Action Selection for Transparent Planning

Aleck M. MacNally, Nir Lipovetzky, Miquel Ramirez and Adrian R. Pearce
The University of Melbourne
Melbourne, Australia
aleck.macnally, nir.lipovetzky, miguel.ramirez, adrianrp@unimelb.edu.au

Abstract

We introduce a novel framework to formalize and solve *transparent planning tasks* by executing actions selected in a suitable and timely fashion. A *transparent planning task* is defined as a task where the objective of the agent is to communicate its true goal to observers, thereby making its intentions and its action selection *transparent*. We formally define and model these tasks as Goal POMDP's where the state space is the Cartesian product of the states of the world and a given set of hypothetical goals. Action effects are deterministic in the world states of the problem but probabilistic in the observer's beliefs. Transition probabilities are obtained from making a call to a model–based plan recognition algorithm, which we refer to as an *observer stereotype*. We propose an action selection strategy via on–line planning that seeks actions to quickly convey the goal being pursued to an observer assumed to fit a given stereotype. In order to keep run–times feasible, we propose a novel model–based plan recognition algorithm that approximates well–known probabilistic plan recognition methods. The resulting on–line planner, after being evaluated over a diverse set of domains and three different observer stereotypes, is found to convey goal information faster than purely goal–directed planners.

Introduction

Understanding the intentions and plans of agents, humans or otherwise, has been identified as crucial in key AI research areas such as intelligent user interfaces, dialogue in natural language, cooperation in multi–agent systems, and assisted cognition (Carberry 2001; Cohen, Perrault, and Allen 1981; Pentney et al. 2006; Yang 2009) and substantial literature exists proposing several different formal and computational frameworks (Avrahami-Zilberbrand and Kaminka 2005; Geib and Goldman 2009; Ramirez and Geffner 2010).

On the other hand, the complementary problem of *assisting* an observer to determine the goal being pursued, has received less attention (Chakraborti et al. 2017). Informally, and from the perspective of model–based approaches to planning and plan recognition, this is a problem of planning – seeking actions – under specific constraints or in response to *feedback* obtained from the observer. In turn, the ability of autonomous systems to generate behaviour that is, according to some measure of efficiency, *easy* to interpret is identified as a critical and complementary component

This paper has been published at AAMAS 2018

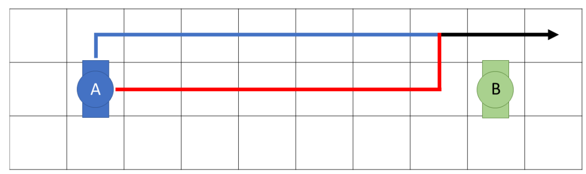

Figure 1: Shows a person, **A,** walking in a corridor who must avoid a collision with a second person, **B**. Two possible paths are shown for person **A**.

to robust intention recognition in human–in–the–loop systems and joint human–robotic teams operating in dynamic environments. Existing approaches reformulate the planning model used by the acting agent (*actor*) in a way that is beneficial to the agent that observes (*observer*). Chakraborti et al (2017) propose a framework for *Model Reconciliation* that aims to elicit changes on-line in planning models so that the cost optimal plans pursued by the actor match those considered by the observer. Moving the focus away from the dialogical dimension of the problem, Keren et al (2014) *Goal Recognition Design* aims at redesigning the environment where plans are executed to guarantee that the degree of ambiguity in optimal plans for a given set of goals is bounded. Interestingly, recent work (Masters and Sardina 2017) aims at *exploiting* inherent ambiguity in the structure of the environment to *delay* the identification of goals. In this paper we present a formal and computational framework for transparent planning, when *on–line co–ordinated* or *off–line* model reformulation is not suitable or possible. We formalize the problem and model it using the Goal POMDP framework (Kaelbling, Littman, and Cassandra 1999), and the Functional STRIPS language to describe states and actions (Geffner 2000; Bonet and Geffner 2000). The goal pursued is then a constraint on the posterior probabilities that the observer assigns to a set of *hypothetical* goals, so that the *actual* goal is the best *predictor* for the actions executed up to a given point in time. We propose an on–line, polynomial and approximate algorithm to select actions, that embeds a model–based plan recognition algorithm to procedurally determine the Goal POMDP transition probabilities required to track the progression of beliefs. We evalu-

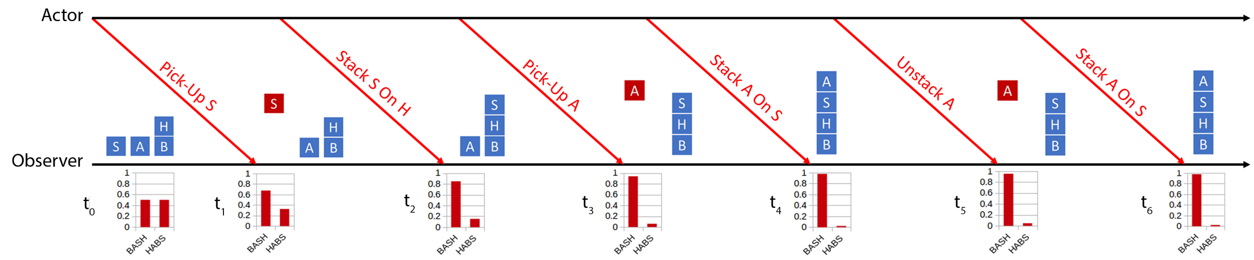

Figure 2: Shows the change in the beliefs of an observer in response to the actions it observes the actor executing. Time flows left to right. The beliefs are shown in the graphs at the base of the figure. Next to each action is a diagram illustrating the state following that action.

ate the resulting planner over several domains, which in this work are restricted to actions whose effects over state variables modelling the environment are deterministic. The results obtained indicate that the proposed scheme, accelerates the convergence of the observer's beliefs towards a probability distribution with the desired properties, and is robust with respect to the assumptions made on the posterior used by the observer to interpret observed actions.

Motivation

It is important in robot-human co-operative teams that each member knows the intentions and roles of their collaborators (Stubbs, Wettergreen, and Hinds 2007). In this paper we focus specifically on how an agent may communicate its goals to its fellow teammates. An obvious method of achieving this is to have the agent explicitly transmit all relevant information to observing agents. This can be achieved through some form of communication protocol, such as, when the observer is human, a computer screen or through natural language. Unfortunately this solution is not as straight-forward as it seems. When the observing agent is human, the internal state of an acting agent must be distilled in such a way that the information relevant to humans is accessible. However, what is accessible may depend on the human. Issues with training, language and culture will affect whether they can correctly comprehend any communication.

An explicit communication protocol requires that both communicating parties understand the protocol in its entirety. *Transparent Planning* represents a method for communicating information implicitly by acting in a way which is recognizable to observers. In comparison to explicit communication methods this form of implicit communication only requires that there is agreement regarding what constitutes natural behaviour, rather than a complete protocol.

We can see *transparent planning* used in human-human collaboration especially in cases where there is no shared communication protocol such as between an adult and a toddler (Warneken and Tomasello 2006). In the work by Warneken et al. (2006) an adult executes actions, which are chosen because they indicate what the adult is attempting to achieve, to communicate to an observing toddler. In this case the toddler does not yet understand languages but can interpret the intentions of the adult by reasoning about its actions, and then it may choose to assist.

In cases where there is an established communication protocol, there is also the problem of communicating over a channel. The form of this channel greatly changes the difficulty of the communication. For instance, if we take the example of a human supervising a team of robots, it is clear that if each of the robots was audibly broadcasting its intentions, the human would be unable to process the barrage of information. Communicating by acting transparently mostly uses a visual communication channel which, whilst subject to its own set of problems, will be useful in cases where other channels are intractable, as in the above example.

Illustration Figure 1 shows a situation in which two people in a corridor must navigate around each other in order to pass. For a successful solution each human must know which side the other will pass on. In this case **A** has decided that they will pass **B** on the their right-hand side. It is imperative that **B** understand **A**'s intention to avoid a collision. Two possible paths are illustrated in Fig. 1 from which **A** may choose. The blue path clearly communicates **A**'s intentions in advance of a possible collision, while the red path does not. In situations like this, humans do not generally broadcast their intentions audibly to other humans but instead act transparently, as in this example by moving to the decided side early on in their plan. We seek to produce this behaviour in human-robot collaborations.

In this work we do not require that the agent who is acting *transparently* commits to achieving its goal *and* communicating it *simultaneously*. The actor's objective is to communicate its goal as efficiently as possible regardless of how far this might remove the actor from the goal. We wish to have the planner use actions as the vocabulary for describing its intentions rather than for their true effects. For example in Figure 2 we can see the interaction of a human and a robot over time. The robot is planning in the BLOCKS WORDS domain (Ramirez and Geffner 2009), in which the robot must assemble alphabet blocks into a goal configuration by picking them up and putting them down. The set of possible goals $\mathcal{G}$ consists of two words: BASH and HABS. The robot's goal is BASH. The initial state can be seen on the far left of Figure 2. The optimal action for both goals is to pick up H and then put it on the table, but if the actor executes these actions the observer will only see a partial plan that is equally good for both goals. So instead it picks up S which only occurs optimally in plans to achieve the word BASH at t_0. It then *composes* the word ASH on top of the B. This clearly conveys the goal of the actor very rapidly as can be seen by the graphs at the bottom of the figure indicating the beliefs of the observer. Once the actor has written ASH it can do no more with its limited vocabulary and therefore un-stacks A as that is its only available action. At this point we may assume that an observer will have gauged the goal of the actor already and the actor may, therefore, act towards achieving the goal. If the actor assumes that the observer is not convinced it may instead again stack A on top of S, in an attempt to convey its goal with a limited vocabulary. This behaviour is reminiscent of that required by the adults in the research by Warneken et al. (2006), in which an adult repeatedly knocks into a closed cupboard to indicate to a toddler that it intended to place an object into the cupboard.

Related Work

The literature combining plan recognition and automated planning generally focuses on recognizing the intentions of exogenous agents and then acting in response to those intentions (Freedman and Zilberstein 2017), this work in contrast focuses on what an observer would gauge from observing the agent under our control. Much of the literature considering the effects actions have on the mental model of observers has been focused in the area of human-aware planning (Alami et al. 2006), in particular in the area of plan explanation. Chakraborti et al. (2017) considered plan explanation by taking into account differences between the actor's action model and the actor's model assumed by the human. They formulated a method for *reconciling* the human and robot action models by suggesting changes to the human model. In contrast, we do not intend to change the human action model but to generate the most unambiguous behaviour given an *observation* model. Zhang et al. (2017) also considered the difference between the actor's action model and the human's. They use this difference to produce plans expected by a human who has full knowledge of the goal of the actor, bypassing the need to explain actions. We produce plans without assuming the observer *knows* the actor's goal, we instead select actions to communicate this goal to the observer.

We take direct inspiration from Keren et al. (2014) work on *Goal Recognition Design*, where a point is made that even if actors plan optimally, it may still be difficult to identify the goal being pursued in a timely fashion. From this observation, we note that seeking a plan which is not necessarily optimal may allow an agent to convey its goal more efficiently.

An orthogonal work to the focus on explaining actions and, more generally, directing observers towards the goals that motivate plans, recent work by Masters and Sardina (2017) explores the active manipulation of observers' beliefs, proposing algorithms to produce *deceptive* plans in the path–planning domain. We manipulate beliefs too, but with the exact *opposite* intent and over *general* planning models.

Finally, this work is inextricably linked to the notion of model–based, probabilistic plan recognition (Ramirez and Geffner 2010). *Transparent Planning* internalizes this idea and uses it as an *element* of the planning model, so the evolution of observer beliefs over time is accounted for *explicitly* while seeking actions and plans.

Planning over Goal POMDPs

Goal POMDPs (Partially Observable Goal MDP) are a well–known framework (Kaelbling, Littman, and Cassandra 1999) for agents that have incomplete state information and receive indirect evidence of the actual state of the world via sensors or by their own reasoning. In this paper we mostly follow Bonet & Geffner's (2013) presentation of Goal POMDPs and define them as a tuple $M = \langle S, b_0, b_G, A, tr, c, \Omega, q \rangle$ whose elements are as follows. S is a non-empty, discrete and finite *state space*. b_0 is the *initial belief* state, a probability distribution $P(s)$ over *every* $s \in S$, representing the probability that the agent is in state s. We seek actions that transform b_0 into a new belief state in which the set of *constraints* on the state probabilities, b_G, is satisfied. The set of *actions* that an agent may execute in a state, s, is given by the function $A(s)$ and $c(a, s)$ is the cost of performing action a in s, which is a positive real number. tr is the state transition function and $tr(s'|s, a)$ gives the probability that the agent will transition into state s' after performing a in s.

Feedback from sensors consists of a finite set of *observation tokens* Ω. The probability of a token, $\omega \in \Omega$, being obtained having performed action a in state s is given by the *sensor model* $q(\omega|s, a)$.

A common way to solve POMDPs is to formulate them as completely observable MDPs over the *belief states* of the agent (Astrom 1965; Sondik 1978). While the effects of an action will have on a state may not be exactly predicted, the effects of actions on *belief states* can. Formally, the belief b_a that results from doing action a in belief b, and belief b_a^ω that result from observing ω after doing a in b, are:

$$b_a(s) = \sum_{s' \in S} tr(s'|s, a) b(s') \,, \tag{1}$$

$$b_a(\omega) = \sum_{s \in S} q(\omega|s, a) b_a(s) \,, \tag{2}$$

$$b_a^\omega(s) = q(\omega|s, a) b_a(s) / b_a(\omega) \quad \text{if } b_a(\omega) \neq 0. \tag{3}$$

As a result, the *partially observable* problem of going from an initial state to a goal state is transformed into the *completely observable* problem of going from one *initial belief state* into a *goal belief state*. We also note that MDPs where $tr(s'|s, a) > 0$ for exactly one state s' map directly onto deterministic state models (Geffner and Bonet 2013).

We use the Functional STRIPS (Geffner 2000) *language* to describe declaratively states and transitions, as Ramirez and Geffner (Ramirez and Geffner 2011) did, for the domains discussed in the Evaluation Section. Out of the full expressiveness of Functional STRIPS we only use a subset, we note the features required by our benchmarks next. State variables X have their domains to be either Boolean, and hence modeling the truth value of some arbitrary atomic formula, or a subset of $\mathbb{N}$ with small cardinality. S then corresponds to the set of possible valuations of X, and the value of x on state s is noted as $[x]^s$. Actions are described by a *precondition* list, $Pre(a)$, which is a list of literals of X, and an *effect* list, $Eff(a)$, which is a list of updates of the form $x := \top$ or $x := \bot$, where x is a variable with Boolean domain.

We refer the reader to (Francès and Geffner 2015) and (Bonet and Geffner 2001) for more detailed accounts of Functional STRIPS, and its extensions to deal with probabilistic effects and partially observable states, respectively.

Probabilistic Plan Recognition

A *probabilistic plan recognition problem* (Ramirez and Geffner 2010) is a tuple $T = \langle X, A, I, \mathcal{G}, O, Prob \rangle$ where:

- X and A are a set of Functional STRIPS variables and actions describing a *planning domain*,
- $I \in 2^X$ is an initial state,
- $\mathcal{G}$ is a set of candidate goal formulae from which a goal will be chosen,
- O is an *observation sequence*,
- $Prob$ is a prior probability distribution over $\mathcal{G}$, which represents previous knowledge of the likelihoods of each goal.

We note that T can be also defined over a *plan library* (Avrahami-Zilberbrand and Kaminka 2005; Geib and Goldman 2009), which is a set of sequences of actions $\pi = (a_0, \ldots, a_n)$, elicited from selecting plans that map initial state I into states that satisfy a goal $G \in \mathcal{G}$.

We assume that an observer's beliefs follow from the probability distribution, $P(G|O)$, where $G \in \mathcal{G}$, which is the solution to a plan recognition problem, T. We also assume that these *posterior* probabilities are computed via the application of Bayes' Rule as:

$$P(G|O) = \alpha \, P(O|G) \, P(G) \tag{4}$$

where α is a normalizing constant, and $P(G)$ is $Prob(G)$. Several definitions of likelihoods $P(O|G)$ have been put forward in the literature on model–based plan recognition, this work examines the following three. The first one, which we refer to as RG09, adopts a hard postulate on the rationality of agents pursuing hypothetical goals (Ramirez and Geffner 2009).

$$P(O|G) = \begin{cases} 1 & |c(G,O) - c(G)| = 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $c(G,O)$ is the cost of a plan for G *constrained* to be consistent with observation sequence O, and $c(G)$ is the cost of a plan for G *without* constraints. $c(*)$ maybe computed with an optimal planner or may be approximated by a satisficing planner or heuristic. The second definition, also by Ramirez & Geffner (2010) and noted as RG10, adopts a softer postulate, preferring hypothetical goals G as an explanation for O when they minimize the difference between $c(G,O)$ and $c(G,\bar{O})$, the latter being the cost of a plan for G which is constrained to be *inconsistent* with O, that is a plan which does not feature the observations in sequence. Assuming a Boltzmann distribution over plans according to their cost, and writing $exp\{x\}$ for e^x, likelihoods $P(O|G)$ are defined as:

$$P(O|G) = \frac{1}{1 + \exp\{-\beta \, \Delta(G)\}} \tag{6}$$

where $\Delta(G) = c(G,O) - c(G,\bar{O})$ and β is a positive constant. It is often noted that $\Delta(G)$ can be sometimes difficult to calculate when the observation sequence O contains actions which are *goal landmarks* (Hoffmann, Porteous, and Sebastia 2004). In that case $c(G,\bar{O})$ can increase substantially – or be ∞ in extreme cases – leading run–times to increase significantly. Also, while Ramirez & Geffner (2010) allows O to deviate from the plans selected by the observer as the *best plan* for G, inherent in the approach is the commitment to *one* specific best plan, amongst potentially many plans with the same associated cost. Various approaches have been proposed to avoid these issues while allowing for a softer stance on rationality (Martin, Moreno, and Smith 2015; Sohrabi, Riabov, and Udrea 2016; Pereira, Oren, and Meneguzzi 2017). In this work we will concentrate our efforts to study the most recent one as it results in run–times similar to library–based approaches, after an off-line and potentially expensive elicitation of landmarks, and has been shown empirically to provide very accurate approximations to RG10. Following from the discussion in Pereira et al (2017) of possible *heuristics* to select hypothetical goals as explanations for O, we define the likelihood of O given a hypothetical goal G in terms of landmarks:

$$P(O|G) \propto \frac{|O \cap \mathcal{L}(G)|}{|\{G' | \mathcal{L}(G') \cap \mathcal{L}(G) \neq \emptyset, G' \in \mathcal{G}\}|} \tag{7}$$

in words, directly proportional to the degree O *covers* the set of landmarks for goal G, $\mathcal{L}(G)$ and inversely proportional to the degree of overlap between sets of landmarks of hypothetical goals $\mathcal{G}$. We will refer to this last approach as POM17.

Planning Transparently

While Plan Recognition is usually described as a multi–agent setting, only recently (Zhang et al. 2017) models and frameworks have been proposed that address explicitly the possible interactions between the agent being observed, or *actor*, and the agent observing the actor, which we will refer to as the *observer*. In particular we drop the key assumption in so–called *keyhole* Plan Recognition (Schmidt, Sridharan, and Goodson 1978; Kautz and Allen 1986) where the actor does not take an interest on what the observer may do or believe about its actions. We will next propose *one possible* planning model for an actor that both *knows* about the observer and *wants* to influence it in some way. In order to do so, the actor may eventually have to follow a course of action that deviates from the most efficient plan for its *actual* goal $G^* \in \mathcal{G}$, or temporarily abandon its pursuit, to pursue instead changing the observer's initial beliefs to consider G^* *more likely*, while driving those *away* from other hypothetical goals $G' \in \mathcal{G}$.

We define a transparent planning task with the tuple $\Pi = \langle X, A, s_0, \Omega, q(\cdot|\cdot), \mathcal{G}, G^*, P(\cdot|\cdot), Prob \rangle$, where:

- X and A are a set of Functional STRIPS variables and actions describing a *planning domain*,
- s_0 is a given valuation of variables X,
- $\mathcal{G}$ is a set of possible goal formulae G over literals of X, with one distinguished member G^* which models the actor's goal,
- $\Omega \subseteq X$ and $q(\phi|\psi, a)$ allow us to specify a *sensor model* when the actor has limited access to the values of X, where $a \in A$, ϕ is a conjunction of literals of every variable in Ω, and ψ is a conjunctive formulae over X,
- $P(\cdot|\cdot)$ is the *definition* posterior goal distribution $P(G|O)$ which is *assumed* by the actor to be used by the observer to make sense of its behaviour,
- $Prob$ is a *prior* probability distribution over $\mathcal{G}$.

We will refer to pairings of distributions $P(\cdot|\cdot)$ and $Prob$ as the *observer stereotype* or *stereotype* for short. Last, Π shares elements both relevant to planning and goal recognition models discussed elsewhere in the paper, this makes apparent that the actor both reasons about his own goal G^*, and influences the observer's beliefs as it generates observation sequences O made of actions $a \in A$ chosen in a suitably defined manner.

Amongst the several possible models discussed by the planning community over the years, we chose that of Kaelbling's (1999) Goal POMDPs. The transparent planning task Π can be compiled into a Goal POMDP $M(\Pi)$ as follows. States S are valuations over variables $X' = X \cup \{y\}$, y is the non–observable state variable used to keep track of the goal the observer *knows* the actor to pursue. The domain of y is given by the set $D(y) = \{KG | G \in \mathcal{G}\}$, where KG are constant symbols that stand for "the observer *knows* $G = G^*$". Such a statement is only true when indeed, the observer has identified *precisely* the goal G^*, as it is false for every goal $G' \neq G^*$, $G' \in \mathcal{G}$. Furthermore, we note that states s such that $[G]^s = \top$ and $y \neq KG$ are valid. The reason for allowing the mental state of the observer to be independent of the actual truth value of the goals $\mathcal{G}$ in a given state, is to avoid the observer beliefs diverging from the distribution $P(\cdot|\cdot)$. Such a thing can occur if we forced $y = KG$ whenever $[G]^s$ is true, such as when there exists a goal $G^{'} \in \mathcal{G}$ which has G as a landmark. Introducing such a functional dependency between X and y can indeed be justified, but it is not of immediate interest to us.

The initial belief b_0 is defined as $b_0(s) = Prob(G)$ for every state $s = s_0 \wedge (y = KG)$ with $G \in \mathcal{G}$ and $b_0(s') = 0$ otherwise. Actions a do not change the value of y so the transition probability distribution follows from

$$tr(s|s', a) = \begin{cases} P(\{a\}|[y]^{s'})P(\sigma|a) & s = f(s', \sigma) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $f(s', \sigma)$ is a function which updates s' with the set of assignments σ, and $P(\sigma|a)$ is the probability of σ given the action a (Bonet and Geffner 2001).

A belief b is a goal belief if b satisfies constraints b_G:

$$\sum_{s \in S,\, [y]^s = G^*} b(s) \geq \frac{1}{|\mathcal{G}|} + \max_{G' \in \mathcal{G} \setminus G^*} \sum_{s \in S'\, [y]^s = G'} b(s) \tag{9}$$

We want to note that alternative definitions of b_G are possible. For instance, fairly intuitive ones such as that of G^* being part of the set

$$\operatorname*{argmax}_{G' \in \mathcal{G}} \sum_{s \in S,\, [y]^s = G'} b(s) \tag{10}$$

could be useful when $Prob$ is *not* a i.i.d. over $\mathcal{G}$, otherwise b_0 would trivially satisfy the constraint. We use equation 9 as it requires that the true goal be salient beyond a normalization factor.

The tracking of beliefs in Equations 1–3 can be simplified in the following way for the domains discussed in the Evaluation section. First, in these domains the agent has access to all values in X and may not directly or indirectly know the value of y, so our sensor model becomes $\Omega = \{\perp\}$, $Q_a(\omega|s) = 1$ for every s and a. As a result of this, Equation 1 and Equation 3 have one and the same expression, i.e. $b_a(s) = b_a^{\perp}$. Second, actions in A have *deterministic* effects so the belief update in Equation 1 becomes

$$b_a(s) = \sum_{s' \in S} tr(s|s', a)b(s') = \sum_{s' \in S} P(\{a\}|[y]^{s'})b(s') \tag{11}$$

The two simplifications above apply to fully observable, deterministic domain theories as in the case of the problems we have used in this paper. For this reason, instead of compiling Π to a Goal POMDP we can instead compile it to a *factored state model* (Frances et al. 2017) and use scalable width–based algorithms such as BFWS (Lipovetzky and Geffner 2017a) to efficiently search for sequences of actions to achieve a goal belief b_G.

The reasons for having chosen Kaelbling's classic framework over recently proposed ones such as Bonet's & Geffner (Bonet and Geffner 2016) that allows belief tracking factorization, are exposed next. First, we note that Goal POMDPs can be compiled into Bonet & Geffner framework without loss of generality. Second, and from a purely empirical standpoint, the models $M(\Pi)$ that result from the domains discussed in the Evaluation Section always have a *causal width* of 1 and hence, no computational advantage of using Bonet's & Geffner framework was to be reported, since time complexity of belief tracking is *linear* over $|\mathcal{G}|$. A recent framework we would have *wanted* to use but could not because it does not support probabilistic beliefs at the time of writing this, is the full–fledged multi–agent on–line planner recently proposed by Kominis & Geffner (2017). We look forward to reconcile this work with that of Kominis in the near future.

Action Selection

The action selection problem over the transparent planning task Π can be addressed as a *net-benefit planning problem* (Keyder and Geffner 2009) where the reward function is defined over beliefs b. The reward $R(b) = -d(b, b_T)$ considered in this work corresponds to the negative of the Euclidean distance between belief b and *target* belief state b_T, that satisfies the constraint b_G in Equation 9. We set b_T so that $P(G^*|O) = 1$, and $P(G|O) = 0$ for $G \in \mathcal{G} \setminus G^*$, which stands for the observer being *absolutely* certain of G^* being the goal the actor pursues. In order to compute the Euclidean distance d, the probabilities assigned by belief b to goals $G \in \mathcal{G}$ are treated as real–valued vectors of dimension $|\mathcal{G}|$.

A valid plan is a sequence of actions $\pi = \langle a_0 \ldots, a_n \rangle$ that induces the sequence of observations $O_\pi = \langle a_0, \ldots, a_n \rangle$ and the sequence of belief states $\langle b_0, \ldots, b_n \rangle$, such that a_0 is applicable in b_0, $b_i = b_{a_i}$ as per Equation 11, and $b_n \models b_G$. The utility of a plan π is defined as

$$u(\pi) = \frac{\sum_i^{|\pi|} R(b_i)}{|\pi|} \quad (12)$$

in words, the average accumulated reward, and a plan π is optimal if there is no other plan π' with higher utility. Note that plan length $|\pi|$ is only used to normalise $u(\pi)$, as we set the task to seek sequences π that maximise the *average* similarity between beliefs b_i and target belief b_T.

Features over Beliefs for Width–Based Search

Width-based algorithms have been shown to scale up for classical planning problems, factored state models, and net-benefit problems (Lipovetzky, Ramirez, and Geffner 2015; Lipovetzky and Geffner 2017a; Frances et al. 2017). In this work we use BFWS, a best-first search that balances exploration and exploitation. BFWS ranks nodes in the open list by novelty (exploration term), and breaks ties by a heuristic function (exploitation term). The novelty of a newly generated state $w(s)$ is the size of the smallest subset (conjunction) Q of atoms $[x]^s$ true in s and false in all the states s' generated before s, i.e. $w(s) = min_{Q \subseteq s, Q \not\subseteq s'} |Q|$ (Lipovetzky and Geffner 2012).

For the transparent planning task we extend the computation of novelty to a belief state by taking into account not only the state variables valuation $[x]^s$, but also the quantities $[Y_G]^s = \sum_{s \in S, [y]^s = G} b(s)$ corresponding with the posterior probability $P(G|O)$ assigned by the observer to each goal $G \in \mathcal{G}$. The observation sequence O corresponds to the partial plan leading to state s. Features Y_G take a precision of two decimal digits[1]. The features used to compute novelty are then $F(s) = \{ [x]^s \mid x \in X \} \cup \{ [Y_G]^s \}$, both the valuation of state variables $x \in X$ and the posteriors Y_G for each goal $G \in \mathcal{G}$. The novelty of a state becomes $w(s) = min_{Q \subseteq F(s), Q \not\subseteq F(s')} |Q|$. Extending the definition of novelty with additional features have already shown benefits in classical planning (Frances et al. 2017). Last, the exploitation term $h(s) = u(\pi)$ is set to be the utility of the partial plan π leading to s, making BFWS to prefer first novel states, breaking ties by utility.

In order to cope with the size of the state-space, and taking advantage that net-benefit problems can be solved by *online* algorithms, we adapt BFWS to solve the online action selection problem, where given a *budget* expressed in terms of time, number of generated states or simply as a search horizon on the maximum plan length, the algorithm has to return the first applicable action leading to the state with highest utility. For this setting we take advantage of a polynomial variant of BFWS where nodes whose novelty $w(s) > 1$ are pruned (Lipovetzky and Geffner 2017b). The benefit of the polynomial version known as 1-BFWS is that it does not require a budget, it simply stops when no more nodes can be expanded. Note that 1-BFWS can generate at most $\mathcal{D}_X \times \mathcal{D}_{Y_G}$ states, where $\mathcal{D}_X = \sum_{x \in X} |D(x)|$ is the sum of domains D cardinality for each state variable, and $\mathcal{D}_{Y_G} = \sum_{G \in \mathcal{G}} |D(Y_G)|$ is the sum of the cardinalities of state features' domains. As soon as 1-BFWS reaches a state s that satisfies the goal belief state b_G, then the search is stopped and the first action leading to s is returned. If no such state exists, then the search stops when no more nodes can be expanded and the action leading to the state with highest utility is returned.

We finish observing that the instance of BFWS above does not *necessarily* select actions (paths) which lead to beliefs b where G^* is true in every state s s.t. $b(s) > 0$, and therefore, guarantee that the true goal has been achieved. In practice, once G^* has been confirmed by the observer in some way, e.g. via some observation token which rules out any state where $y \neq KG^*$, one could switch to a purely goal–directed planner.

Approximating $\mathrm{P(O|G)}$ for Belief Tracking

The computational bottleneck to solve the transparent planning problem is the derivation of Δ values. We extend Ramirez & Geffner (2009) approximation for $C(G, O)$, based on Hoffmann's h_{FF} heuristic (2001), to approximate $C(G, \bar{O})$. While the value of h_{FF} on the tasks that result from compiling away O may be reasonably informative, as noted by Ramirez & Geffner (2010), it generally is much less informative for the planning task constrained to avoid plans consistent with O, as the constraints are satisfied already in the initial state.

The cost of achieving goal G without satisfying the observation sequence O, is the cost of a plan which either removes one or more observations, or changes their sequence order. Hence, in order to break the sequence of operators it is sufficient to *force* an observation to appear with a different index. The following transformation extends Ramirez & Geffner (2009) in order to achieve this.

Given an observation sequence O and observation $a_i \in O$, we construct the FSTRIPS planning problem $\Pi = \langle X', s, A', G \rangle$ required to evaluate $P(\cdot|\cdot)$ in Equation 11 as follows. The set of variables X' results from augmenting X with Boolean state variable $next$. A' is made up of copies a' of each $a \in A$, where $Pre(a') = Pre(a) \wedge next$ when $a = a_i$, $Pre(a') = Pre(a)$ otherwise. Effects of copies a' for actions $a \neq a_i$ are $Eff(a') = Eff(a) \cup \{next := \top\}$. Finally the initial state $s = f(s_0, O')$ is set to the state resulting from the application of the observation sequence $O' = \langle a_0, \ldots, a_{i-1} \rangle$ preceding observation a_i. The transformed planning problem ensures that the observation a_i is out of sequence as it is not applicable on the ith step of a plan, and importantly not even in the delete-relaxed plan that induces the value of h_{FF}. This is achieved by making sure that a different observation action that adds the variable $next$

[1]Features need to be bounded, otherwise their domain becomes non-finite.

is used first. This transformation is applied to every observation $a_i \in O$. The cost of an optimal plan in Π_i is approximated by h_{FF}. The cost $C(G, \bar{O})$ is then set to minimum cost for breaking the observation sequence O:

$$C(G, \bar{O}) = \min_{i=0,\dots,|O|} c(\pi_{i-1}) + h_{FF}(\Pi_i, I_i) \qquad (13)$$

where $c(\pi_{i-1})$ is the cost of the observation sequence leading to a_i. Note that for each search node we do not require computing h_{FF} $|O|$ times, $C(G, \bar{O})$ can be computed from the best $C(G, \bar{O})$ value leading to the parent node.

Evaluation

Domain	RG10	RG09	POM17	IGC T	LAMA T
BLOCKS WORD	0.565	0.813	1.028	0.344	0.193
GRID NAVIGATION	0.523	0.523	1.000	0.431	0.21
GRID NAVIGATION + OBS	0.649	0.649	1.000	0.433	0.21
TICKET TO RIDE	0.571	0.634	0.959	0.654	0.386
CAMPUS	0.854	0.828	1.344	0.437	0.353
INTRUSION-DETECTION	0.774	0.774	1.919	0.535	0.244
LOGISTICS	0.442	0.369	1.011	0.21	0.556
KITCHEN	0.778	0.444	1.756	0.237	0.19
ROVER	0.750	0.597	1.775	0.34	0.23

Table 1: Evaluation of IGC and LAMA. IGC T and LAMA T is average time in seconds to select an action. Columns RG10, RG09, POM17 are average ratios $Q_v(IGC)/Q_v(LAMA)$ for the three stereotypes considered.

In order to test the usefulness of the proposed framework, algorithms and approximations discussed in the previous Sections, we have implemented an on–line planner for transparent, IGC (implicit goal communication), using the LAPKT planning framework (Ramirez, Lipovetzky, and Muise 2015). This planner, sets $P(\cdot|\cdot)$ to the approximation discussed in the previous Section, fixing $\beta = 1$ in Equation 6. A general planner for Π trivially follows from attaching procedures to the planner which provide the denotation of FSTRIPS terms corresponding to the posterior goal probabilities (Frances et al. 2017).

It may be that a regular goal-directed planner communicates its goal efficiently in the course of its plan execution, in this case there would be very little utility in the proposed methods. To determine whether this is true we evaluate IGC against the satisficing classical planner LAMA (Richter and Westphal 2010) with which we project away the variable y. Action selection in LAMA is driven mainly by the goal–directed reachability and landmark–based heuristics used to determine *helpful actions/preferred operators*. Run–times, quantity T in the tables, are put forward for reference only, as action selection in LAMA requires the computation of a *full* plan for goal G^* and then to commit to its first action.

We have tested IGC and LAMA over 9 domains: GRID NAVIGATION with and without obstacles, TICKET TO RIDE, BLOCKS WORD, CAMPUS, INTRUSION-DETECTION, LOGISTICS, KITCHEN and ROVER. CAMPUS, INTRUSION-DETECTION and KITCHEN are domains which have been compiled from plan libraries (Ramirez 2012), whereas the BLOCKS WORD, LOGISTICS and ROVER are domains from the International Planning Competition discussed in the literature on plan recognition as planning (Ramirez and Geffner 2010; Pereira, Oren, and Meneguzzi 2017). The GRID NAVIGATION domain specifies a simple navigation task on a 2–dimensional, 4–connected grid. The TICKET TO RIDE domain is a simplified version of the popular board game of the same name[2]. The task requires the connection of nodes representing in–game locations. Connecting two nodes is only possible if enough game pieces of a given colour have been drawn. Actions in every domain have deterministic effects and unitary costs. Experiments were conducted on a i7-6700HQ CPU running at 2.60GHz and the physical memory available to the planners was limited to 8Gb of RAM.

Experiments *simulate* the effects of the actions selected by IGC and LAMA on the beliefs of the observer. For each instance, and domain, three *observer stereotypes* are tested, one for each of the three methods, RG09, RG10 and POM17, discussed in the Section "Probabilistic Plan Recognition" to define $P(O|G)$, along with uniform priors $Prob$. The difference between the landmark graph used by LAMA and POM17 observer is that POM17 uses *causal* landmarks and no disjunctive landmarks (Keyder, Richter, and Helmert 2010; Richter 2011). It is an open question to measure the sensitivity of the observed results to the choice of algorithm to approximate action landmark sets. The simulator progresses $M(\Pi)$ as actions are requested from the planner and executed, this process continues up to the point that the current belief b satisfies b_G in Equation 9, producing a sequence of actions π. We measure performance comparing the length of sequences π generated, we denote this value as $Q_v(X)$, where v is the *observer stereotype* (RG09, RG10, POM17) and X is a planner (IGC, LAMA).

Table 1 compares the average lengths of generated action sequences π between IGC and $LAMA$, and the results clearly show a stark difference in the performance of IGC when switching the cost–based stereotypes (RG09, RG10) for the landmark–based observer stereotype. For the former, IGC, accelerates the convergence of the observer beliefs towards those satisfying b_G. For these observer stereotypes IGC acts far more transparently and manages to convey the goal faster than the standard goal-directed $LAMA$. In the case of the POM17 stereotype in most domains the lengths are very similar, and in the specific case of ROVERS and the plan library compilations (CAMPUS, KITCHEN and INTRUSION DETECTION) are worse. The reason for this difference of performance can be traced directly back to the "loopy" behaviour demonstrated in Figure 2 which *delays* the selection of landmarks. The number of plans in the plan library compilations are quite small[3] and the set of (action) landmarks contain most if not every action used relevant to valid plans. In contrast, Richter and Westphal's use of the landmark heuristic (2010) naturally directs LAMA towards including action landmarks in plans. While LAMA may seem

[2]https://boardgamegeek.com/boardgame/9209/ticket-ride

[3]We refer the reader to the figures in pages 73–78 of M. Ramirez's thesis (2012), where a graphical representation of the plan libraries is discussed.

to have been a poor choice for a baseline, in fact it serves us to illustrate the potential of an implementation of IGC using likelihoods derived from landmarks.

Table 2 produces a ranking between IGC and $LAMA$ on the basis of how often either planner induce beliefs consistent with b_G faster than the other. Again, for the cost–based stereotypes, the proposed IGC planner clearly outperforms $LAMA$, sometimes by a wide margin (see entry for Logistics with RG10 stereotype). Interestingly, this Table also shows how the performance of IGC degrades as we move from RG10 to stereotypes which *strongly* penalize plans which do not match their definition of rationality – $P(G|O)$. This is made manifest by the results on the ROVERS domain, where we see RG09 and POM17 react in *opposite* ways to IGC. RG09 assigns null posteriors $P(G|O)$ as soon as O deviates from *the* (suboptimal) best plan, POM17 in contrast penalises IGC as it avoids executing landmark actions common to many of the hypothetical goals $\mathcal{G}$, yet instrumental to achieve G^*. Included in Table 2 is an analysis which uses a χ^2 test where the null hypothesis (H_0) is that the number of IGC wins is equal to the number of times IGC doesn't win. A result with an *, **, *** or **** indicates a rejection of H_0 with a significance level of $0.05, 0.01, 0.001$ or 0.0001 respectively. Results which are significant in the opposite direction of the hypothesis are denoted **ns**, these are considered not significant because they only suggest that IGC did not perform better than LAMA.

While the results clearly show that our proposed instantiation of IGC is overfitting RG09 and RG10, and more generally, cost–based goal recognisers, we note that the hypothesis that maximising the number of landmarks achieved by plans as a good strategy for selecting actions for cost–based recognisers, does not seem to follow from the results in Tables 1 and 2. LAMA does not convey the goal faster in most domains.

Finally, we note that assumptions on rationality from the actor and observer perspective are best tested over domains that present opportunities for divergences in terms of which action conveys more information. Such domains are essentially partially–observable domains like those discussed by Ramirez & Geffner (2011), yet to try our ideas on these, additional research is necessary to integrate algorithms for action selection over partially–observable planning models, such as $LW(1)$ (Bonet and Geffner 2014), with the width–based search methods discussed in this paper.

			RG10			RG09			POM17		
Domain	I	P	F	Δ_m	Δ_a	F	Δ_m	Δ_a	F	Δ_m	Δ_a
BLOCKS WORD	23	IGC	22^{****}	7	2.7	14	4	2.4	3^{ns}	2	1.3
		LAMA	–	–	–	3	3	1.7	4	2	1.2
CAMPUS	16	IGC	4	2	1.5	4	3	2.2	$–^{ns}$	–	–
		LAMA	–	–	–	–	–	–	5	2	1.8
GRID NAVIGATION	21	IGC	16^{*}	16	5.1	16^{*}	16	5.1	$–^{ns}$	–	–
		LAMA	–	–	–	–	–	–	–	–	–
GRID NAVIGA–TION (W. OBST)	24	IGC	15	7	3.0	15	7	3.0	$–^{ns}$	–	–
		LAMA	–	–	–	–	–	–	–	–	–
INTRUSION DETECTION	23	IGC	19^{***}	4	1.4	18^{**}	4	1.3	3^{ns}	3	3.0
		LAMA	–	–	–	–	–	–	17	8	3.8
KITCHEN	18	IGC	12	2	1.5	18^{****}	2	1.3	6	2	2.0
		LAMA	–	–	–	–	–	–	12	10	3.0
LOGISTICS	20	IGC	20^{****}	17	4.5	18^{***}	20	11.0	13	17	7.3
		LAMA	–	–	–	2	11	10.0	5	10	7.0
ROVER	18	IGC	13	1	1.0	14^{*}	9	5.2	2^{ns}	1	1.0
		LAMA	1	1	1.0	3	1	1.0	13	4	1.1
TICKET TO RIDE	21	IGC	20^{****}	11	3.0	19^{***}	11	2.9	3^{ns}	4	3.3
		LAMA	–	–	–	–	–	–	–	–	–

Table 2: Comparison between IGC and LAMA over stereotypes RG09, RG10 and POM17. Column F is the number of times $Q_v(IGC) < Q_v(LAMA)$ (and vice versa). Columns Δ_m and Δ_a report maximal and average difference $Q_v(X_2) - Q_v(X_1)$, where X_2 and X_1 are the slowest and fastest planner. – entries correspond to cases where IGC (or LAMA) were never quicker than the other. Number of ties follows from subtracting the sum of F values from I (# instances) for every domain and stereotype. *, **, *** and **** indicate the statistical significance of the result representing *p-values* less than 0.05, 0.01, 0.001 and 0.0001 respectively. **ns** denotes results which are significant in the opposite direction of the hypothesis, which are those where IGC did not perform better than LAMA.

Future Work

In this paper we have formalised the notion of planning for communicating goals to an observer, making some strong assumptions on the nature of the reasoning process used to make sense out of the actions in the plan. One obvious and appealing line of future work is to conduct investigations with human observers and see to what extent human cognitive processes fit the behaviours predicted by our computational model. This study would need to investigate whether this fit would vary based upon if the human knows that the agent is communicating with them or not (whether the human believes that it is in a keyhole recognition scenario or an intended recognition scenario). Such a study would verify the suitability of our framework to inform the design of systems where humans and robots need to co–operate directly or indirectly in the pursuit of a common goal.

Communicating goals is a limited form of communication between agents which does not require a previous agreement on a communication protocol to be used by all the parties involved. Another future line of research is to explore the possibilities opened up by this work to enable *implicit co–ordination* in multi–agent planning domains.

The experimental setting we use in this paper to evaluate our approach implicitly assumes that every action can be perceived by the observer. This may well be not the case in realistic, yet simple settings. Examples of these are domains where the environment in some form *prevents* the observer from sensing the actors' actions.

Last, planning to *obfuscate* the goal being pursued seems intuitively a matter of setting b_T in a suitable manner, yet finding settings with an entirely *passive* observer where doing so is meaningful seems challenging to us.

References

Alami, R.; Clodic, A.; Montreuil, V.; Sisbot, E. A.; and Chatila, R. 2006. Toward human-aware robot task planning. In *AAAI spring symposium*, 39–46.

Astrom, K. 1965. Optimal control of markov decision processes with incomplete state estimation. *J. Math. Anal. Appl.* 10:174–205.

Avrahami-Zilberbrand, D., and Kaminka, G. A. 2005. Fast and complete symbolic plan recognition. In *Proc. IJCAI*.

Bonet, B., and Geffner, H. 2000. Planning with incomplete information as heuristic search in belief space. In *Proc. of AIPS-2000*, 52–61. AAAI Press.

Bonet, B., and Geffner, H. 2001. Gpt: A tool for planning with uncertainty and partial information. In *Proc. IJCAI Workshop on Planning with Uncertainty and Partial Information*.

Bonet, B., and Geffner, H. 2014. Belief tracking for planning with sensing: Width, complexity and approximations. *JAIR* 50:923–970.

Bonet, B., and Geffner, H. 2016. Factored probabilistic belief tracking. In *Proc. IJCAI*.

Carberry, S. 2001. Techniques for plan recognition. *User Modeling and User-Adapted Interaction* 11(1-2):31–48.

Chakraborti, T.; Sreedharan, S.; Zhang, Y.; and Kambhampati, S. 2017. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *Proc. IJCAI*.

Cohen, P. R.; Perrault, C. R.; and Allen, J. F. 1981. Beyond question answering. In Lehnert, W., and Ringle, M., eds., *Strategies for Natural Language Processing*. LEA.

Francès, G., and Geffner, H. 2015. Modeling and computation in planning: Better heuristics from more expressive languages. In *Proc. ICAPS*, 70–78.

Frances, G.; Ramirez, M.; Lipovetzky, N.; and Geffner, H. 2017. Purely declarative action representations are overrated: Classical planning with simulators. In *Proc. IJCAI*.

Freedman, R. G., and Zilberstein, S. 2017. Integration of planning with recognition for responsive interaction using classical planners. In *AAAI*, 4581–4588.

Geffner, H., and Bonet, B. 2013. *A Concise Introduction to Models and Methods for Automated Planning*. Morgan & Claypool.

Geffner, H. 2000. Functional strips. In Minker, J., ed., *Logic-Based Artificial Intelligence*. Kluwer. 187–205.

Geib, C. W., and Goldman, R. P. 2009. A probabilistic plan recognition algorithm based on plan tree grammars. *Artificial Intelligence* 173(11):1101–1132.

Hoffmann, J., and Nebel, B. 2001. The FF planning system: Fast plan generation through heuristic search. *JAIR* 14:253–302.

Hoffmann, J.; Porteous, J.; and Sebastia, L. 2004. Ordered landmarks in planning. *JAIR*.

Kaelbling, L. P.; Littman, M.; and Cassandra, A. R. 1999. Planning and acting in partially observable stochastic domains. *Artificial Intelligence* 101:99–134.

Kautz, H., and Allen, J. F. 1986. Generalized plan recognition. In *Proc. AAAI*, 32–38.

Keren, S.; Gal, A.; and Karpas, E. 2014. Goal recognition design. In *Proc. ICAPS*.

Keyder, E., and Geffner, H. 2009. Soft goals can be compiled away. *JAIR* 36:547–556.

Keyder, E.; Richter, S.; and Helmert, M. 2010. Sound and complete landmarks for and/or graphs. In *Proc. of European Conference in Artificial Intelligence (ECAI)*, 335–340.

Kominis, F., and Geffner, H. 2017. Multiagent online planning with nested beliefs and dialogue. In *Proc. ICAPS*.

Lipovetzky, N., and Geffner, H. 2012. Width and serialization of classical planning problems. In *Proc. ECAI*.

Lipovetzky, N., and Geffner, H. 2017a. Best-first width search: Exploration and exploitation in classical planning. In *Proc. AAAI*.

Lipovetzky, N., and Geffner, H. 2017b. A polynomial planning algorithm that beats lama and ff. In *Proc. ICAPS*.

Lipovetzky, N.; Ramirez, M.; and Geffner, H. 2015. Classical planning with simulators: Results on the atari video games. In *Proc. IJCAI*.

Martin, Y.; Moreno, M. D.; and Smith, D. E. 2015. A fast goal recognition technique based on interaction estimates. In *Proc. IJCAI*.

Masters, P., and Sardina, S. 2017. Deceptive path-planning. In *Proc. IJCAI*.

Pentney, W.; Popescu, A.; Wang, S.; Kautz, H.; and Philipose, M. 2006. Sensor-based understanding of daily life via large-scale use of common sense. In *Proc. AAAI-06*.

Pereira, R. F.; Oren, N.; and Meneguzzi, F. 2017. Landmark-based heuristics for goal recognition. In *Proc. AAAI*.

Ramirez, M., and Geffner, H. 2009. Plan recognition as planning. In *Proc. IJCAI*, 1778–1783. AAAI Press.

Ramirez, M., and Geffner, H. 2010. Probabilistic plan recognition using off-the-shelf classical planners. In *Proc. AAAI*.

Ramirez, M., and Geffner, H. 2011. Goal recognition over POMDPs. In *Proc. IJCAI*.

Ramirez, M.; Lipovetzky, N.; and Muise, C. 2015. Lightweight Automated Planning ToolKiT. `http://lapkt.org/`. Accessed: 2016-12-11.

Ramirez, M. 2012. *Plan recognition as planning*. Universitat Pompeu Fabra.

Richter, S., and Westphal, M. 2010. The LAMA planner: Guiding cost-based anytime planning with landmarks. *JAIR* 39:127–177.

Richter, S. 2011. *Landmark-based heuristics and search control for automated planning*. Griffith University.

Schmidt, C.; Sridharan, N.; and Goodson, J. 1978. The plan recognition problem: an intersection of psychology and artificial intelligence. *Artificial Intelligence* 11(1:2):45–83.

Sohrabi, S.; Riabov, A. V.; and Udrea, O. 2016. Plan recognition as planning revisited. In *Proc. IJCAI*, 3258–3264.

Sondik, E. 1978. The optimal control of partially observable markov decision processes over the infinite horizon: discounted costs. *Operations Research* 26(2).

Stubbs, K.; Wettergreen, D.; and Hinds, P. H. 2007. Autonomy and common ground in human-robot interaction: A field study. *Intelligent Systems, IEEE* 22(2):42–50.

Warneken, F., and Tomasello, M. 2006. Altruistic helping in human infants and young chimpanzees. *Science* 311(5765):1301–1303.

Yang, Q. 2009. Activity Recognition: Linking low-level sensors to high-level intelligence. In *Proc. IJCAI-09*, 20–26.

Zhang, Y.; Sreedharan, S.; Kulkarni, A.; Chakraborti, T.; Zhuo, H. H.; and Kambhampati, S. 2017. Plan explicability and predictability for robot task planning. In *In Proc. ICRA*, 1313–1320. IEEE.