author.
E-mail address: qz5@cse.sc.edu (Q. Zhang).

https://doi.org/10.1016/j.artint.2023.103864
0004-3702/

Inspired by how people coordinate in many settings, one important form of information agents can exchange is that of a *commitment*, which refers to an agent making credible and prolonged promises about various aspects of the consequences of its future actions, addressing the coordination problem in a principled manner. For example, consider the scenario of multiple cars navigating through an intersection. While there is a large amount of information that each car could share (e.g., the status of its battery and fuel supply, its maintenance history, how alert its driver is, its driver's level of impatience, its destination and deadline for arriving there, etc.), what is most germane for intersection safety is that other cars know which road out of the intersection it intends to take. Thus, turn signals can be viewed as credible commitments that other cars can use to anticipate the consequences of their own actions, and hence cars can coordinate their actions to move through an intersection more safely and efficiently by each choosing which turn signals to and not to activate.

In this article, we specifically focus on the two-agent scenario, where the agent that makes a commitment (e.g., signals its intended turn) is referred to as the commitment *provider*, and the other agent that uses the commitment to improve its own plans (e.g., avoid a collision) as the commitment *recipient*. In a general sense, a commitment is an abstraction of the provider's future behavior, summarizing the most important (to the recipient) effects the provider's future actions will have on the recipient's environment, and hence on the recipient's decisions.

When stochasticity is inherent in the environment, the provider cannot always firmly guarantee to bring about particular outcomes, and in fact could discover after committing that its plan to pursue those outcomes is more costly or risky than it had previously realized. Under such circumstances, commitments are conditional [1]. While in principle the provider could enumerate and include in the specification to the recipient the conditions upon which meeting the commitment's outcomes are predicated, in practice this is not generally sensible: even if the provider can afford the expense of enumerating and communicating them, the recipient will generally be unable to observe them, much less predict ahead of time how likely they are to hold. Therefore, a provider in such settings should instead provide a *probabilistic* commitment, using its local model to arrive at a summary probability of the necessary conditions holding for the outcomes to be realized.

1.2. Contributions

Prior work has focused on semantics and mechanisms for the provider to follow to faithfully pursue its commitments despite uncertainty [2–5]. That work held that a probabilistic commitment should be considered fulfilled if the provider's actions would have brought about the desired outcome by the promised time with at least the promised probability, even if in a particular instance the desired outcome was not realized. In this vein, the focus was largely on the provider's pursuit of *achievement* commitments [6–9], where the provider commits to changing some features of the state in a way desired by the recipient with some probability by some time. For example, the recipient plans to take an action (e.g., move from one room to another) with a precondition (e.g., the door separating rooms is open) that the provider has promised to likely enable by some deadline. In this article, we also consider another form of commitment, a *maintenance* commitment, where the provider instead commits to a course of action that, up until a promised time, is sufficiently unlikely to change features that are already the way the recipient wants them maintained. After that time, the provider can freely change the features. For example, a door the recipient wants open might initially be so, but the provider wants to close it to clean behind it during housekeeping tasks. The provider could postpone closing it (clean elsewhere first), but by changing other doors while cleaning elsewhere it might accidentally introduce a draft that could prematurely close the door the recipient wants left open.

Even though decision-theoretic formulations of, and reasoning methods for, achievement and maintenance commitments are nearly identical, prior work has found it much harder to successfully coordinate for maintenance than achievement [10–12]. That is, maintenance commitments have appeared to be riskier to trust. In the past, it has been assumed that the difficulty lies on the provider's side—that it might be inherently harder for a provider to find good policies that maintain a feature than to change it. However, in this article we claim and justify that instead the challenge actually lies on the recipient's side. Although a commitment abstracts the provider's behavior to summarize information about the timing and likelihood of the key effects, such abstraction is intentionally incomplete, omitting details that leave the recipient with residual uncertainty. Our core claim is that ***it is fundamentally harder for the recipient to manage the risk associated with the residual uncertainty in a maintenance commitment than in an achievement commitment.*** Specifically, here we study the type of abstraction where the probabilistic commitment specifies the provider's behavior at a single time step, which is a widely-adopted abstraction because with it computationally efficient solutions for choosing optimal commitments exist [13,14], and it gives a provider that is learning aspects of its local world model during execution more latitude for improving its behavior based on what it learns, as illustrated in Section 2.2 and more comprehensively investigated in [9,15,16].

We substantiate our core claim both theoretically and empirically. We theoretically frame the recipient's problem of risk-aware interpretation of an achievement or maintenance commitment as a robust planning problem against the incomplete information about the provider's impacts on the environment specified in the commitment. We then begin by analyzing a straightforward strategy to deal with such incomplete information, adopted in previous work, where the recipient minimizes risk by modeling an achievement commitment pessimistically—it assumes the feature will not (probabilistically) attain its desired value any earlier than the commitment's promised time. We show analytically that the worst-case suboptimality induced by such pessimism can be bounded fairly tightly. For the maintenance counterpart, however, we show that adopting the same form of pessimistic model fails to bound the risked suboptimality. We advance our theoretical analysis by motivating several alternative interpretation strategies that attempt to bound risk, yet for maintenance the suboptimality induced by any of the strategies remains effectively unbounded. We then empirically measure the realized suboptimality

for the alternative interpretation strategies in various settings, and the results show that there is no choice of risk-aware interpretation strategy the recipient can adopt for maintenance commitments that reliably limits the suboptimality of coordination with the provider. Our results suggest that successful maintenance commitments will generally require that the provider's and recipient's plans be more tightly coupled than for achievement commitments, intentionally sacrificing some of the provider's autonomy in order to reduce the recipient's risk.

1.3. Related work

Others have adopted alternative frameworks, such as conditional commitments [17,18] and contracting frameworks [19], for managing uncertainty when a commitment is being pursued. In this vein, there has been substantial work for developing protocols for agents who are modeling and communicating about commitments. The focus is on the lifecycle of a commitment [20–22], from its initial proposed creation, to the mutual agreement to adopt it, to determining whether it has been fulfilled, to whether it is time to abandon it or replace it with a better commitment. Over the lifecycle, it is important that interacting agents engage in a communication protocol that ensures their beliefs about the status of a shared commitment are aligned. In our work, we adopt the probabilistic commitment framework to study both achievement and maintenance commitments, and focus just on the "detached" stage of the commitment lifecycle where an agreed-upon commitment is being actively pursued, and the pursuit requires a sequence of actions, where some might not have desired outcomes, or an agent's priorities could change in the midst of executing the sequence.[1]

The probabilistic commitment framework [6,13,25] summarizes the likelihood the commitment will be successfully discharged by a given time, versus violated due to bad luck or a better option appearing. Probabilities let a decision-theoretic recipient optimally hedge for violations while waiting for the provider. Existing work in probabilistic commitments mostly focuses on the provider's side, e.g., developing semantics and planning methods for the provider to faithfully fulfill its commitment despite inherent uncertainty in its environment [16], yet largely overlooks the recipient's risk in modeling a commitment that only partially specifies the provider's behavior. We close this loop in this article by systematically studying how the recipient should robustly plan to manage the risk associated with modeling a probabilistic achievement or maintenance commitment.

While we discuss the notions of achievement and maintenance in multi-agent systems under the framework of probabilistic commitment, there has been other work characterizing achievement and maintenance in alternative frameworks for intelligent agent systems absent multi-agent commitments, primarily based on the Belief-Desire-Intention (BDI) model that categorizes goals for an agent into achievement ones and maintenance ones. Kaminka et al. [26] models teamwork in agents using the BDI model, and, similar to our work, notice that teamwork models focus much more on achievement goals than maintenance goals. They implement mechanisms for collaborative maintenance in two teamwork architectures for situated agent teams on top of Soar [27] and BITE [28]. Baral et al. [29] reveal the limitations of earlier characterizations of maintenance using the notion of stabilizability [30,31], and propose an improved characterization where the agent is situated in an adversarial environment. Duff [32] notices that existing agent systems for maintenance are mostly reactive, and discusses methods for proactive maintenance where the agent acts before a maintenance condition is violated. Our research similarly focuses on proactivity in the sense that a commitment provider chooses actions that are sufficiently unlikely to violate a maintenance commitment over the committed timeframe.

As will be detailed in Section 2, we adopt a subclass of Dec-POMDPs [33], Transition-Decoupled POMDPs (TD-POMDPs) [34], to formulate and analyze the interaction between the provider, the recipient, and their environment, where the commitment is specifying the transition dependence between the two agents. We note here that the framework of commitment can be generalized into other Dec-POMDP subclasses that characterize other forms of inter-agent dependence. For example, Becker et al. [35] considered a form of reward independence that couples multiple agents. In another formulation of transition dependence, Varakantham et al. [36] use the so-called task state values to characterize their weak inter-agent transition dependence. Varakantham et al. mainly discuss the two task state values: *Done* and *Not Done*, which translate to our binary values of the commitment feature as will be detailed in Section 2, but they did not discuss the distinction between achievement and maintenance as in this work. The commitments in this work are formulated for the inter-agent (transition) dependence in TD-POMDPs, and incorporating commitments into other types of DEC-(PO)MDPs would similarly need them to have inter-agent dependence. Our two agents exist in a common MDP but each has its own local MDP and reward, where the reward structure is analogous to that in stochastic games [37] where agents have their own reward functions to model the general self-interested case. In our setting, commitments are motivated when the two agents are cooperative to maximize their joint reward, or even in some non-cooperative cases (e.g., if side payments are possible in the environment). Our analysis focuses only on the recipient's local performance and therefore applies to both cooperative and non-cooperative cases, as long as the agents (for whatever reason) have agreed on a commitment.

[1] A common example used to illustrate the commitment lifecycle is the purchase of a good such as a book [23,24]. When a buyer accepts a seller's offer to sell the book, the commitment becomes active, and then when the seller's antecedent conditions are met (e.g., payment has been made and a shipping address provided by the buyer), the commitment enters the "detached" stage. Within this stage, the seller pursues the commitment by taking multiple steps involving the book, such as retrieving it from a warehouse, packaging it, handing it off for shipment, transporting it, and delivering it to the address. If all of the steps are successful, the commitment transitions to a satisfied state, and otherwise to a violated state. Hence, it is during the "detached" stage that decisions are made about how/whether the active commitment will be met, and the focus of this article is on reasoning about these decisions.

2. Preliminaries

In this section, we begin by describing the decision-theoretic setting we adopt for analyzing probabilistic commitments for the recipient and the provider, including both achievement commitments and maintenance commitments.

The **recipient**'s environment is modeled as a Markov Decision Process (MDP) defined by the tuple $M = (\mathcal{S}, \mathcal{A}, P, R, H, s_0)$ where $\mathcal{S}$ is the finite state space, $\mathcal{A}$ is the finite action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ ($\Delta(\mathcal{S})$ denotes the set of all probability distributions over $\mathcal{S}$) is the transition function, $R : \mathcal{S} \to \mathbb{R}$ is the reward function, H is the finite horizon, and s_0 is the initial state. The state features are explicitly augmented with the time step, such that the state space is partitioned into disjoint sets by the time step, $\mathcal{S} = \bigcup_{h=0}^{H} \mathcal{S}_h$, where states at time step h in $\mathcal{S}_h$ only transition to states at time step $h+1$ in $\mathcal{S}_{h+1}$. The MDP starts in s_0 and terminates in $\mathcal{S}_H$. Given a policy $\pi : \mathcal{S} \to \mathcal{A}$ and starting in the initial state, a random sequence of transitions $\{(s_h, a_h, r_{h+1}, s_{h+1})\}_{h=0}^{H-1}$ is generated by $a_h = \pi(s_h), s_{h+1} \sim P(s_h, a_h), r_{h+1} = R(s_{h+1})$. The value function of policy π is $V_M^\pi(s) = \mathbb{E}[\sum_{h'=h+1}^{H} r_{h'} | \pi, s_h = s]$ where h is such that $s \in \mathcal{S}_h$. The optimal policy for M, denoted as π_M^*, maximizes V_M^π for all $s \in \mathcal{S}$, and its value function $V_M^{\pi_M^*}$ is abbreviated as V_M^*. The value of the initial state is abbreviated as $v_M^\pi = V_M^\pi(s_0)$.

Similarly, the **provider**'s environment is modeled as another MDP with a finite state space, a finite action space, and a finite horizon. As one way to model the interaction between the provider and the recipient, we adopt the Transition-Decoupled POMDP (TD-POMDP) framework [34] where both the recipient's state and the provider's state can be factored into state features. The recipient's state is factored as $s = (l, u)$, where l is the set of all the recipient's state features locally controlled by the recipient, and u is the set of state features shared with the provider. The provider's state features, including u, are all locally controlled by the provider. The provider and the recipient are weakly coupled in the sense that the shared state features u are only controllable by the provider. (For example, u could be features associated with the status of a door that the recipient wants open but can only be opened/closed by the provider.)

Formally, the dynamics of the recipient's state can be factored as

$$\begin{aligned} P(s_{h+1}|s_h, a_h) =& P\left((l_{h+1}, u_{h+1})|(l_h, u_h), a_h\right) \\ =& P_u(u_{h+1}|u_h) P_l\left(l_{h+1}|(l_h, u_h), a_h\right). \end{aligned}$$

We refer to P_u as the true *influence* that the provider exerts on the recipient's environment dynamics [34,38,39], which is the transition function of u that is fully determined by the provider's policy (it is not a function of a_h).

2.1. Commitment semantics

A commitment is concerned with state features u that are shared by both agents but only controllable by the provider. Intuitively, a commitment provides partial information about P_u from which the recipient can plan accordingly. For simplicity in this article, we focus on the setting where the set u contains a single, binary state feature. In a slight abuse of notation, we henceforth refer to this feature as u, where binary value u^+, as opposed to u^-, is the value of u that is desirable for the recipient. Intuitively, u^+ (u^-) stands for an enabled (disabled) precondition needed by the recipient. We will refer to u as the commitment feature. Further, we assume that u can be toggled at most once [40,8,9]. In transactional settings (e.g., [41,42]), a feature changing only once is common, as it is in multiagent planning domains where one agent enables a precondition needed by an action of another. Some cooperative agent work requires agents to return changed features to prior values, and in extreme cases where toggling reliably repeats there may be no need for explicit commitments. While, in general, toggling more than once can be modeled by a series of alternating achievement and maintenance commitments, the fundamental differences between these commitment types are most readily revealed and understood without such complications, and so in what follows we consider the two types separately.

2.1.1. Achievement commitments

Let the initial state be factored as $s_0 = (l_0, u_0)$. For achievement commitments, the initial value of the commitment feature is u^-, i.e. $u_0 = u^-$. The provider commits to pursuing a course of action that can bring about the commitment feature value desirable to the recipient with some lower bound on probability. Formally, an achievement commitment is defined by tuple $c_a = (T_a, p_a)$, where T_a is the achievement commitment time, and p_a is the achievement commitment probability [8,9]. The commitment semantics is that the provider is to follow a policy that sets u to u^+ by time step T_a with *at least* probability p_a, i.e.

$$\Pr(u_{T_a} = u^+ | u_0 = u^-) \geq p_a. \tag{1}$$

When planning with the achievement commitment, the provider finds an optimal policy (one that maximizes its local value) that respects the commitment's semantics. A straightforward way of doing so adopted in prior work solves the provider's planning problem using linear programming (LP) [43], where the commitment semantics are captured simply by adding the above inequality as an additional constraint to the LP [13,44]. Specifically, the provider's planning problem can be solved with the linear program in Equation (2) [13], where $(s^{\mathrm{p}}, a^{\mathrm{p}})$ denotes the provider's state-action pair; $R^{\mathrm{p}}(s^{\mathrm{p}}, a^{\mathrm{p}})$ and $P^{\mathrm{p}}(s^{\mathrm{p}}, a^{\mathrm{p}})$ are the provider's reward and transition function, respectively; decision variable x^{p} is the provider's occupancy

measure (where $x^{\mathrm{p}}(s^{\mathrm{p}}, a^{\mathrm{p}})$ is the expected number of times action a^{p} will be taken in state s^{p}); $\delta_{i,j}$ is the Kronecker delta which is 1 if variables i and j are equal, and 0 otherwise; constraints (2b)(2c) guarantee that x^{p} is a valid occupancy measure; and constraint (2d) expresses the semantics of the achievement commitment from Equation (1), noting that, since the state features are explicitly augmented with the time step and therefore each state is never re-visited, the occupancy measures can be viewed as probability measures, as similarly treated in prior work [43,45].

$$\max_{x^{\mathrm{p}}} \sum_{s^{\mathrm{p}}, a^{\mathrm{p}}} x^{\mathrm{p}}(s^{\mathrm{p}}, a^{\mathrm{p}}) R^{\mathrm{p}}(s^{\mathrm{p}}, a^{\mathrm{p}}) \tag{2a}$$

$$\text{s.t. } \forall s^{\mathrm{p}}, a^{\mathrm{p}} \quad x^{\mathrm{p}}(s^{\mathrm{p}}, a^{\mathrm{p}}) \geq 0; \tag{2b}$$

$$\forall s^{\mathrm{p}\prime} \quad \textstyle\sum_{a^{\mathrm{p}\prime}} x^{\mathrm{p}}(s^{\mathrm{p}\prime}, a^{\mathrm{p}\prime}) = \sum_{s^{\mathrm{p}}, a^{\mathrm{p}}} x^{\mathrm{p}}(s^{\mathrm{p}}, a^{\mathrm{p}}) P^{\mathrm{p}}(s^{\mathrm{p}\prime} | s^{\mathrm{p}}, a^{\mathrm{p}}) + \delta_{s^{\mathrm{p}\prime}, s_0^{\mathrm{p}}}; \tag{2c}$$

$$\textstyle\sum_{s^{\mathrm{p}}_{T_a}:\, u^+ \in s^{\mathrm{p}}_{T_a}} \sum_{a^{\mathrm{p}}} x^{\mathrm{p}}(s^{\mathrm{p}}_{T_a}, a^{\mathrm{p}}) \geq p_a \qquad \text{for achievement } c_a = (T_a, p_a) \tag{2d}$$

$$\text{or } \textstyle\sum_{s^{\mathrm{p}}_{T_m}:\, u^+ \in s^{\mathrm{p}}_{T_m}} \sum_{a^{\mathrm{p}}} x^{\mathrm{p}}(s^{\mathrm{p}}_{T_a}, a^{\mathrm{p}}) \geq p_m \qquad \text{for maintenance } c_m = (T_m, p_m) \tag{2e}$$

2.1.2. Maintenance commitments

As a reminder, a maintenance commitment is appropriate in scenarios where the initial value of state feature u is desirable to the recipient, who wants it to maintain its initial value for some interval of time (e.g., [40,46]), but where the provider might want to take actions that could change it. Formally, a maintenance commitment is defined by tuple $c_m = (T_m, p_m)$, where T_m is the maintenance commitment time, and p_m is the maintenance commitment probability. Given such a maintenance commitment, the provider is constrained to follow a policy that keeps u unchanged for the first T_m time steps with *at least* probability p_m. Since u can be toggled at most once, this is equivalent to probabilistically guaranteeing that u is still u^+ at the commitment time T_m, i.e.

$$\Pr(u_{T_m} = u_0 | u_0 = u^+) \geq p_m. \tag{3}$$

As with an achievement commitment, the provider with a maintenance commitment finds a policy that optimizes its local value while respecting the commitment semantics, again by including the commitment constraint in its LP. Specifically, in the LP of Equation (2), one only needs to replace constraint (2d) with constraint (2e) that expresses the semantics in Equation (3) for the maintenance commitment. Hence, from the provider's perspective, achievement and maintenance commitments are treated essentially identically.

2.2. Exploiting abstraction in the commitment specification

As we have seen, the commitment specification and semantics constrain the provider's policy based on a single future time step: at that time step, the value of u will (still) be u^+ with at least the promised probability. By abstracting away the probabilities at intervening (and subsequent) time steps, the commitment specification allows the provider to retain flexibility to revise its policy on the fly as it learns more about its environment.

Our prior work has shown the value to the provider of retaining such flexibility [9,15,16], and because the rest of this article focuses on challenges the recipient faces due to the abstraction in the commitment specification, we here briefly illustrate its benefits to the provider using the classic partially-observable MDP problem known as RockSample [47]. In a RockSample(n, s) problem instance, a rover agent is tasked to explore an unknown environment of an $n \times n$ grid containing s rocks, as shown in Fig. 1 for $(n = 2, s = 2)$ and $(n = 4, s = 4)$. Some rocks are of type `good`; the others are of type `bad`. The type of each rock has a uniformly random prior, and the rover can make noisy sensor observations detecting the rock type. The task is to determine which rocks are valuable, approach and take samples of valuable rocks, and exit the map as soon as possible. Our prior work [16] adapts the original problem such that the rover commits to exit the map by a predefined time T_a horizon with a certain probability p_a; that is, the rover can be viewed as an achievement commitment provider with u^+ denoting that it has exited.

If the rover knew exactly which rocks were valuable, it could at the outset formulate an optimal policy to follow from beginning to end, where it could not only meet its commitment but could also predict if and when it might in fact exit the map earlier than the committed time. However, its uncertainty about which rocks are valuable prevents this. Formulating a comprehensive policy that accounts for all possible combinations of rock valuations and sensor observations is intractable in all but the simplest settings, so instead the rover should iteratively modify its policy over the course of execution based on what it has learned about its environment so far [16]. Since the rover has provided a probabilistic commitment for a time by which it will exit, its modifications must always respect the commitment, but even so the rover can flexibly re-optimize its trajectory, including whether and when to exit earlier than the commitment time, based on its evolving knowledge of its environment.

A consequence of such flexibility is that a commitment provider (in this case, the rover) in general might not know, at the time when it and the recipient agree on the commitment, what exact policy it will ultimately follow, and thus what true influence P_u it will exert on the recipient's environment. The *only* information both agents know with certainty about

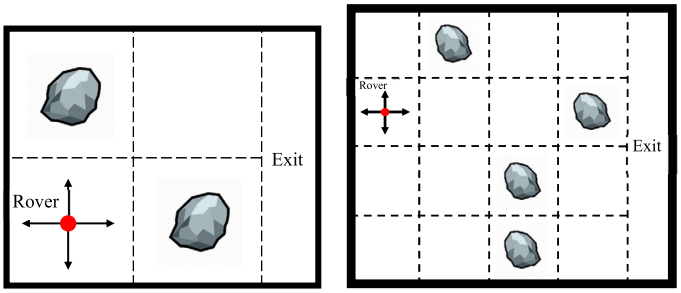


Fig. 1. *Left:* RockSample(2,2). *Right:* RockSample(4,4).

P_u at the outset is the commitment specification, and that the probability will change monotonically (due to u toggling at most once). Therefore, to plan its policy, the recipient must interpret the abstract commitment by fleshing out the rest of the influence in some way, into an approximation of the true influence.

2.3. The approximate influence

We notate the approximate influence that the recipient uses for its planning as $\widehat{P}_u$. Because u toggles at most once, any (true or approximate) influence P_u is fully specified as a vector of probabilities for each time $h = 0, \ldots, H$. We thus denote the value at time step h as $P_u[h]$. Fig. 2 illustrates two influences for an achievement commitment in Fig. 2a and two influences for a maintenance commitment in Fig. 2b. The gray area in each of the figures represents the admissible time-probability combinations based on the commitment semantics (Equation (1) for achievement and Equation (3) for maintenance). Any monotonically non-decreasing (non-increasing) function, for achievement (maintenance), falling entirely within the gray area is an admissible (commitment-respecting) influence, where again two examples of the many possible such influences are drawn.

In principle, the recipient could approximate the provider's influence with any admissible influence, but we are specifically interested in how the recipient can adopt an approximate influence $\widehat{P}_u$ that, in expectation, maximizes the quality of its plan when evaluated in (true) influence P_u. Formally, given $\widehat{P}_u$, let $\widehat{M} = (\mathcal{S}, \mathcal{A}, \widehat{P}, R, H, s_0)$ be the approximate model that only differs from M in terms of the dynamics of u, i.e. $\widehat{P} = (P_l, \widehat{P}_u)$. The quality of $\widehat{P}_u$ is evaluated using the difference between the value of the optimal policy for $\widehat{M}$ and the value of the optimal policy for M when both policies are evaluated[2] in M starting in s_0, i.e.

$$\text{Suboptimality}(\widehat{P}_u; P_u) = V^*_M(s_0) - V^{\pi^*_{\widehat{M}}}_M(s_0) = v^*_M - v^{\pi^*_{\widehat{M}}}_M.$$

Because P_u is entirely determined (perhaps iteratively) by the provider, the recipient should limit its risk of miscoordination by adopting a strategy for formulating an approximate influence $\widehat{P}_u[\cdot]$ that robustly induces low suboptimality for *any* admissible P_u given the commitment. This problem of *risk-aware interpretation of the commitment* is the focus of the rest of this article.

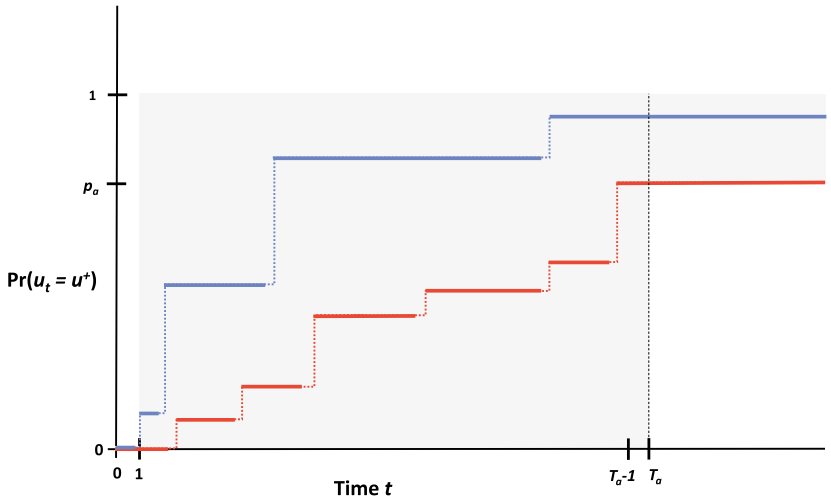
2.4. Summary

In this section, we have shown that, from the provider's perspective, achievement and maintenance commitments are treated essentially identically. Further, from the recipient's perspective, the notions of approximate influence and suboptimality also identically apply to the two types of commitment. Even though decision-theoretic formulations of, and reasoning methods for, achievement and maintenance commitments are nearly identical, prior work has found it much harder to successfully coordinate for maintenance than achievement [10–12]. In the past, it has been assumed that the difficulty lies on the provider's side—that it might be inherently harder for a provider to find good policies that maintain a feature than to change it. However, in the remainder of this article, we show that the challenge instead actually lies on the recipient's side: that a maintenance commitment is fundamentally harder for the recipient to interpret robustly to bound risk than an achievement commitment is. We now substantiate this claim theoretically in Section 3 and empirically in Section 4.

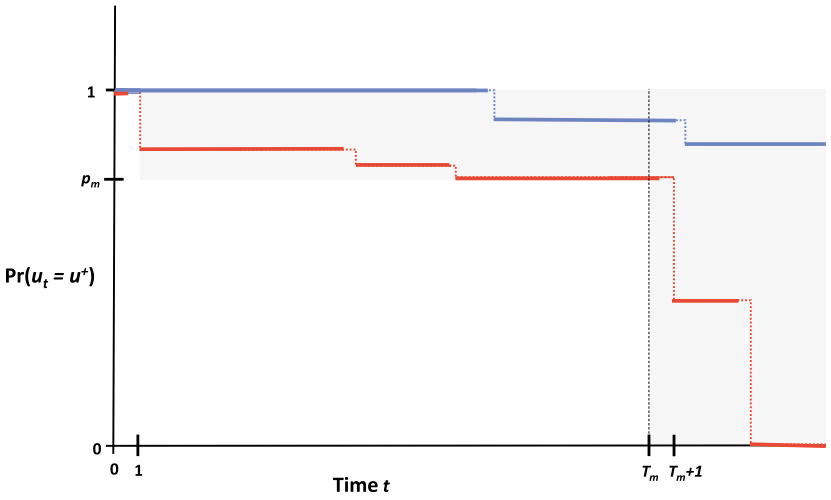
3. Bounds on interpretation suboptimality

In this section, we develop several interpretation strategies for the recipient to approximate the true influence, and present theoretical analyses that bound the worst-case suboptimality of some of these strategies, and prove no such bounds

[2] Note that when the support of P_u is not fully contained in the support of $\widehat{P}_u$, the recipient's policy $\pi^*_{\widehat{M}}$ can associate zero occupancy (hence plan no action) for certain states when executed in M, which makes $V^{\pi^*_{\widehat{M}}}_M$ ill-defined. In this work, we resolve this by re-planning: during the execution of $\pi^*_{\widehat{M}}$ in M, the recipient re-plans from any zero occupancy state that it happens to reach.



(a) Example Influences for Achievement



(b) Example Influences for Maintenance

Fig. 2. Example admissible true influences for an achievement commitment and a maintenance commitment. Because time is discrete, these are step functions. As indicated using the solid lines for the values of the functions, the probability takes the higher value at a "step" for the achievement commitment, and the lower value of a "step" for the maintenance commitment.

exist for the others. As stated earlier, to analyze each type of commitment separately, we need to ensure that only one commitment can hold between the agents. For that reason, as explained in Section 2.1, we assume that u can toggle at most once, since toggling more than once implies more than one commitment. Concretely, if $u = u^-$ initially and can toggle to u^+ and then back to $u-$, the recipient wants a commitment by the provider to achieve u^+ and also a commitment to maintaining u^+ long enough for the recipient to make use of it. Alternatively, if $u = u^+$ initially and can toggle to u^- and back to u^+, the recipient wants commitments both about maintaining u^+ initially and about re-achieving it in case the maintenance did not last long enough.

Assuming at most a single toggling, however, is not enough, because even so it is still possible that multiple commitments can be required if the recipient's preferred value of u can change. That is, if the recipient sometimes prefers $u = u^-$ to $u = u^+$, and (by definition) also sometimes prefers $u = u^+$ to $u = u^-$, then even with a single toggling it could want both a commitment that the initial value be maintained for some period of time, and also a commitment that the other value be achieved by some later time. To ensure that this cannot happen requires the following two assumptions, one about the

reward function (that reward for a reachable state with $u = u^-$ is never higher than for the same state with u^+), and one about the transition function (that the recipient will never prefer $u = u^-$ over $u = u^+$ because it causes some actions to reach preferable outcomes).

Assumption 1. For the recipient's reward function R, we assume

$$R(s,a) = R(s) = R((l,u)) \text{ and } R((l,u^+)) \geq R((l,u^-))$$

for all (s,a) and l.

Again, in words, this assumption ensures multiple preferences for the value of u cannot occur due to the reward function because there is no state the recipient can reach where its reward will be higher if $u = u^-$ than if $u = u^+$. This assumption is, for example, trivially met in domains where the reward function is designed to only consider the state features l that the recipient can actually control, as well as in domains where the reward function can be factored as $R((l,u)) = R(l) + R(u)$.

Assumption 2. Let $s^- = (l,u^-)$ and $s^+ = (l,u^+)$ be a pair of states that only differ in u. For any M with arbitrary influence P_u, there exists an optimal policy π^*_M such that

$$P_l\left(\cdot|s^-, \pi^*_M(s^-)\right) = P_l\left(\cdot|s^+, \pi^*_M(s^-)\right).$$

Again, in words, this assumption avoids different preferences for u due to the recipient's transition function by requiring that any action in its optimal policy assuming $u = u^-$ has identical dynamics even if $u = u^+$ instead. Note that if this assumption does not hold, then u^+ and u^- could induce different dynamics (make different states more or less likely to be reached), and thus the recipient could sometimes prefer that $u = u^-$ while executing its policy.

As a simple example of a domain where this assumption holds, consider an indoor robot whose only actions move it around its environment, and there is a door that can be open u^+ or closed u^-. If it believes that the door is closed, the robot's optimal policy will only include actions that move it between locations in whichever room it begins in. Note though that none of these actions are actually affected by the status of the door, and so the policy will behave identically even if the door is open.

Note further that the assumption can also hold even when $u = u^-$ is in fact a precondition to the success of some actions in the domain, as long as those actions are not in the optimal policy for the specific problem instance. Continuing the example, it could be that the door being closed improves the success for some actions (e.g., preventing the dog from sleeping on the bed), but the assumption still holds if such actions are not included in the policy over the finite time horizon H.

To derive bounds on achievement and maintenance commitments, we will make use of the following lemma, where M^+ (M^-) is defined as the recipient's MDP identical to M except that u is always set to u^+ (u^-). Lemma 1 directly follows from Assumption 2, stating that the value of M^- is no more than that of M^+ and the value of any M is between the two.

Lemma 1. *For any M with arbitrary influence P_u and initial value of u, we have $v^*_{M^-} \leq v^*_M \leq v^*_{M^+}$.*

Proof. Let's first consider the case in which P_u toggles u only at a single time step. We show $v^*_{M^-} \leq v^*_M$ by constructing a policy in M for which the value is at least $v^*_{M^-}$ by executing $\pi^*_{M^-}$. Whether u is initially u^- and later toggled to u^+ or *vice versa*, we can construct a policy π_M that chooses the same actions as $\pi^*_{M^-}$ assuming $u = u^-$ throughout the episode. Formally, for any $s^- = (l,u^-)$, letting $s^+ = (l,u^+)$,

$$\pi_M(s^+) = \pi_M(s^-) = \pi^*_{M^-}(s^-).$$

By Assumption 2, π_M in M yields the same distribution over the trajectory of l as $\pi^*_{M^-}$ in M^-, and therefore $v^{\pi_M}_M \geq v^*_{M^-}$ by Assumption 1.

Similarly, we show $v^*_M \leq v^*_{M^+}$ by constructing a policy π_{M^+} in M^+ for which the value is at least v^*_M by executing π^*_M. Formally, for time steps when $u = u^-$ in M, let $\pi_{M^+}(s^+) = \pi^*_M(s^-)$. For time steps when $u = u^+$ in M, let $\pi_{M^+}(s^+) = \pi^*_M(s^+)$, where $s^- = (l,u^-)$, $s^+ = (l,u^+)$.

When P_u is such that the single toggling of u could occur at any of $K > 1$ time steps, we can decompose the value function for P_u as the weighted average of K value functions corresponding to the K influences that toggle u at a single time step, and the weights of the average are the toggling probabilities of P_u at these K time steps. □

3.1. Minimal enablement duration

We begin by analyzing an intuitive and straightforward interpretation strategy that has been adopted in previous work [34,9] to create approximate influences for achievement commitments. The strategy models the influence with a single

transition, *at the commitment time*, where u^- probabilistically toggles to u^+. Approximating the influence as toggling to u^+ at the latest possible time ignores the possibilities of being enabled earlier than the deadline and of being enabled serendipitously after the deadline. We say that such an approximate influence interprets the achievement commitment *pessimistically*, in the sense that it tries to reduce the risk of prematurely expecting the condition to hold by minimizing the expected duration of u being enabled over all admissible influences (Equation (1)):

$$\min_{P_u \sim (1)} \mathbb{E}_{P_u}\left[\sum_{t=0}^{H} 1_{\{u_t = u^+\}}\right]$$

where $P_u \sim (1)$ means influence P_u satisfies Equation (1), and 1_E is the indicator function that takes value one if event E occurs and zero otherwise. We refer to this minimizing approximation as the *minimal enablement duration* influence, as formalized in Definition 1 and illustrated in Fig. 3a.

Definition 1. Given achievement commitment $c_a = (T_a, p_a)$, its minimal enablement duration influence $\widehat{P}_u^{\min+}(c_a)$ toggles u in the transition from time step $t = T_a - 1$ to $t = T_a$ with probability p_a, and does not toggle u at any other time step.

For maintenance commitments, the counterpart approximation minimizes the expected enablement duration over all influences that respect the maintenance commitment semantics (Equation (3)):

$$\min_{P_u \sim (3)} \mathbb{E}_{P_u}\left[\sum_{t=0}^{H} 1_{\{u_t = u^+\}}\right].$$

This minimizing approximation thus models a probabilistic toggling to u^- at the earliest possible time, and a deterministic toggling to u^- (if it had not toggled earlier) after the commitment time, as formalized in Definition 2 and illustrated in Fig. 3b.

Definition 2. Given maintenance commitment $c_m = (T_m, p_m)$, its minimal enablement duration influence $\widehat{P}_u^{\min+}(c_m)$ toggles u in the transition from time step $t = 0$ to $t = 1$ with probability $1 - p_m$, and (unless already toggled) from $t = T_m$ to $t = T_m + 1$ with probability one. It does not toggle u at any other time step.

3.2. Bounding suboptimality for achievement commitments

Suboptimality in the recipient's performance given the provider's actual behavior with respect to a commitment can arise from two sources: (1) when the recipient poorly approximates the provider's true influence on u as the provider pursues the promised commitment; and/or (2) when the provider actually pursues a better commitment than what was promised, in which case the recipient likely acts suboptimally to the better commitment because it is unaware of it. For our analyses of interpretation strategies for how the recipient chooses to approximate commitment influences, we focus only on suboptimality due to the first reason. To ensure that suboptimality due to the provider exceeding what was promised in the commitment cannot arise, we introduce Assumption 3.

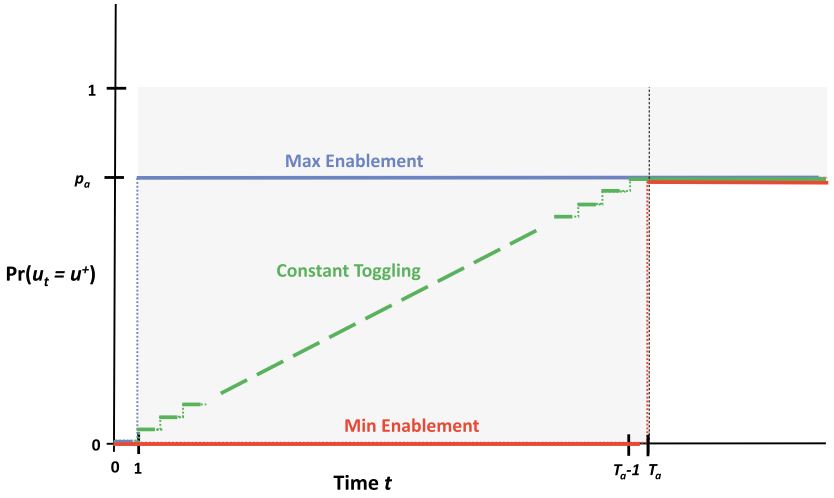
Assumption 3. $P_u[u_{T_a}] = p_a$ and $P_u[u_{T_m}] = p_m$ for probabilistic achievement and maintenance commitments, respectively. Further, $u_h = u_{T_a}$ for $h \geq T_a$ and $u_h = u^-$ for $h > T_m$ for probabilistic achievement and maintenance commitments, respectively.

In words, the assumption is that the provider's actual influence does not exceed the probabilities for achievement or maintenance promised in the commitment, and that it will not achieve (toggle from u^- to u^+) or maintain (prevent toggling from u^+ to u^-) later than the time promised in the commitment. With this assumption, note that P_u will behave identically to the minimal-enablement-duration approximate influence (and, per Fig. 3, the other approximate influences we will consider) after the commitment time, meaning that suboptimality only arises due to the approximation up to the commitment time.

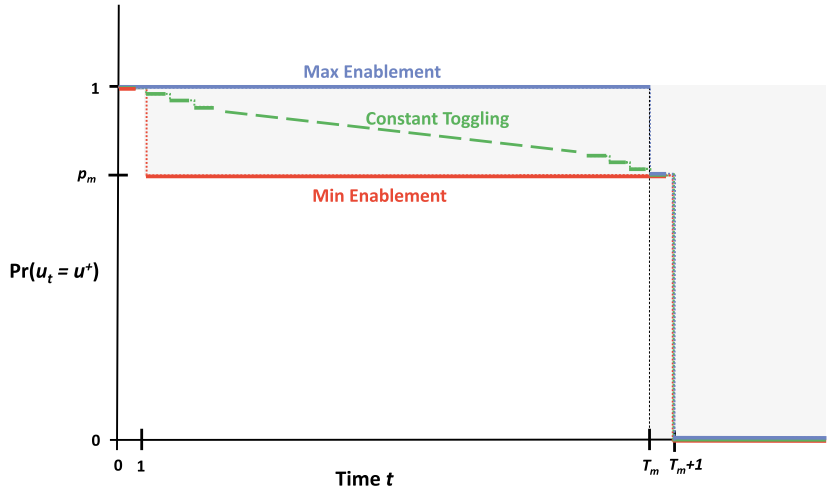
We should emphasize that Assumption 3 actually typically holds in domains where commitments are useful as a coordination strategy. Commitments are useful if, without a commitment, the provider would not otherwise choose to incur the costs of achieving/maintaining u^+. Given these costs, therefore, a rational provider will never choose to pursue u^+ harder (seek a higher probability of success) or longer than promised. Moreover, if the provider's MDP is such that, to meet the commitment, it cannot help but attain a probability higher than p_a (p_m) or achieve u^+ later (maintain u^+ longer) than T_a (T_m), then the agents would have adopted the better commitment in the first place. For example, elsewhere we have proven that the recipient's value is monotonically non-decreasing as a function of the achievement probability [14].

With Assumption 3 in place, we can now prove Lemma 2 which states that, for achievement commitments, the possible ways the true influence differs from the minimal enablement duration approximation can only improve the expected value.

Lemma 2. *Given achievement commitment $c_a = (T_a, p_a)$, let $\widehat{P}_u = \widehat{P}_u^{\min+}(c_a)$, then we have $v_M^{\pi^*_{\widehat{M}}} \geq v_{\widehat{M}}^{\pi^*_{\widehat{M}}}$ where influence P_u in M respects the commitment semantics of c_a.*



(a) $\widehat{P}_u^{\min+}[\cdot]$ for Achievement is the bottom function.



(b) $\widehat{P}_u^{\min+}[\cdot]$ for Maintenance is the lowest step function.

Fig. 3. Minimal enablement duration influences are depicted for an achievement and a maintenance commitment, along with other heuristic strategies described in Section 3.4.

Proof. For achievement commitments, the initial value of u is u^-. Let $P_u[t]$ be the probability that u is not enabled to u^+ until time step t in influence P_u, and $\overline{v}_t^{\pi}$ be the initial state's value under π when u is enabled from u^- to u^+ at t with probability one. By Assumption 3, $v_M^{\pi^*_{\widehat{M}}}$ and $v_{\widehat{M}}^{\pi^*_{\widehat{M}}}$ can be decomposed as

$$v_M^{\pi^*_{\widehat{M}}} = \sum_{t=1}^{T_a} P_u[t]\overline{v}_t^{\pi^*_{\widehat{M}}} + (1-p_a)v_{M^-}^{\pi^*_{\widehat{M}}},$$
$$v_{\widehat{M}}^{\pi^*_{\widehat{M}}} = p_a\overline{v}_{T_a}^{\pi^*_{\widehat{M}}} + (1-p_a)v_{M^-}^{\pi^*_{\widehat{M}}}.$$

When u is enabled at t in M, $\pi^*_{\widehat{M}}$ can be executed as if u is not enabled, by Assumption 2, yielding identical trajectory distribution of l and therefore no less value by Assumption 1 as in $\widehat{M}$. Further, the recipient's re-planning at t when $u = u^+$ will derive a better policy if possible. Therefore, the value of executing $\pi^*_{\widehat{M}}$ in M is no less than that in $\widehat{M}$, i.e. $\overline{v}_t^{\pi^*_{\widehat{M}}} \geq \overline{v}_{T_a}^{\pi^*_{\widehat{M}}}$. Therefore,

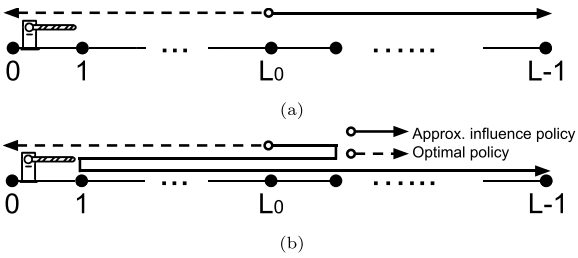


Fig. 4. 1D Walk. *Top:* Example in the proof of Theorem 1. *Bottom:* Example in the proof of Theorem 2.

$$\begin{aligned} v_M^{\pi^*_{\widehat{M}}} &= \textstyle\sum_{t=1}^{T_a} P_u[t]\overline{v}_t^{\pi^*_{\widehat{M}}} + (1-p_a)v_{M^-}^{\pi^*_{\widehat{M}}} \\ &\geq \textstyle\sum_{t=1}^{T_a} P_u[t]\overline{v}_{T_a}^{\pi^*_{\widehat{M}}} + (1-p_a)v_{M^-}^{\pi^*_{\widehat{M}}} \\ &\geq p_a\overline{v}_{T_a}^{\pi^*_{\widehat{M}}} + (1-p_a)v_{M^-}^{\pi^*_{\widehat{M}}} \quad \text{(commitment semantics)} \\ &= v_{\widehat{M}}^{\pi^*_{\widehat{M}}}. \quad \square \end{aligned}$$

Finally, with Assumptions 2 and 3, and Lemma 2, we derive Theorem 1 that bounds the suboptimality of the approximate influence based on the minimal enablement duration strategy, when applied to achievement commitments, as the difference between $v^*_{M^-}$ and $v^*_{M^+}$.

Theorem 1. *Given achievement commitment c_a, let $\widehat{P}_u = \widehat{P}_u^{\min+}(c_a)$. The suboptimality can be bounded as*

$$v^*_M - v_M^{\pi^*_{\widehat{M}}} \leq v^*_{M^+} - v^*_{M^-} \tag{4}$$

where influence P_u in M respects the commitment semantics of c_a. Further, there exists an achievement commitment for which the equality is attained.

Proof. The derivation of the bound in Equation (4) is straightforward from Lemma 2:

$$v^*_M - v_M^{\pi^*_{\widehat{M}}} \leq v^*_{M^+} - v_{\widehat{M}}^{\pi^*_{\widehat{M}}} \leq v^*_{M^+} - v^*_{M^-}.$$

Next, to prove that the bound is tight, we use a simple illustrative example as an existence proof of an achievement commitment for which the equality is attained.

Example: An Achievement Commitment in 1D Walk. Consider the example of a 1D walk of L locations on $[0, L-1]$, as shown in Fig. 4(top), where the recipient starts at L_0 and can move right, left, or stay still. There is a gate between 0 and 1 for which u^+ denotes the state of open and u^- closed. The provider toggles the gate stochastically according to P_u. For each time step the recipient is at neither end, it gets a reward of -1. Hence, the optimal policy is to reach either end as soon as possible in expectation. Note that the reward function makes Assumptions 1 and 2 hold.

Here, we derive an achievement commitment for which the bound in Theorem 1 is attained. Consider $L = 10, L_0 = 3, H = 10$, achievement commitment $(T_a = L-1-L_0 = 6, p_a = 1)$, and the true influence P_u in M that toggles the gate to open at $t = L_0 - 1 = 2$ with probability $p_a = 1$. The optimal policy in M is to move left to 0. Therefore, $v^*_M = v^*_{M^+} = -L_0 = -3$. Given the minimal enablement duration influence, moving right to L (arriving at time $L-1-L_0 = 6$) is faster than waiting for the gate to toggle at $T_a = 6$ and then reaching location 0 at time $T_a + 1 = 7$. Had the recipient known the gate would toggle at time $t = L_0 - 1 = 2$, it would have moved left, but by the time the gate toggles the recipient is at location $L_0 + L_0 - 1 = 5$, and continuing on to L is the faster choice. Therefore, $v_M^{\pi^*_{\widehat{M}}} = v^*_{M^-} = -(L-1-L_0) = -6$, and the bound in Theorem 1 is attained. $\square$

3.3. Failing to bound suboptimality for maintenance commitments

We next ask if the bound in Equation (4) on suboptimality in achievement commitments also holds for maintenance commitments. Unfortunately, as stated in Theorem 2, the optimal policy of the minimal enablement duration influence for maintenance commitments can be arbitrarily bad when evaluated in the true influence, incurring a suboptimality exceeding the bound in Equation (4). We give an example for an existence proof.

Theorem 2. *Consider $\widehat{P}_u = \widehat{P}_u^{\min+}(c_m)$ to be the approximate influence when modeling the maintenance commitment in $\widehat{M}$. There exists an MDP M and a maintenance commitment c_m, such that the true influence P_u in M respects the commitment semantics of c_m, $v^*_M = v^*_{M^+}$, $v_M^{\pi^*_{\widehat{M}}} < v^*_{M^-}$, and therefore the suboptimality*

$$v^*_M - v^{\pi^*_{\widehat{M}}}_M > v^*_{M^+} - v^*_{M^-} \tag{5}$$

exceeds the bound in Equation (4).

Proof. As an existence proof, we give an example of a maintenance commitment in 1D Walk for which $v^*_M = v^*_{M^+}$ and $v^{\pi^*_{\widehat{M}}}_M < v^*_{M^-}$. Consider 1D Walk with the same $L = 10, L_0 = 3, H = 10$ as in the example for Theorem 1. Consider maintenance commitment $(T_m = L_0 + 1 = 4, p_m = 0^+)$ where we use $p_m = 0^+$ to denote an infinitesimally small positive probability (i.e., very close to 0 yet strictly larger than 0), and P_u toggles the gate to closed at $T_m = 4$ with probability $1 - p_m$. As shown in Fig. 4(bottom), the optimal policy should take L_0 steps to move directly to 0, and since probability $1 - p_m$ approaches one, for the policy's value approaches $v^*_M = v^*_{M^+}$. Meanwhile, the optimal policy assuming the gate is closed is to simply move right, so $v^*_{M^-} = -(L - 1 - L_0) = 6$ (which is the same as we computed for Theorem 1 since L and L_0 are the same in both examples).

The optimal policy with the minimal enablement duration approximation is as follows. Because it models that with near certainty the gate will close at time step 1, the first move is to the right. Then, it observes that the gate did not close. With $p_m = 0^+$, the policy includes a branch for this unlikely but possible situation, and the approximate influence models the dynamics in this case to be that the gate is expected to be open until $T_m + 1 = 5$, at which point it will be shut with certainty. Since the agent can just barely get through the gate by then, the policy directs it to move left. However, with the true influence the gate is almost surely closed at $T_m = 4$, just when the agent would have moved through it. At that point, re-planning occurs (since the policy did not model the possibility of the gate being closed at time 4), and the optimal policy is to move right. Thus, $v^{\pi^*_{\widehat{M}}}_M \cong -H = -10 < v^*_{M^-}$, where the approximate equality ($\cong$) accounts for the extremely unlikely case of the gate remaining untoggled with probability $p_m = 0^+$. □

In the example used in the preceding existence proof, the maximum suboptimality is incurred with maintenance commitment probability $p_m = 0$ (a no-guarantee commitment), because this is when the recipient is most uncertain about the influence and will be most negatively affected by the uncertainty. Note that for achievement, a no-guarantee commitment still falls within the Theorem 1 bound.

Comparing the bound Equation (4) in Theorem 1 with the bound Equation (5) in Theorem 2 reveals a fundamental difference between achievement and maintenance commitments: maintenance commitments are inherently less tolerant to an unexpected change in the commitment feature. For achievement commitments, the easily-constructed minimal enablement duration influence has the property of being pessimistic, in that any unexpected changes to the feature, if they impact the recipient at all, *can only improve the expected value*. Thus, if despite its minimal enablement duration influence approximation, a recipient has chosen to follow a policy that exploits the commitment, it never risks experiencing a true admissible influence that would lead it to regret having done so. *The same cannot be said for maintenance commitments.* There, the easily-constructed minimal enablement duration influence is *not* pessimistic—it does not guarantee that any deviations from the influence can only improve the expected value. As our theoretical results show, the minimal enablement duration influence assuming toggling from u^+ to u^- right away still risks negative surprises, since if the toggling does not immediately occur the influence suggests that it is safe to assume no toggling until T_m, but that is not true since toggling could happen sooner, after the recipient has incurred cost for a policy that would need to be abandoned. In the example for Theorem 2, the worst time for toggling to u^- is not right away, but right before the precondition would be used, where the gate shuts just as the recipient is about to pass through it, and now the recipient needs to go all the way to the other end.

3.4. Alternative influence approximations

There are of course many other candidate approximations that are just as legitimate as the minimal enablement duration strategy, in terms of obeying the commitment semantics. Visually, this equates to staying within the "gray" region depicted in Fig. 3 and being monotonically non-decreasing (non-increasing) for achievement (maintenance) commitments. Two obvious alternative candidate approximations, also shown in that figure, are described next.

Maximal enablement duration As opposed to the minimal enablement duration strategy, the maximal enablement duration strategy optimistically toggles u right after the initial time step for achievement commitments, and at the commitment time for maintenance commitments, as shown in Fig. 3. Formally, given achievement commitment $c_a = (T_a, p_a)$, the maximal enable duration strategy, denoted as $\widehat{P}^{\max+}_u[\cdot]$, chooses the influence $\widehat{P}^{\max+}_u(c_a)$ that toggles u in the transition from time step $t = 0$ to $t = 1$ with probability p_a, and does not toggle u at any other time step; given maintenance commitment $c_m = (T_m, p_m)$, the maximal enablement duration strategy chooses the influence $\widehat{P}^{\max+}_u(c_m)$ that toggles u in the transition from time step $t = T_m - 1$ to $t = T_m$ with probability $1 - p_m$, and (unless already toggled) from $t = T_m$ to $t = T_m + 1$ with probability one. It does not toggle u at any other time step.

Constant toggling The constant toggling strategy, denoted as $\widehat{P}^{\text{const}}_u[\cdot]$, chooses the influence $\widehat{P}^{\text{const}}_u(c)$, for either an achievement or a maintenance commitment c, that linearly interpolates between the probability of u^+ at time $t = 0$ and the commitment time. That is, it toggles u at every time step up to the commitment time with a constant probability, and

the probability is chosen such that the overall probability of toggling by the commitment time matches the commitment probability. The influence $\widehat{P}_u^{\text{const}}(c)$ agrees with the minimal enablement duration influence after the commitment time. This strategy is also depicted in Fig. 3.

While the preceding alternative approximation strategies make sense as complements to the minimal enablement duration strategy, they are not formulated so as to directly address the shortcomings of the minimal enablement duration strategy for maintenance commitments illustrated in the proof of Theorem 2. As shown in that proof, sometimes the worst time for toggling does not correspond to either the minimal or maximal enablement duration. Instead, it is when the recipient has invested effort in a plan that will utilize the u^+ precondition, and is just about to use that precondition when the condition toggles (e.g., the gate closes just before it is to be passed through).

The following strategies therefore try to be even more pessimistic by trying to find the worst possible toggling time, and build the approximate influence accordingly. That is, like the minimal and maximal enablement duration strategies, they model toggling at a single time step and agree with Assumption 3 thereafter. We denote the set of such influences as $\mathcal{P}_u^1(c)$ for either an achievement or a maintenance commitment c.

Minimal value timing The approximate influence based on the minimal value timing strategy, denoted as $\widehat{P}_u^{\text{minV}}[\cdot]$, chooses the influence from $\mathcal{P}_u^1(c)$ that has the minimal optimal value. Formally, for either an achievement or a maintenance commitment c, its minimal value timing influence $\widehat{P}_u^{\text{minV}}(c)$ is $\arg\min_{\widehat{P}_u \in \mathcal{P}_u^1(c)} v^*_{\widehat{M}}$ where $\widehat{P}_u$ is the influence in $\widehat{M}$. Thus, to find $\widehat{P}_u^{\text{minV}}(c)$, the recipient builds approximate influences that toggle for every time up to the commitment time, computes the value of its optimal policy for each of these candidate approximate influences, and adopts the approximate influence with the worst value. (Note that the first and last of these approximations correspond to the minimal and maximal enablement duration approximations, where which is minimal and which is maximal depends on the commitment type.)

Minimax regret timing The minimax regret timing strategy $\widehat{P}_u^{\text{minimax}}[\cdot]$ chooses an influence from $\mathcal{P}_u^1(c)$ based on the minimax regret principle. Formally, for either an achievement or a maintenance commitment c, its minimax regret timing influence $\widehat{P}_u^{\text{minimax}}(c)$ is

$$\arg\min_{\widehat{P}_u \in \mathcal{P}_u^1(c)} \max_{P_u \in \mathcal{P}_u^1(c)} v^*_M - v^{\pi^*_{\widehat{M}}}_M$$

where $P_u, \widehat{P}_u$ are the influences in $M, \widehat{M}$, respectively. The straightforward algorithmic realization of this strategy is to once again build an approximation for each time up to the commitment time and compute an optimal policy based on the approximation, but then to evaluate that policy against the toggling happening at all of the possible times (again up to the commitment time), and choosing the approximation that minimizes the maximum regret (how badly the approximation's optimal policy performs compared to how well the policy optimal for that toggling time can perform, in the worst case).

We have therefore described a total of five influence approximation strategies: minimal enablement duration, maximal enablement duration, constant toggling, minimal value timing, and minimax regret timing. Recall that our overarching objective is to find one (or more) strategies that the recipient can use to bound the degree of suboptimality that it risks due to uncertainty about the true influence. A secondary objective is that the strategy be computationally inexpensive, in computing the approximation and/or in computing a policy based on the approximation. Of the five strategies, the maximal enablement duration strategy, while computationally inexpensive to compute and use, seems intuitively unlikely to help the recipient avoid suboptimal decisions, since it would often lead to the recipient being overoptimistic in the provider's behavior. On the other hand, the minimal duration strategy as initially motivated is pessimistic, as well as computationally inexpensive. The minimal value timing and the minimax regret timing strategies are considerably more computationally expensive, but because they explicitly search for the worst case based on their respective criteria, the hope would be that they support the overarching objective of bounding suboptimality, especially in the case of maintenance commitments where the simple minimal enablement duration strategy fell short. Finally, the constant toggling strategy occupies a middle ground computationally: it is inexpensive to compute the approximation, but the recipient must compute substantially more trajectories than for the other strategies where toggling only happens once before the commitment time. In its favor is that, by modeling possible toggling at each possible time, it can avoid the problem that arose for the minimal enablement duration strategy for maintenance commitments (seen in the proof of Theorem 2) where, if the toggling did not occur right at the outset, the recipient did not model toggling as even being possible before the commitment time.

Unfortunately, despite the intuitions above, Theorem 3 proves that, while the minimal value timing influence coincides with the minimal enablement duration for achievement and thus enjoys the same suboptimality bound as in Equation (4), *the bound does not hold for any of the other alternative strategies in either achievement or maintenance.*

Theorem 3. *For an achievement commitment c_a, the minimal value timing influence coincides with the minimal enablement duration, i.e. $\widehat{P}_u^{\text{minV}}(c_a) = \widehat{P}_u^{\text{min}+}(c_a)$, and thus the bound in Equation (4) holds for $\widehat{P}_u^{\text{minV}}(c_a)$. Except for this, the bound does not hold, i.e. for $\widehat{P}_u \in \{ \widehat{P}_u^{\text{max}+}(c_a), \widehat{P}_u^{\text{const}}(c_a), \widehat{P}_u^{\text{minimax}}(c_a), \widehat{P}_u^{\text{max}+}(c_m), \widehat{P}_u^{\text{const}}(c_m), \widehat{P}_u^{\text{minV}}(c_m), \widehat{P}_u^{\text{minimax}}(c_m)\}$, there exists an MDP M, and an achievement or maintenance commitment, such that the true influence P_u in M respects the commitment semantics of $c \in \{c_m, c_a\}$, and the suboptimality*

Table 1
1D Walk examples for Theorem 3. We use $p_m = 0^+$ to denote that $p_m > 0$ is an infinitesimally small positive probability.

	Achievement	Maintenance
Min Enablement	The bound in Eq. (4) holds	$L = 10, L_0 = 3, r_{\text{left}} = 0$ $T_m = 4, p_m = 0^+$ $v^*_{M^+} - v^*_{M^-} = -3 - (-6) = 3$ $P_u \in \mathcal{P}^1_{u,c}$ toggles at t =3 Suboptimality = 8.8
Max Enablement	$L = 10, L_0 = 6, r_{\text{left}} = 7$ $T_a = 4, p_a = 0.9$ $v^*_{M^+} - v^*_{M^-} = 1 - (-3) = 4$ $P_u \in \mathcal{P}^1_{u,c}$ toggles at t =3 Suboptimality = 4.7	$L = 10, L_0 = 3, r_{\text{left}} = 0$ $T_m = 3, p_m = 0^+$ $v^*_{M^+} - v^*_{M^-} = -3 - (-6) = 3$ $P_u \in \mathcal{P}^1_{u,c}$ toggles at t =1 Suboptimality = $4 - 0^+$
Constant Toggling	$L = 10, L_0 = 3, r_{\text{left}} = 0$ $T_a = 7, p_a = 0.9$ $v^*_{M^+} - v^*_{M^-} = -3 - (-6) = 3$ $P_u \in \mathcal{P}^1_{u,c}$ toggles at t =6 Suboptimality = 4.0	$L = 10, L_0 = 3, r_{\text{left}} = 0$ $T_m = 7, p_m = 0.1$ $v^*_{M^+} - v^*_{M^-} = -3 - (-6) = 3$ $P_u \in \mathcal{P}^1_{u,c}$ toggles at t =1 Suboptimality = 3.3
Min Value	The bound in Eq. (4) holds	$L = 10, L_0 = 3, r_{\text{left}} = 9$ $T_m = 7, p_m = 0.3$ $v^*_{M^+} - v^*_{M^-} = 6 - (-6) = 12$ $P_u \in \mathcal{P}^1_{u,c}$ toggles at t =5 Suboptimality = 14.9
Minimax Regret	$L = 10, L_0 = 6, r_{\text{left}} = 7$ $T_a = 5, p_a = 1.0$ $v^*_{M^+} - v^*_{M^-} = 1 - (-3) = 4$ $P_u \in \mathcal{P}^1_{u,c}$ toggles at t =4 Suboptimality = 5.8	$L = 10, L_0 = 3, r_{\text{left}} = 0$ $T_m = 4, p_m = 0^+$ $v^*_{M^+} - v^*_{M^-} = -3 - (-6) = 3$ $P_u \in \mathcal{P}^1_{u,c}$ toggles at t =3 Suboptimality = 8.8

$$v^*_M - v^{\pi^*_{\widehat{M}}}_M > v^*_{M^+} - v^*_{M^-}$$

exceeds the bound in Equation (4)*.*

Proof. We first show that the minimal value timing influence coincides with the minimal enablement duration for achievement commitments, i.e. $\widehat{P}^{\text{minV}}_u(c_a) = \widehat{P}^{\text{min}+}_u(c_a)$. Consider achievement commitment $c_a = (T_a, p_a)$, and $\widehat{P}_u, \widehat{P}'_u \in \mathcal{P}^1_u(c_a)$ that toggles u at T and T' respectively with $T' < T \leq T_a$. We can construct a recipient's policy for the earlier toggling $\widehat{P}'_u$ that mimics the optimal policy for $\widehat{P}_u$, and hence the optimal value for T' is at least that for T, i.e. $v^*_{\widehat{M}} \leq v^*_{\widehat{M}'}$ where $\widehat{P}_u$ and $\widehat{P}'_u$ are the influences in $\widehat{M}$ and $\widehat{M}'$, respectively. Specifically, let $\pi^*_{\widehat{M}}$ be the optimal policy for $\widehat{P}_u$ and $\pi^*_{\widehat{M}}(\cdot|s)$ be the action probability distribution of $\pi^*_{\widehat{M}}$ in state s. For the earlier toggling time $T' < T$, we construct a policy $\pi_{T'}$ that mimics $\pi^*_{\widehat{M}}$: it chooses actions as if $u = u^-$ until T. Formally, for time steps $t < T$, $\pi_{T'}(\cdot|s^-) = \pi^*_{\widehat{M}}(\cdot|s^-)$ for any state $s^- = (l, u^-)$ in which $u = u^-$, and $\pi_{T'}(\cdot|s^+) = \pi^*_{\widehat{M}}(\cdot|s^-)$ where $s^+ = (l, u^+)$ and $s^- = (l, u^-)$ only differ in u; for time steps $t \geq T$, $\pi_{T'}(\cdot|s^r) = \pi^*_{\widehat{M}}(\cdot|s^r)$. Because $\pi_{T'}$ and $\pi^*_{\widehat{M}}$ yield the same trajectories of l and the reward only depends on l, they achieve the same value, and therefore $v^*_{\widehat{M}} \leq v^*_{\widehat{M}'}$. Because the values are monotonically non-increasing with time, the minimal value timing strategy chooses either the same (last admissible) toggling time as the minimal enablement duration strategy, or an earlier time than that that gives the recipient no relative benefit.

We now prove that the bound does not hold for the other strategies. As an existence proof, Table 1 summarizes examples for which the bound in Equation (4) does not hold for $\widehat{P}_u \in \{\ \widehat{P}^{\text{max}+}_u(c_a),\ \widehat{P}^{\text{const}}_u(c_a),\ \widehat{P}^{\text{minimax}}_u(c_a),\ \widehat{P}^{\text{max}+}_u(c_m),\ \widehat{P}^{\text{const}}_u(c_m),\ \widehat{P}^{\text{minV}}_u(c_m),\ \widehat{P}^{\text{minimax}}_u(c_m)\ \}$. All the examples are in the 1D Walk domain with fixed $L = 10, H = 20$. Besides the -1 reward for every time step until reaching either end, the recipient also gets a one-time reward r_{left} when reaching the left end, which is an integer chosen from interval $[0, 10]$. We compute the suboptimality for a commitment c, achievement or maintenance, with initial location L_0 chosen from $\{1, 2, 3, ..., 8\}$, commitment time chosen from $\{1, 2, ..., H\}$ and commitment probability chosen from $\{0^+, 0.1, 0.2, ..., 1\}$, with the provider's true influence P_u chosen from $\mathcal{P}^1_u(c)$. For all possible combinations of r_{left}, c, L_0, and P_u, the corresponding suboptimality is evaluated, and Table 1 reports combinations for which the bound in Equation (4) does not hold. □

The analysis used for the proof above purposely chooses the provider's true influence from $\mathcal{P}^1_u(c)$ in an adversarial manner (from the recipient's perspective). One might thus question whether such influences contrived for the proof could actually realistically arise. We answer this question by asserting that a rational provider that maximizes its value can indeed induce such an influence in $\mathcal{P}^1_u(c)$ for any given commitment c, as formally stated in Theorem 4.

Theorem 4. *For any commitment c, achievement or maintenance, and any influence $P_u \in \mathcal{P}_u^1(c)$, there exists an MDP for the provider such that the optimal policy induces influence P_u.*

Proof. For achievement commitment $c = c_a = (T_a, p_a)$ and influence $P_u \in \mathcal{P}_u^1(c_a)$ that toggles at time step $T \leq T_a$, consider 1D Walk of $T_a + 1$ locations on $[0, T_a]$ as the provider's MDP, where the provider starts at location 0. The provider gets a reward of +1 for each time step at location T_a, and a reward of 0 everywhere else. For each time step at location T, the provider toggles the value of u from u^- to u^+ with probability p_a. Obviously, the provider's optimal policy is to move to and then stay at location T_a, which induces influence P_u.

Similarly, for maintenance commitment $c = c_m = (T_m, p_m)$ and influence $P_u \in \mathcal{P}_u^1(c_m)$ that toggles at time step $T \leq T_m$, consider the same 1D Walk of $T_m + 1$ locations as the provider's MDP, except that the provider toggles the value of u from u^+ to u^- with probability $1 - p_m$ at location T. The provider's optimal policy remains the same, which induces influence P_u. □

3.5. Summary

As a brief summary, in this section we have developed several interpretation strategies for the recipient to create the approximate influence, and theoretically analyzed their worst-case suboptimalities for both achievement and maintenance commitments. Our theoretical results show that there exists a strategy, minimal enablement duration, such that its worst-case suboptimality is reasonably bounded for achievement commitments. However, such a guarantee does not hold for maintenance commitments for *any* of the strategies we have considered. This not only includes the counterpart minimal enablement duration strategy but also the strategies that are purposely developed using insights about worst-case timing of the toggling, as well as the constant toggling strategy that models the toggling at every time step. While we cannot assert with certainty that a bounded strategy does not exist for maintenance commitments, we have shown that strategies specifically developed to account for the shortcomings of others nonetheless can still induce the worst-case unbounded suboptimality, and for this reason believe that no such bounded strategy exists.

4. Empirical study

In Section 3, we motivated and developed several strategies for the recipient to use to create its approximate influence for a given (achievement or maintenance) commitment, and analyzed their worst-case suboptimalities. Specifically, we formulated MDPs for the recipient in the 1D Walk domain, commitments between the agents, and true influences for the provider that respect the commitment semantics, that together maximize the suboptimality induced by the approximate influences. We showed that, for achievement, the worst-case suboptimality of the minimal enablement duration influence (or equivalently the minimal value timing influence) strategy can be bounded fairly tightly, while for maintenance the worst-case suboptimality of every approximate influence strategy that we could devise is effectively unbounded.

In this section, we conduct empirical evaluations of the suboptimality induced by the approximate influence strategies in settings that were not specifically designed to realize the worst case. In Section 4.1, we measure suboptimality for general (achievement or maintenance) commitments in the 1D Walk domain, sweeping through the possible combinations of commitment times and probabilities. In Section 4.2, we narrow the investigation specifically to combinations that agents would be most likely to adopt because they either maximize the provider's or the recipient's local commitment value, or maximize the joint commitment value.

Our experiments are designed to supplement the theoretical analyses by measuring the suboptimality of the alternative influence approximations. To our knowledge, there really are no other approaches to approximating a commitment's influence, and therefore no benchmarks for us to compare with. Note that, although there are indeed other formulations/methods for coordinating dependent agents that could be applied to the problems studied in the empirical section, this work is not about comparing the effectiveness of commitment-based coordination to other methods, but rather it is asking how to model commitments well assuming that a commitment-based coordination approach is being employed.

4.1. Suboptimality for general commitments

Here, we measure the suboptimality of the interpretation strategies developed in Section 3 for a general achievement commitment $c_a = (T_a, p_a)$ or maintenance commitment $c_m = (T_m, p_m)$ in the 1D Walk domain, where the commitment time $T_a, T_m \in \{1, 2, ..., H\}$ can be any time step up to the horizon and the commitment probability $p_a, p_m \in \{\frac{i}{n}\}_{i=0}^{n}$ is chosen from the interval $[0, 1]$ evenly discretized with $n = 10$. For a given (achievement or maintenance) commitment c, we measure the suboptimality with respect to all the influences in $\mathcal{P}_u^1(c)$ as the provider's true influence. The parameters for the 1D Walk domain are the same as the example for Theorem 1 except that the horizon is longer, $L = 10, L_0 = 3, H = 20$.

Fig. 5 shows the mean, minimum, and maximum suboptimality over all realizations of the provider's true influence $P_u \in \mathcal{P}_u^1$ for commitment time $T_a, T_m \in \{1, 5, 10, 15\}$. We see that for achievement commitments, the minimal enablement duration (or equivalently the minimal value timing) influence incurs the lowest suboptimality. The more expensive minimax regret timing influence has comparable suboptimality. The other two, maximal enablement duration and the constant toggling influences, incur the most suboptimality overall. For maintenance commitments, the minimal enablement duration

and the minimax regret influences incur the *most* suboptimality overall, and, among the other three approximate influences, it is difficult to identify a single best influence that reliably reduces the suboptimality for all the maintenance commitments. The maximal enablement duration strategy has the lowest mean suboptimality overall, yet the maximum suboptimality it induces over admissible true influences can be quite high especially when the probability of toggling $(1 - p_m)$ is close to one. On the other hand, the constant toggling strategy incurs higher mean suboptimality than the maximal enablement duration, yet its maximum suboptimality is consistently lower. The suboptimality of the minimum value timing strategy is the median among the five.

For both achievement and maintenance commitments, higher suboptimality tends to be induced with a larger commitment time and a larger probability of toggling (i.e., larger p_a for toggling from u^- to u^+ for achievement commitments, and larger $(1 - p_m)$ for toggling from u^+ to u^- for maintenance commitments). This is because the recipient has more uncertainty about the provider's true influence when both the commitment time and probability of toggling are larger. (Note that, to make comparing easier, the x-axes of the graphs go from lower to higher probability of toggling, which again are p_a and $(1 - p_m)$ for achievement and maintenance commitments, respectively.)

In the extreme, as shown in Figs. 5a and 5b, for commitment time $T_a = T_m = 1$ the recipient has no uncertainty about the toggling time, and hence the suboptimality is zero given that the provider's true influence $P_u \in \mathcal{P}_u^1$ must match the one consistent with the commitment probability. Similarly, for any commitment time, when the probability of toggling is $p_a = (1 - p_m) = 0$, the suboptimality is also zero since the recipient's approximate influence must match the provider's true influence in the sense that there is no toggling in either of the two influences.

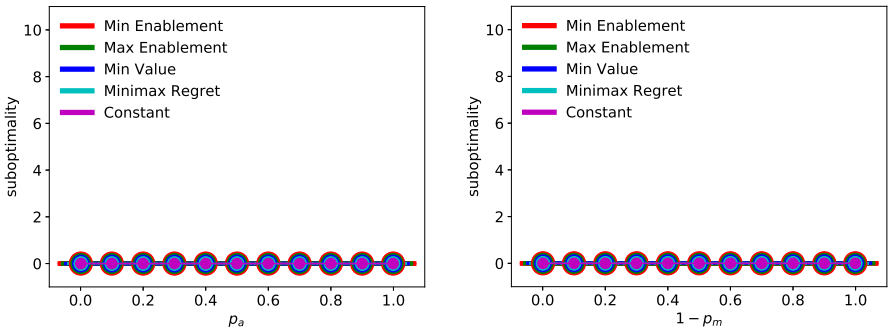
The same reasoning explains why the largest suboptimality occurs at $p_a = 1 - p_m = 1$. Intuitively, uncertainty (hence potential suboptimality) is larger when the "gray" area between $t = 0$ and the commitment time is larger in Figs. 2 and 3, which is when the commitment time is later and/or the toggling probability is larger.

4.2. Suboptimality for value maximizer commitments

In Section 4.1, we measured the suboptimality that is induced by general commitments across a spectrum of commitment times and probabilities. This is informative over the space of *possible* commitments, but of course some commitments are less likely to induce useful cooperation (e.g., achievement commitments with very low probabilities and/or very late times) than others. We now complement those results by looking more directly at suboptimality for commitments that are more likely to be adopted. Here we introduce an environment that explicitly incorporates the provider's commitment value, and we examine commitments that are rationally chosen to be value maximizers, which either maximize the provider's commitment value $v^{\mathrm{p}}(c)$, the recipient's commitment value $v^{\mathrm{r}}(c)$, or the joint commitment value $v^{\mathrm{p}}(c) + v^{\mathrm{r}}(c)$ [14]. These kinds of rationally-chosen commitments are more likely to be the ones adopted by the agents, and they are not chosen in favor of a particular type of commitment, nor in favor of a particular approximate influence strategy. Moreover, in both Sections 3 and 4.1 we modeled the virtual provider's true influence $P_u \in \mathcal{P}_u^1(c)$ toggling u at a single time step no later than the commitment time. In this section, we are concerned with the more general situation in which the true influence P_u is not restricted to be an element in $\mathcal{P}_u^1(c)$; instead, P_u is naturally determined by the provider's policy that maximizes its own value while respecting the commitment semantics. We first describe the recipient's and the provider's environments below.

The recipient's environment The recipient's environment is the same 1D Walk domain used for the previous experiments in Section 4.1, with a horizon again set to $H = 20$ and the starting location L_0 randomly chosen from locations $1 - 8$. Like in the proof of Theorem 3, there is additionally a one-time reward of r_{left}, an integer value in $[0, 10]$. In a specific instantiation of the recipient's MDP, L_0 and r_{left} are fixed, and they are randomly chosen to create various MDPs for the recipient. Since the left end has higher rewards than the right end, if the recipient's start position is close enough to the left end and the provider commits to opening the gate early enough (or keeping the gate open late enough) with a high enough probability, the recipient should utilize the commitment by checking if the gate is open by the commitment time, and pass through it if so; otherwise, the recipient should simply ignore the commitment and move to the right end. Thus, the various instances of the recipient's MDP include diverse preferences regarding the commitments.

The provider's environment The provider's MDP also has a horizon $H = 20$, and is randomly generated from a distribution designed such that, in expectation, the provider's value when enabling the precondition is smaller than when not enabling it. This introduces tension in the provider between enabling the precondition to help the recipient, versus increasing its own reward. We now describe the provider's MDP-generating distribution. The MDP has 10 states the provider can be in at any time step, one out of which is an absorbing state denoted as s^+, and where the initial state is chosen from the non-absorbing states. There are 3 actions. For each state-action pair $(s^{\mathrm{p}}, a^{\mathrm{p}})$ where $s^{\mathrm{p}} \neq s^+$, the transition function $P^{\mathrm{p}}(\cdot|s^{\mathrm{p}}, a^{\mathrm{p}})$ is determined independently by filling the 10 entries with values uniformly drawn from $[0, 1]$, and normalizing $P^{\mathrm{p}}(\cdot|s^{\mathrm{p}}, a^{\mathrm{p}})$. For achievement commitments, feature u takes the value of u^+ only in the absorbing state, i.e. $u^+ \in s^{\mathrm{p}}$ if and only if $s^{\mathrm{p}} = s^+$, and the reward $R^{\mathrm{p}}(s^{\mathrm{p}}, a^{\mathrm{p}})$ for a non-absorbing state $s^{\mathrm{p}} \neq s^+$ is sampled uniformly and independently from $[0, 1]$, and for the absorbing state $s^{\mathrm{p}} = s^+$ is zero, meaning the provider prefers to avoid the absorbing state, but that state is the only one that enables the precondition and realizes the achievement commitment. For maintenance commitments, feature u takes



(a) Achievement, $T_a = 1$ (b) Maintenance, $T_m = 1$

(c) Achievement, $T_a = 5$ (d) Maintenance, $T_m = 5$

(e) Achievement, $T_a = 10$ (f) Maintenance, $T_m = 10$

(g) Achievement, $T_a = 15$ (h) Maintenance, $T_m = 15$

Fig. 5. Suboptimality in 1D Walk. Please view in color. The results are for the recipient with $L = 10, L_0 = 3, H = 20$. Markers on the curves show the mean suboptimality over possible true influences that toggles at a single time step before the commitment time, $P_u \in \mathcal{P}_u^1(c)$. Bars show the minimum and maximum.

Table 2
Suboptimality for maximizer commitments (without action a^+ for the provider). The suboptimality is normalized by $v^*_{M^+} - v^*_{M^-}$. The results are means and standard errors (in parentheses). Mean + standard error below 5% are underlined, and below 1% are in bold.

		Suboptimality (%)		
		Provider Value Maximizer	Joint Value Maximizer	Recipient Value Maximizer
Achv.	Min Enablement Min Value	**0.21 (0.03)**	**0.27 (0.03)**	**0.40 (0.03)**
	Max Enablement	26.51 (0.61)	29.25 (0.65)	28.75 (0.65)
	Minimax Regret	6.44 (0.23)	7.78 (0.26)	7.20 (0.26)
	Constant Toggling	**0.03 (0.01)**	**0.06 (0.01)**	**0.97 (0.04)**
Maint.	Min Enablement	9.93 (0.83)	4.00 (0.56)	1.55 (0.18)
	Max Enablement	11.04 (0.74)	11.04 (0.74)	11.04 (0.74)
	Min Value	15.02 (1.11)	10.82 (1.06)	8.74 (0.96)
	Minimax Regret	10.17 (0.83)	7.24 (0.62)	7.63 (0.58)
	Constant Toggling	9.47 (1.01)	7.56 (0.91)	**0.02 (0.01)**

Table 3
Suboptimality for maximizer commitments ($p^+ = 0$). The suboptimality is normalized by $v^*_{M^+} - v^*_{M^-}$. The results are means and standard errors (in parentheses). Mean + standard error below 5% are underlined, and below 1% are in bold.

		Suboptimality (%)		
		Provider Value Maximizer	Joint Value Maximizer	Recipient Value Maximizer
Achv.	Min Enablement Min Value	**0.01 (0.01)**	**0.01 (0.01)**	**0.31 (0.02)**
	Max Enablement	2.69 (0.21)	14.81 (0.33)	34.28 (0.69)
	Minimax Regret	**0.66 (0.09)**	6.01 (0.25)	10.15 (0.37)
	Constant Toggling	**0.01 (0.01)**	**0.01 (0.01)**	**0.46 (0.02)**
Maint.	Min Enablement	6.31 (0.67)	**0.68 (0.21)**	**0.01 (0.01)**
	Max Enablement	8.12 (0.63)	8.12 (0.63)	4.80 (0.49)
	Min Value	14.42 (1.15)	6.62 (0.87)	0.97 (0.33)
	Minimax Regret	7.30 (0.65)	7.56 (0.60)	4.45 (0.46)
	Constant Toggling	6.33 (0.83)	2.56 (0.91)	**0.01 (0.01)**

the value of u^+ only in the non-absorbing states, i.e. $u^+ \in s^{\mathrm{p}}$ if and only if $s^{\mathrm{p}} \neq s^+$, and the reward $R^{\mathrm{p}}(s^{\mathrm{p}}, a^{\mathrm{p}})$ for a non-absorbing state $s^{\mathrm{p}} \neq s^+$ is sampled uniformly and independently from $[-1, 0]$, and for the absorbing state $s^{\mathrm{p}} = s^+$ is zero, meaning the provider prefers to reach the absorbing state, but that state disables the precondition and fails the maintenance commitment.

We observe that, for small values of commitment time, the provider's maximum feasible probability of toggling u, or equivalently reaching s^+, by the commitment time is fairly low. Hence, in some experiments we also introduce a fourth action for the provider, a^+, such that, after taking a^+ in any non-absorbing state $s^{\mathrm{p}} \neq s^+$, the provider will transit to the absorbing state s^+ with probability $p^+_{s^{\mathrm{p}}}$, and will stay in the current state s^{p} with probability $1 - p^+_{s^{\mathrm{p}}}$. For each non-absorbing state $s^{\mathrm{p}} \neq s^+$, $p^+_{s^{\mathrm{p}}}$ is sampled from a Gaussian distribution and then clipped into $[0, 1]$. In a specific instantiation of the provider's MDP, the mean of the Gaussian distribution, denoted as p^+, is chosen from $\{0, 0.5, 0.9\}$, and the standard deviation is fixed to 0.1.

Results Tables 2, 3, 4, and 5 show the suboptimality for the value maximizer commitments without action a^+, and with action a^+ and $p^+ = 0$, 0.5, and 0.9, respectively, each reporting the means and standard errors over 2500 randomly-generated pairs of the provider's MDP and the recipient's MDP. Since the problem instances have different reward scales, the suboptimality is normalized by the bound in Equation (4), i.e. $v^*_{M^+} - v^*_{M^-}$. The tables highlight strategies that induce low suboptimality for certain types of value maximizer commitments, with mean+error $\leq 5\%$ underlined and mean+error $\leq 1\%$ in bold.

For achievement commitments, the minimal enablement duration (or equivalently the minimal value timing) strategy consistently induces suboptimality below 1% with or without action a^+, for all three types of maximizer commitment, while the maximal enablement duration and the minimax regret often induce suboptimality higher than 5%. Table 2 shows that, without action a^+, the constant toggling influence also induces suboptimality below 1% for all three types of maximizer commitment, and this also holds with action a^+ and a small $p^+ = 0$ as shown in Table 3. However, as p^+ increases, the constant toggling influence can induce suboptimality higher than 5%, especially for the joint value maximizer commitments, as shown in Tables 4 and 5. Generally, the provider value maximizers are "weak" achievement commitments with late

Table 4
Suboptimality for maximizer commitments ($p^+ = 0.5$). The suboptimality is normalized by $v^*_{M^+} - v^*_{M^-}$. The results are means and standard errors (in parentheses). Mean + standard error below 5% are underlined, and below 1% are in bold.

		Suboptimality (%)		
		Provider Value Maximizer	Joint Value Maximizer	Recipient Value Maximizer
Achv.	Min Enablement Min Value	**0.16(0.02)**	**0.12 (0.01)**	**0.14 (0.01)**
	Max Enablement	28.00 (0.63)	31.69 (0.69)	38.08 (0.63)
	Minimax Regret	6.82 (0.23)	9.67 (0.34)	4.53 (0.30)
	Constant Toggling	**0.01 (0.01)**	10.66 (0.43)	**0.02 (0.01)**
Maint.	Min Enablement	22.66 (0.70)	3.08 (0.36)	1.74 (0.20)
	Max Enablement	45.09 (1.54)	45.09 (1.54)	45.09 (1.54)
	Min Value	6.33 (0.39)	2.81 (0.35)	2.17 (0.31)
	Minimax Regret	22.99 (0.70)	4.72 (0.35)	10.62 (0.65)
	Constant Toggling	4.36 (0.37)	2.48 (0.34)	**0.01 (0.01)**

Table 5
Suboptimality for maximizer commitments ($p^+ = 0.9$). The suboptimality is normalized by $v^*_{M^+} - v^*_{M^-}$. The results are means and standard errors (in parentheses). Mean + standard error below 5% are underlined, and below 1% are in bold.

		Suboptimality (%)		
		Provider Value Maximizer	Joint Value Maximizer	Recipient Value Maximizer
Achv.	Min Enablement Min Value	**0.16 (0.02)**	**0.10 (0.01)**	**0.01 (0.01)**
	Max Enablement	27.89 (0.63)	32.40 (0.67)	32.36 (0.67)
	Minimax Regret	6.71 (0.23)	9.58 (0.36)	5.15 (0.31)
	Constant Toggling	**0.01 (0.01)**	52.79 (1.48)	**0.01 (0.01)**
Maint.	Min Enablement	10.40 (0.48)	**0.71 (0.16)**	1.72 (0.20)
	Max Enablement	46.83 (1.32)	50.00 (1.34)	50.00 (1.34)
	Min Value	1.66 (0.13)	**0.52 (0.11)**	**0.45 (0.10)**
	Minimax Regret	9.17 (0.42)	5.62 (0.24)	13.13 (0.64)
	Constant Toggling	1.80 (0.13)	**0.45 (0.10)**	**0.01 (0.01)**

commitment time T_a and low commitment probability p_a, while the recipient value maximizers are "strong" commitments with early T_a and high p_a. Since later commitment time T_a and higher commitment probability p_a often cause the recipient more uncertainty about the true influence and therefore higher suboptimality (as evidenced by the results in Figs. 5a, 5c, 5e, and 5g), it is difficult to predict which type of value maximizer induces higher suboptimality. Thus, it should be unsurprising that some strategies work well for one type of value maximizer achievement commitment but not for another. Nonetheless, the minimal enablement duration (or equivalently the minimal value timing) strategy consistently induces low suboptimality for all types of value maximizer achievement commitment.

For maintenance commitments, the results show that none of the five strategies has suboptimality below 1% consistently for all three types of maximizer commitment, with or without action a^+. Overall, the suboptimality of all five strategies for maintenance is significantly higher than the suboptimality of the minimal enablement duration strategy for achievement. It is worth noting that, while the maximal enablement duration was an above-average strategy for maintenance commitments if the true influence is chosen from $\mathcal{P}^1_u(c_m)$ which toggles only at a single time step (shown in Figs. 5b, 5d, 5f, and 5h), here we see that the maximal enablement duration is overall the worst among the five strategies, confirming that being overly optimistic does not result in robust risk-aware interpretation of maintenance commitments. Similarly to achievement commitments, it is difficult to predict which type of value maximizer maintenance commitment is harder for the recipient to model, and a strategy can work well for one value maximizer but not for another. For example, constant toggling induces the lowest suboptimality for recipient value maximizer maintenance commitments, suggesting that, when the commitment time T_m is late and the toggling probability $\leq 1 - p_m$ is low, it is empirically better to model the toggling more often than a single time step. However, such a claim about the constant toggling strategy does not hold for joint value maximizers, as shown in Table 3.

4.3. Summary

While our theoretical analyses in Section 3 showed that suboptimality could not be bounded in the worst case for most of the interpretation strategies (and for all of the interpretation strategies when applied to maintenance commitments), this does not necessarily answer the question of whether the strategies can nonetheless work well for some regions of the

commitment space, or for commitments that are most likely to arise. The empirical results in this section begin to answer such questions. Perhaps the most important (if not altogether surprising) high-level insight from the results is that a longer commitment time and larger difference between the initial and the commitment probabilities (or equivalently, in Fig. 2, the larger the "gray" area up to the commitment time) is, the larger the risk and degree of suboptimality there generally is.

This in turn suggests that one way to improve recipient performance for maintenance commitments is for the commitment to specify intermediate probabilities, so as to reduce the recipient's uncertainty. That is, based on our results, the question arises, both within our research project and more broadly for researchers studying coordination mechanisms, as to whether commitment specifications for maintenance commitments should differ from those for achievement commitments. Of course, any changes to the specifications would also impact the provider too, and in the case of needing to specify and adhere to intermediate probabilities, the provider's flexibility over choices of policies would be reduced. Thus, the question of finding the right compromise between abstraction for the provider's flexibility and detail for the recipient's risk-aware interpretation poses an interesting joint optimization problem for future research.

5. Conclusion

In this article, we focus on how the recipient of a commitment should robustly plan against the risk associated with interpreting a probabilistic achievement or maintenance commitment. We have formally characterized and analyzed an algorithmic and representational strategy that had previously only been described and justified intuitively. In so doing, we have carefully explained why it has succeeded for probabilistic commitments of achievement but has failed for those of maintenance, despite the fact that the two types of commitments are identical except in their directions of precondition toggling. Contrary to intuitions, the difficulty lies not on the provider's side, but as we have analytically and empirically shown it lies in the recipient's uncertainty in how to approximate the provider's influence that probabilistically changes the precondition over time. Although this uncertainty is present for both commitment types, we have proven that the suboptimality it induces is effectively unbounded only in the case of maintenance commitments. Further, we have defined alternative new interpretation strategies for approximating the uncertain influences, including in sophisticated ways intended to overcome these limitations, but we have analytically and empirically shown that they still fall short.

Our work thus proves how a risk-aware recipient of an achievement commitment can easily approximate the provider's uncertain influence in a manner that bounds suboptimality regardless of the provider's true influence. Further, by proving that such risk-aware interpretation is elusive for a maintenance commitment, our work in this article encourages future research in coordination especially for maintenance. One immediate next step is to try to develop and investigate better interpretation strategies than the ones we studied in this work. Possibly, instead of developing environment-agnostic heuristics (e.g., minimal/maximal enablement duration), one can also approach the recipient's modeling problem by directly minimizing the worst-case suboptimality in a domain-specific manner, such that the suboptimality minimizer strategy will depend on both the commitment and the recipient's MDP environment.

Another avenue for future research is to relax the assumptions of only a single toggling, and that $u = u^-$ can never be preferred (Assumptions 1 and 2), which together ensured the presence of only a single commitment. Since these assumptions were largely made to avoid confounding factors in the analyses and experiments, we speculate that suboptimality bounds (or lack thereof) for the different types of commitments and influence approximations will apply to each of the commitments individually in a multi-commitment setting. However, without Assumptions 1 and 2, calculating the bounds will be more challenging than in Lemma 1, likely drawing on minimax regret concepts like those used to compute the $\widehat{P}_u^{\text{minimax}}(c_a)$ and $\widehat{P}_u^{\text{minimax}}(c_m)$ approximate influences. Further, such settings could introduce other, possibly overriding, concerns besides suboptimality risk when choosing among influence approximations. As a simple example, let's say that u can toggle twice, from u^- to u^+ and back to u^-, and that the window is short between when the provider has committed to achieve u^+ and when it has committed to maintaining u^+ until. In this case, the recipient might be better off approximating the achievement influence more optimistically—anticipating that u^+ could hold earlier than the promised time—because the expected value of using u^+ over a larger time interval outweighs the risks of being wrong. Studying the interplay between influence approximations for combinations of commitments remains future work.

Finally, another direction for future study that was suggested in Section 4.3 is that instead of trying to build better approximations for the given abstract commitment specification, one can explore new specifications particularly for maintenance commitments that are more detailed than the single time step specifications for achievement commitments. These added details would reduce the recipient's uncertainty when creating its approximate influence, but, as a potential cost, could also reduce the flexibility of the provider to adapt its policy as it learns to improve its own model of the environment. As a first step in pursuing this problem, one can focus on the recipient and ask: to ensure a desired bound on the suboptimality for a given commitment, how many (and which) time steps should be specified for the provider's influence as part of the commitment? Then, after identifying what additional details would be most valuable to the recipient, one would need ways to estimate the expected costs to the provider of being locked in by having to satisfy those details that it would now be committed to.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

We would like to thank the anonymous reviewers for their many helpful suggestions for improving this article. This work was supported in part by the Air Force Office of Scientific Research under grant FA9550-15-1-0039, the Open Philanthropy Project to the Center for Human-Compatible AI, and National Science Foundation grant IIS 2154904. Opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] M.P. Singh, Semantical considerations on dialectical and practical commitments, in: Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, 2008, pp. 176–181.
[2] N.R. Jennings, Commitments and conventions: the foundation of coordination in multi-agent systems, Knowl. Eng. Rev. 8 (3) (1993) 223–250.
[3] J. Xing, M.P. Singh, Formalization of commitment-based agent interaction, in: Proc. of the 2001 ACM Symposium on Applied Computing, 2001, pp. 115–120.
[4] M. Winikoff, Implementing flexible and robust agent interactions using distributed commitment machines, Multiagent Grid Syst. 2 (4) (2006) 365–381.
[5] E.H. Durfee, S. Singh, On the trustworthy fulfillment of commitments, in: N. Osman, C. Sierra (Eds.), Autonomous Agents and Multiagent Systems: AAMAS 2016 Workshops Best Papers, Springer, 2016, pp. 1–13.
[6] P. Xuan, V.R. Lesser, Incorporating uncertainty in agent commitments, in: International Workshop on Agent Theories, Architectures, and Languages, Springer, 1999, pp. 57–70.
[7] R. Maheswaran, P. Szekely, M. Becker, S. Fitzpatrick, G. Gati, J. Jin, R. Neches, N. Noori, C. Rogers, R. Sanchez, et al., Predictability & criticality metrics for coordination in complex environments, in: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems, 2008, pp. 647–654.
[8] S.J. Witwicki, E.H. Durfee, Commitment-based service coordination, Int. J. Agent-Oriented Softw. Eng. 3 (2009) 59–87.
[9] Q. Zhang, E.H. Durfee, S. Singh, A. Chen, S.J. Witwicki, Commitment semantics for sequential decision making under reward uncertainty, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016, pp. 3315–3323.
[10] B.J. Clement, S.R. Schaffer, Exploiting C-TÆMS models for policy search, in: ICAPS Workshop on Multiagent Planning, 2008.
[11] R.P. Goldman, D.J. Musliner, E.H. Durfee, M.S. Boddy, Coordinating highly contingent plans: biasing distributed MDPs towards cooperative behavior, in: ICAPS Workshop on Multiagent Planning, 2008.
[12] L.M. Hiatt, Probabilistic plan management, Ph.D. thesis, Carnegie Mellon University, 2009.
[13] S.J. Witwicki, E.H. Durfee, Commitment-driven distributed joint policy search, in: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems, 2007, pp. 480–487.
[14] Q. Zhang, E.H. Durfee, S. Singh, Efficient querying for cooperative probabilistic commitments, in: Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, 2021, pp. 11378–11386.
[15] Q. Zhang, S. Singh, E. Durfee, Minimizing maximum regret in commitment constrained sequential decision making, in: Proceedings of the Twenty-Seventh International Conference on Automated Planning and Scheduling, 2017, pp. 348–356.
[16] Q. Zhang, E.H. Durfee, S. Singh, Semantics and algorithms for trustworthy commitment achievement under model uncertainty, Auton. Agents Multi-Agent Syst. 34 (1) (2020) 19.
[17] M.P. Singh, Commitments in multiagent systems: some history, some confusions, some controversies, some prospects, in: The Goals of Cognition. Essays in Honor of Cristiano Castelfranchi, London, 2012, pp. 601–626.
[18] J. Vokrínek, A. Komenda, M. Pechoucek, Decommitting in multi-agent execution in non-deterministic environment: experimental approach, in: Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems, 2009, pp. 977–984.
[19] T. Sandholm, V.R. Lesser, Leveled commitment contracts and strategic breach, Games Econ. Behav. 35 (2001) 212–270.
[20] N. Desai, N.C. Narendra, M.P. Singh, Checking correctness of business contracts via commitments, in: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems, 2008, pp. 787–794.
[21] A. Günay, Y. Liu, J. Zhang, Promoca: probabilistic modeling and analysis of agents in commitment protocols, J. Artif. Intell. Res. 57 (2016) 465–508.
[22] R.F. Pereira, N. Oren, F. Meneguzzi, Detecting commitment abandonment by monitoring sub-optimal steps during plan execution, in: Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems, 2017, pp. 1685–1687.
[23] P. Telang, F.R. Meneguzzi, M. Singh, Hierarchical planning about goals and commitments, in: Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems, 2013.
[24] T.C. King, A. Günay, A.K. Chopra, M.P. Singh, Tosca: operationalizing commitments over information protocols, arXiv preprint, arXiv:1708.03209.
[25] H. Bannazadeh, A. Leon-Garcia, A distributed probabilistic commitment control algorithm for service-oriented systems, IEEE Trans. Netw. Serv. Manag. 7 (4) (2010) 204–217.
[26] G.A. Kaminka, A. Yakir, D. Erusalimchik, N. Cohen-Nov, Towards collaborative task and team maintenance, in: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems, 2007, pp. 1–8.
[27] A. Newell, Unified Theories of Cognition, Harvard University Press, 1994.
[28] G.A. Kaminka, I. Frenkel, Towards flexible teamwork in behavior-based robots, in: Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems, 2005, pp. 1355–1356.
[29] C. Baral, T. Eiter, M. Bjäreland, M. Nakamura, Maintenance goals of agents in a dynamic environment: formulation and policy construction, Artif. Intell. 172 (12–13) (2008) 1429–1469.
[30] F. Bacchus, F. Kabanza, Planning for temporally extended goals, Ann. Math. Artif. Intell. 22 (1) (1998) 5–27.

[31] C.M. Özveren, A.S. Willsky, P.J. Antsaklis, Stability and stabilizability of discrete event dynamic systems, J. ACM 38 (3) (1991) 729–751.
[32] S. Duff, J. Thangarajah, J. Harland, Maintenance goals in intelligent agents, Comput. Intell. 30 (1) (2014) 71–114.
[33] F.A. Oliehoek, C. Amato, A Concise Introduction to Decentralized POMDPs, Springer, 2016.
[34] S.J. Witwicki, E.H. Durfee, Influence-based policy abstraction for weakly-coupled Dec-POMDPs, in: Proceedings of the Twentieth International Conference on Automated Planning and Scheduling, 2010, pp. 185–192.
[35] R. Becker, S. Zilberstein, V. Lesser, C.V. Goldman, Transition-independent decentralized Markov decision processes, in: Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems, 2003, pp. 41–48.
[36] P. Varakantham, J.-y. Kwak, M. Taylor, J. Marecki, P. Scerri, M. Tambe, Exploiting coordination locales in distributed pomdps via social model shaping, in: Nineteenth International Conference on Automated Planning and Scheduling, 2009.
[37] L.S. Shapley, Stochastic games, Proc. Natl. Acad. Sci. 39 (10) (1953) 1095–1100.
[38] F.A. Oliehoek, S.J. Witwicki, L.P. Kaelbling, Influence-based abstraction for multiagent systems, in: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012, pp. 1422–1428.
[39] F.A. Oliehoek, M.T. Spaan, S.J. Witwicki, Influence-optimistic local values for multiagent planning, in: Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, 2015, pp. 1703–1704.
[40] K.V. Hindriks, M.B. van Riemsdijk, Satisfying maintenance goals, in: 5th Int. Workshop Declarative Agent Languages and Technologies (DALT), 2007, pp. 86–103.
[41] M.P. Singh, Multiagent systems as spheres of commitment, in: Proceedings of the ICMAS Workshop on Norms, Obligations, and Conventions, Citeseer, 1996.
[42] J. Xing, M.P. Singh, Engineering commitment-based multiagent systems: a temporal logic approach, in: Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems, 2003, pp. 891–898.
[43] E. Altman, Constrained Markov Decision Processes: Stochastic Modeling, Routledge, 1999.
[44] M. Steinmetz, J. Hoffmann, O. Buffet, Goal probability analysis in probabilistic planning: exploring and enhancing the state of the art, J. Artif. Intell. Res. 57 (2016) 229–271.
[45] C. Jin, T. Jin, H. Luo, S. Sra, T. Yu, Learning adversarial Markov decision processes with bandit feedback and unknown transition, in: International Conference on Machine Learning, 2020, pp. 4860–4869.
[46] S. Duff, J. Thangarajah, J. Harland, Maintenance goals in intelligent agents, Comput. Intell. 30 (1) (2014) 71–114.
[47] T. Smith, R. Simmons, Heuristic search value iteration for POMDPs, in: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, 2004, pp. 520–527.