Explicability? Legibility? Predictability? Transparency? Privacy? Security? The Emerging Landscape of Interpretable Agent Behavior

Tathagata Chakraborti[1] · **Anagha Kulkarni**[2] · **Sarath Sreedharan**[2]
David E. Smith · **Subbarao Kambhampati**[2]

[1]IBM Research AI, Cambridge 02142 USA
[2]Arizona State University, Tempe 85281 USA

tchakra2@ibm.com, anaghak@asu.edu, ssreedh3@asu.edu, david.smith@psresearch.xyz, rao@asu.edu

Abstract

There has been significant interest of late in generating behavior of agents that is interpretable to the human (observer) in the loop. However, the work in this area has typically lacked coherence on the topic, with proposed solutions for "explicable", "legible", "predictable" and "transparent" planning with overlapping, and sometimes conflicting, semantics all aimed at some notion of understanding what intentions the observer will ascribe to an agent by observing its behavior. This is also true for the recent works on "security" and "privacy" of plans which are also trying to answer the same question, but from the opposite point of view – i.e. when the agent is trying to hide instead of reveal its intentions. This paper attempts to provide a workable taxonomy of relevant concepts in this exciting and emerging field of inquiry.

Introduction

There has been significant interest in the robotics and planning community lately in developing algorithms that can generate behavior of agents that is interpretable to the human (observer) in the loop. This notion of interpretability can be in terms of goals, plans or even rewards that the observer is able to ascribe to the agent based on observations of the latter. Interpretability remains a significant challenge in the design of human-aware AI agents, such as assistive agents, as emphasized in the *Roadmap for U.S. Robotics* (Christensen et al. 2009) – *"humans must be able to read and recognize agent activities in order to interpret the agent's understanding"*. However, the work in this area has typically lacked coherence on the topic from the community as a whole, even if not in the research agenda of different research groups (Chakraborti et al. 2017a; Dragan 2017; MacNally et al. 2018), per se. Indeed, a quick scan of the existing literature reveals algorithms for "explicable", "legible", "predictable" and "transparent" planning with overlapping, and sometimes conflicting, semantics. The same can be said of a parallel thread of work on the "deception", "privacy" and "security" of plans. This paper thus attempts to provide a workable taxonomy of relevant concepts that can hopefully provide some clarity and guidance to future researchers looking to work on the topic.

The rest of the paper is organized as follows: We first introduce a general framework for describing problems in the space of "plan interpretability" and outline how existing works have addressed different aspects of this problem in cooperative settings. We will then turn the tables and explore complementary manifestations in adversarial settings. Finally, we end with a discussion on aspects of the proposed framework that have not been explored in existing literature.

Model Differences with the Observer

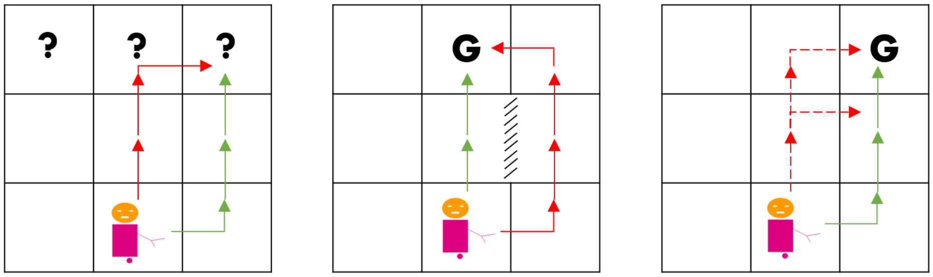
The key challenge in generating interpretable behavior is the ability to account for the model of the observer. This can be summarized as follows –

- An agent's actions may be uninterpretable when it does not conform to the expectations or predictions engendered by the observer model. Thus, the agent, to plan for interpretable behavior, must not only consider its own model but also the observer model and the differences thereof. (Chakraborti et al. 2017a; Dragan 2017)

This "model" can include the beliefs or state information of the agent, its goals and intentions, its capabilities or even its reward function. It can also include the observation model as well as the computational capability of the observer. A misunderstanding or mismatch on any of those accounts will mean that the plan or policy, as expected by the observer (given their cognitive capabilities), will not be the same as that computed by the agent, and will thus be difficult to interpret from the observer's point of view. We will outline in the rest of this writeup how existing work on the topic addresses one or more of these contributing factors, especially the goals and plans[1] ascribed to the agent by an observer.

Table 1 formalizes these considerations in the modeling of the agent A and the observer Θ in terms of –

[1] In this paper, we talk of behavior and plan in the same breath. In general, behavior can be seen as a particular instantiation of a plan or policy (which, in its general form, can have loops, contingencies, abstractions, etc.). However, most of the works surveyed here have used the term plan to refer to behavior. We will also stick to that convention – i.e. all the discussion here is confined to behaviors observed or ascribed to the agent by the observer.



(a) Legibility/Transparency. (b) Explicability. (c) Predictability.

Figure 1: A simple illustration of the differences between plan explicability, legibility and predictability. In this Gridworld, the agent can travel across cells, but cannot go backwards. Figure 1a illustrates a legible plan (green) in the presence of 3 possible goals of the agent, marked with **?**s. The red plan is not legible since all three goals are likely in its initial stages. In the parlance of *transparent planning*, the first action in the green plan can be part of a transparent plan (having conveyed the goal). Figure 1b illustrates an explicable plan (green) which goes straight to the goal **G** as we would expect. The red plan may be more favorable to the agent due to its internal constraints (the arm sticking out might hit the wall), but is inexplicable (i.e. sub-optimal) in the observer's model. Finally, Figure 1c illustrates a predictable plan (green) since there is only one possible plan after it performs the first action. In the parlance of *t-predictability*, this is a 1-predictable plan. The red plans fail to disambiguate among two possible completions of the plan. Note that all the plans shown in Figure 1c are explicable (optimal in the observer's model) but only one of them is predictable – i.e. explicable plans may not be predictable. Similarly, in Figure 1b, the red plan is predictable after the first action (even though not optimal, since there is only one likely completion) but not explicable – i.e. a predictable plan in the online setting may not be explicable in the offline setting. Similarly, in the offline case without the prefix (Figure 1b) the green plan is the only predictable plan and is also explicable.

- **Planning Problem** $\Pi = \langle \text{Domain Theory} = \mathcal{M}, \text{ Current State} = \mathcal{I}, \text{ Goal State} = \mathcal{G} \rangle$. The planning problem espoused by agent A is referred to as Π^A.
- **Plan** π is a solution to the planning problem Π. A behavior is one instantiation of a plan or policy – as we noted before, a plan refers to a behavior in this paper, unless otherwise mentioned. $\tilde{\pi}$ is a partial plan whose completion set is denoted by $\{\tilde{\pi}\}$.
- **Computational Model** $\chi \in \{S = \text{Sound}, SF = \text{Satisficing}, O = \text{Optimal}, C = \text{Complete}\}$ defines under what criterion an agent solves a planning problem. For example, if the observer has a complete computational model (C), they would find a solution if there was one. Note that most of these are features of the observer model that has been explored in existing literature and are certainly not meant to be exhaustive. Also, some of these possibilities are not disjoint.
- **Completion Function** $\delta(s, \pi, \chi) \mapsto \hat{s}$ captures whether a state $\hat{s}$ is reachable from the state s following a plan π subject to the computation model χ. For example, $\delta(\mathcal{I}, \pi, O) \mapsto \mathcal{G}$ implies π is an optimal solution to Π).
- **Observation Model** $\Omega : a \times s \mapsto o$ associates a token (o) with an action (a) and next state (s) pair. An observation sequence $\langle o \rangle$ produced by π is represented by $\langle o \rangle \models \pi$.

Interpretability? Plans versus Goals

An agent model (and the observer model) accounts for its beliefs, goals, capabilities and computation power. The computational model and completion function given above capture what plans the observer can understand, and the quality of those plans. This is key to questions of interpretability. Whether a plan is good or even sound from the point of view of the agent does not matter if the observer does not think so according to their computational model.

The exact nature of the interpretation task may vary. For example, we will see later that an optimal completion (O) in the observer model is an explicable plan while satisficing completions (SF) may be used to obfuscate. Many of the distinctions between different types of interpretability [2] are related to whether we are concerned with *goals* or *plans* (Dragan, Lee, and Srinivasa 2013).

Explicability We begin with "plan explicability" as introduced in (Chakraborti, Sreedharan, and Kambhampati 2018a; Zhang et al. 2016; 2017; Kulkarni et al. 2019).

> *Explicability measures how close a plan is to the expectation of the observer, given a goal / planning problem.*

Thus the objective of explicability is to be in the set of solutions to the observer's understanding of a planning problem. In Table 1, the explicable plan is one that has a completion in both the agent and the observer model. The first constraint requires that the solution solves the agent's planning problem while the latter requires that there exists a solution satisfying the emitted observations in the observer model – e.g. a plan looks optimal to the observer (Chakraborti, Sreedharan, and Kambhampati 2018a). When the observer model is not known (Zhang et al. 2017;

[2] Explicability, legibility and predictability of plans is a spectrum, i.e. one plan can have more "X"-ability than another. In the rest of the paper, unless otherwise stated, we refer to the end of that spectrum, whenever such a plan exists, (e.g. most explicable plan) when we mention an explicable, legible or predictable plan.

Concept	Setting / Agent Perspective		Formulation / Existing Literature
Explicability	Agent	$\Pi^A = \langle \mathcal{M}^A, \mathcal{I}^A, \mathcal{G}^A \rangle, \chi^A, \Omega$	Find: $\tilde{\pi}$
	Observer	$\Pi^\Theta = \langle \mathcal{M}^\Theta, \mathcal{I}^\Theta, \mathcal{G}^\Theta \rangle, \chi^\Theta$	Subject to: ${}_{\exists \pi \in \{\tilde{\pi}\}} \delta(\mathcal{I}^A, \pi, \chi^A) \models \mathcal{G}^A$
	Target	Solve Π^A with completion in Π^Θ	and ${}_{\exists \pi \in \{\tilde{\pi}\}, \langle o \rangle \models \tilde{\pi}} \delta(\mathcal{I}^\Theta, \pi, \chi^\Theta) \models \mathcal{G}^\Theta$ This involves finding the *expected* plan (prefix) that satisfies the observer model. Note that the $\exists$ can be switched to $\forall$ to model a more pessimistic observer model that requires all possible completions be explicable.
	(Zhang et al. 2017)		Π^Θ, χ^Θ unknown; $\mathcal{G}^A = \mathcal{G}^\Theta; \Omega : a \times s \mapsto a$ Here Π^Θ and χ^Θ are learned from human feedback in terms of a pre-defined labeling scheme.
	(Kulkarni et al. 2019)		Π^Θ, χ^Θ partially known; $\mathcal{G}^A = \mathcal{G}^\Theta; \Omega : a \times s \mapsto a$ Here Π^Θ and χ^Θ are learned from human feedback in terms of known plan distance measures $\Delta(\pi_1, \pi_2)$.
	(Chakraborti, Sreedharan, and Kambhampati 2018a)		$\Pi^A \neq \Pi^\Theta, \chi^\Theta = O, \Omega : a \times s \mapsto a$ This algorithm has the ability (via explanations) to deal with cases where $\nexists \pi : \delta(\mathcal{I}^\Theta, \pi, \chi^\Theta) \models \mathcal{G}^\Theta$, i.e. when explicable plans are not feasible in the agent model.
Predictability	Agent	$\Pi^A = \langle \mathcal{M}^A, \mathcal{I}^A, \mathcal{G}^A \rangle, \chi^A, \Omega$	Explicability +
	Observer	$\Pi^\Theta = \langle \mathcal{M}^\Theta, \mathcal{I}^\Theta, \mathcal{G}^\Theta \rangle, \chi^\Theta$	$\min \lVert \{\pi \mid \pi \in \{\tilde{\pi}\}, \langle o \rangle \models \tilde{\pi}, \delta(\mathcal{I}^\Theta, \pi, \chi^\Theta) \models \mathcal{G}^\Theta\} \rVert$
	Target	Solves Π^A with fewest completions in Π^Θ	This involves finding the most disambiguated or easy to predict plan (suffix) – all plan prefixes (even though belonging to explicable plans) are not easy to complete.
	(Dragan, Lee, and Srinivasa 2013)		Π^Θ implicit, $\chi^\Theta = O, \Omega : a \times s \mapsto a$ This deals with motion planning in continuous spaces where the mental model is often implicit – e.g. shortest path.
	(Fisac et al. 2018)		Π^Θ implicit, $\chi^\Theta = SF, \Omega : a \times s \mapsto o$ This explores predictability in discrete spaces but is still confined to motion / semi-task planning.
	(Kulkarni, Srivastava, and Kambhampati 2019)		$\Pi^A = \Pi^\Theta, \chi^\Theta = C, \Omega : a \times s \mapsto o$ This looks at *m-similar* solutions (instead of most predictable) with similarity d such that $\lVert \mathbb{S} \rVert \geq m$ and ${}_{\forall \pi_1, \pi_2 \in \mathbb{S}} \Delta(\pi_1, \pi_2) \leq d$, where $\mathbb{S} = \{\pi \mid \delta(\mathcal{I}^\Theta, \pi, \chi^\Theta) \models \mathcal{G}^\Theta\}$.
Legibility or Transparency	Agent	$\Pi^A = \langle \mathcal{M}^A, \mathcal{I}^A, \mathcal{G}^A \rangle, \chi^A, \Omega$	Find: $\tilde{\pi}$
	Observer	$\Pi^\Theta = \langle \mathcal{M}^\Theta, \mathcal{I}^\Theta, \{\mathcal{G}^\Theta\} \rangle \equiv \{\Pi_i^\Theta\}, \chi^\Theta$	Subject to: ${}_{\exists \pi \in \{\tilde{\pi}\}} \delta(\mathcal{I}^A, \pi, \chi^A) \models \mathcal{G}^A$
	Target	Solve Π^A and least number of Π_i^Θs	and $\min \lVert \{g \mid g \in \{\mathcal{G}^\Theta\} \wedge_{\exists \pi \in \{\tilde{\pi}\}, \langle o \rangle \models \tilde{\pi}} \delta(\mathcal{I}^\Theta, \pi, \chi^\Theta) \models g\} \rVert$ This involves finding plans that disambiguate possible goals – this is a property of the goal and not the plan. Interestingly, $\mathcal{G}^A$ may not be in $\{\mathcal{G}^\Theta\}$ as long as there is a mapping between them.
	(Dragan, Lee, and Srinivasa 2013)		Π^Θ implicit, $\chi^\Theta = O, \Omega : a \times s \mapsto a$
	(MacNally et al. 2018)		$\Pi^A = \Pi^\Theta, \chi^\Theta = O, \Omega : a \times s \mapsto a$
	(Kulkarni, Srivastava, and Kambhampati 2019)		$\Pi^A = \Pi^\Theta, \chi^\Theta = C, \Omega : a \times s \mapsto o$ Similar to *m-similarity*, this work looks for *j-legible* solutions in the offline sense such that $\lVert \{g \mid \delta(\mathcal{I}^\Theta, \pi, \chi^\Theta) \models g\} \rVert \leq j$

Table 1: A summary of concepts in the cooperative setting.

Kulkarni et al. 2019), as is most often the case, the completion in the observer model is difficult to guarantee. As such, *explicability is a spectrum*, where closer to completed plans in the observer model can be deemed to be more explicable. Note that this formulation does not require that the agent and observer have the same goal. As long as it looks like the agent's plan (prefix) achieves the observer's goal in the expected manner, the plan (prefix) is explicable.

Predictability Plan predictability, on the other hand, looks for non-ambiguous completions of a plan prefix (Dragan, Lee, and Srinivasa 2013; Fisac et al. 2018; Kulkarni, Srivastava, and Kambhampati 2019).

> *Plan predictability reduces ambiguity over possible plans, given a goal / planning problem.*

Table 1 highlights this distinction with the additional minimization term over the cardinality of the possible plan set (that satisfies the emitted observations) with completions in the observer model. This makes it clear that predictability is, again, a spectrum and –

> *An explicable plan can be unpredictable.*

An example would be when there are multiple explicable plans, i.e. many completions in the observer model, so that there is still work to be done in making sure that the observer can anticipate which plan it is that the agent is going to execute. If this can be achieved, then that specific plan would be both explicable and predictable. Consider the example in Figure 1c: if the user expects optimal plans ($\chi = O$), the two red plans are explicable because they are optimal. However, they are not predictable until after step two because either one is still possible before that. Similarly –

> *A predictable plan (in the online setting) can be inexplicable in the offline setting.*

This is possible when, given a prefix (during online plan execution), the observer can tell exactly what plan the agent is executing but the entire plan is still not one that s/he might expect it to (i.e. it does not follow the completion model of the observer). For example, the red plan in Figure 1b is completely predictable (since there is only one completion path), but it is not optimal, and therefore inexplicable if the user expects optimal plans ($\chi = O$). In (Fisac et al. 2018) the actions in the plan prefix of length t can be arbitrary and inexplicable as long as the postfix is predictable. This is also true for transparent (MacNally et al. 2018) plans as well. This phenomenon is readily seen in (Chakraborti et al. 2018) where the agent produces suboptimal plans that are easier to predict[3]. Figure 1 provides another example. More on this later in the discussion on online versus offline interactions.

Legibility So far we have discussed explicability and predictability of plans under the condition of known goals only. Plan legibility, in contrast, is defined as follows –

> *Plan legibility reduces ambiguity over possible goals that might be achieved.*

The observer model now includes a set of possible goals or equivalently a set of possible models parameterized by the goal, as shown in Table 1. In addition to solving the planning problem of the agent (first constraint), a legible solution requires that the set of observer models where a plan (satisfied by the observations) has completions is minimized.

The notion of legibility of goals has remained consistent across existing literature (Dragan and Srinivasa 2013; Dragan, Lee, and Srinivasa 2013; Kulkarni, Srivastava, and Kambhampati 2019) and is equivalent to the notion of *transparency* of plans (MacNally et al. 2018). To the best of our knowledge, plan explicability / predictability and legibility have not been considered together (i.e. with ambiguity over goals and plans simultaneously).

Interestingly, as Table 1 highlights, even though both predictability and explicability assume known goals, the goal known to the observer may not be the actual true goal of the agent and yet plans may be predictable or explicable. For example, the agent could really be doing something else but achieve the expected goal in the process, or the observer might think that their goal was being achieved due to limited observability or model differences. The ability to communicate enables Chakraborti, Sreedharan, and Kambhampati (2018a) to handle expectations under conditions of misunderstood goals as well. However, the notion of explicability remains identical as one of generating expected behavior. Similarly, for legibility, there needs to be only *some* mapping between the agent goal and the possible goal set which may not actually contain the real goal of the agent. In this sense, it is useful to recognize that all these behaviors model interpretable behavior given the entire planning problem and not specifically just the goal in the observer model.[4]

Online versus Offline Interactions

The actual setup of the interaction – i.e. online or offline – makes a big difference to the explicability versus predictability discussion. This is because explicability and predictability of a plan are *non-monotonic*, a plan prefix deemed inexplicable can become explicable with the execution of more actions and vice versa, either due to the observer being an imperfect planner due to computational limitations or due to implicit updates to the mental model based on the observations. The online case of explicability can then be seen in terms of the plan prefix – i.e. if its completion belongs to one of the explicable (completions in the observer model) or not. On the other hand, the offline case does not exist for plan predictability, which is a property of the plan suffix. However, in the online case, before the execution starts (i.e. with no prefix) a predictable plan has to be one of the explicable plans. With a prefix, that may no longer be the

[3] Fisac et al. (2018) use a fixed length of the plan prefix to generate predictable plan suffixes. In general, a planner can be allowed to determine this organically as done in (MacNally et al. 2018; Chakraborti et al. 2018).

[4] Though in the context of goal-directed behavior, from the perspective of the observer, these can end up being *perceived* as behavior given a goal rather than all components of the planning problem. In fact, recent works have explored how humans assign intentionality (de Graaf and Malle 2019) and unequal importance (Zahedi et al. 2019) to model artifacts that may be equivalent theoretically.

case, as discussed above (this is considering the definition of explicability in the existing work on the entire plan).

Note that, similar to predictability, legibility of plans is more useful in the online setting since it may be easy to deduce the real goal from the final state after completion of the plan. Though, even in such cases, when the goals (which are not usually fully specified) are not mutually exclusive, legible plans can help. Like explicability and predictability, legibility also shares the non-monotonicity property.

Motion versus Task Planning

One of the biggest points of difference in many of these works is in the nature of the target domain i.e. **motion planning** (Dragan and Srinivasa 2013; Dragan, Lee, and Srinivasa 2013; Dragan et al. 2015) versus **task planning** (Zhang et al. 2017; Kulkarni et al. 2019; Zakershahrak and Zhang 2018; MacNally et al. 2018). From the algorithmic perspective, this means continuous versus discrete state variables. However, the notion of plan interpretability engenders additional challenges. This is because a reasonable mental model for motion planning[5] can be assumed to be one that prefers shorter plans and thus need not be modeled explicitly (and thus does not need to be acquired or learned). For task planning in general, this is less straightforward. In fact, work on explicable task planning (Zhang et al. 2017; Kulkarni et al. 2019; Zakershahrak and Zhang 2018) has aimed to learn this implicit model using feedback from humans on the agent's behavior. A particular instance of this is when these model are assumed to be identical (MacNally et al. 2018; Kulkarni, Srivastava, and Kambhampati 2019) (this is usually the case in path planning, by default).

Given how humans can have vastly different expectations in the case of task planning, it is unclear how useful mental models learned from crowd feedback (as done in (Zhang et al. 2017; Kulkarni et al. 2019; Zakershahrak and Zhang 2018)) can be in the case of individual interactions.

Computational Capability

Interpretability is, of course, contingent on the computational capability of the observer, i.e. the completion function. There has been surprisingly little work on this. Fisac et al. (2018) approximated the human model with Boltzmann noisy rationality. Motion planning can permit the assumption of *"top-K"* rationality. However, for task planning (i.e. domains with combinatorial properties) the computational model of the observer is less clear – one can imagine something like depth or time bounded inferential capability that constrains the space of plans in the mental model. While almost all related work (Chakraborti, Sreedharan, and Kambhampati 2018a; MacNally et al. 2018) assumes a perfectly rational observer, models learned using feedback from human-subjects (Zhang et al. 2017; Kulkarni et al. 2019) are likely to implicitly model computational limitations.

Some recent works (Zhang and Zakershahrak 2019; Zakershahrak, Gong, and Zhang 2019) have started exploring these directions in this multi-model setting, especially from the point of view of explanation generation as a model reconciliation process (more on this later).

Discussion

Learning the Observation Model. The original work on explicability in task planning (Zhang et al. 2016; 2017) and subsequent works that build on it (Zakershahrak and Zhang 2018; Gong and Zhang 2018) attempt to learn the observer model when it is unknown. To the best of our knowledge, this is the only attempt to do so in the existing literature, in the context of plan explicability. They postulate that the explicability[6] can be measured in terms of whether the human observer is able to associate higher level semantics to actions in the plan. While this approach has its merits (e.g. in taking into account computational limitations), it also arguably conflates explicability with predictability. For example, just because an observer is able to assign task labels to individual actions in a plan does not necessarily mean they would have expected that plan.

Recent work (Choudhury et al. 2019) has highlighted the merits of being able to learn such models from data, even though explicit theories on the observer model can provide an early advantage in terms of accuracy and robustness due to the sample complexity of learning such models.

Observability. The concepts of explicability, predictability and legibility are intrinsically related to what is observable. In most of the existing work, the plan has been assumed to be completely observable. When this is not the case, the agent can try to ensure that unexpected actions are not observable and thus still be explicable. Interestingly most of the work in cooperative settings have worked with full observability while highlighting model differences. Later we will see that in the adversarial setting existing work mostly focuses on the observation model while assuming the rest of the agent's model is aligned with that of the observer.

Longitudinal effects. All of the work on the topic of interpretable behavior has, unfortunately, revolved around single, and one-off, interactions and little attention has been given to the impact of evolving expectations in longer term interactions. There is some reason to suspect that the need for explicable behavior will diminish as the observer becomes accustomed to the "quirks" of the agent. After all, to paraphrase George Bernard Shaw, *"the world conforms to the unreasonable man"!* This is, however, not a concern for legible and predictable behavior since, even with complete model alignment, the topic of coordination remains relevant.

Explanatory actions. In recent work (Sreedharan et al. 2018), authors have explored the notion of *"explanatory actions"* as actions that can have epistemic effects. These are actions that can affect the observer model. Plans that are made explicable with the use of explanatory actions are,

[5] While this is true for path planning in general, complex trajectory plans of manipulators with high degrees of freedom might still require modeling of observer expectations.

[6] Zhang et al. (2017) use "explicability" and "predictability" as measures towards achieving the same objective of producing plans closer to human expectation. This is somewhat confusing. The notion of predictability used there for the disambiguation of the plan suffix remains consistent.

of course, never predictable – i.e. one cannot *predict* that an explanatory action will occur during a behavior, but its presence can make the whole behavior explicable. Thus, in this view, in the set of explicable plans, not all plans are predictable. But, as we discussed before, all the predictable plans at the start of plan execution have to be explicable.

Human-agent Collaboration. Note that most of the discussion till now has assumed a passive observer. However, in most scenarios, the observer is likely to be a collaborator or, at the least, their behavior is going to be contingent on that of the agent. While explicability helps this cause, predictable behavior can negatively affect the observer when considered in isolation since such behavior, even though predictable, can leave the collaborator with little room to plan around. Indeed, human factors studies of plan predictability versus legibility (Dragan et al. 2015) are consistent with this concern, demonstrating that legibility is more desirable in a collaborative setting. Recent work (Zakershahrak and Zhang 2018) has started to take these considerations into account.

On preferences versus expectations. There is considerable prior art on incorporating human preferences in a robot's behavior, or plans in general. Indeed, the distinction between *preferences* and *expectations* is rather subtle. The former can be seen as constraints imposed on the plan generation process if the agent wants to contribute to the observer's utility – *"What would Jesus want me to do?"* – while the latter looks at how the agent can adapt its behavior in a manner that the observer would expect, as required by the observer's mental model of the agent – *"What would Jesus expect me to do?"*. As we mentioned before, in the case of motion planning, there is often no such distinction. Even in the case of task planning – for example, in "human-aware" planning where an agent decides not to vacuum while the elderly are asleep (Köckemann, Pecora, and Karlsson 2014) – sometimes it may be hard to identify where exactly the constraints lie, with preferences ("I don't want vacuuming while I am asleep") or expectations ("I don't expect the agent to be designed to vacuum at odd hours"). Ultimately this distinction might not make a difference algorithmically. The agent would need some process of performing multi-model argumentation (with its own model and the observer model) during its planning process (Chakraborti, Sreedharan, and Kambhampati 2018a).

Turning the Tables

So far we have talked about work that aims to reveal the intentions of the agent to an observer. The agent can also use the observer mental model and/or the observation model to hide its intentions. In the following, we compare and contrast recent work in the planning community in this direction. Many of the distinctions carry over from our discussion of plan explicability, predictability and legibility.

Goals versus Plans Similar to the previous discussion on predictability/explicability versus legibility, an agent can consider obfuscation of its goals and/or its plans. The goal obfuscation problem is the inverse of the legibility problem, while plan obfuscation is the inverse of the predictability problem discussed previously. Also similar to the previous discussion, it is easy to see that obfuscation of one (goal or plan) may not necessarily obfuscate the other. Unsurprisingly, they can be viewed under a unified framework, as explored recently in (Kulkarni, Srivastava, and Kambhampati 2019). Most of the existing work in this area has revolved around goal obfuscation (under the various names of privacy, deception and security) as outlined in Table 2. Interestingly, these ideas has evolved out of two parallel threads of research – one (Keren, Gal, and Karpas 2015; 2016; Masters and Sardina 2017b; MacNally et al. 2018) from the seminal work on *goal recognition design* (Keren, Gal, and Karpas 2014) and the other (Kulkarni, Srivastava, and Kambhampati 2019; Kulkarni et al. 2018) from the earlier work on plan explicability (Zhang et al. 2017; Chakraborti, Sreedharan, and Kambhampati 2018a; Kulkarni et al. 2019). The connections between these topics have hopefully become more apparent at this point.

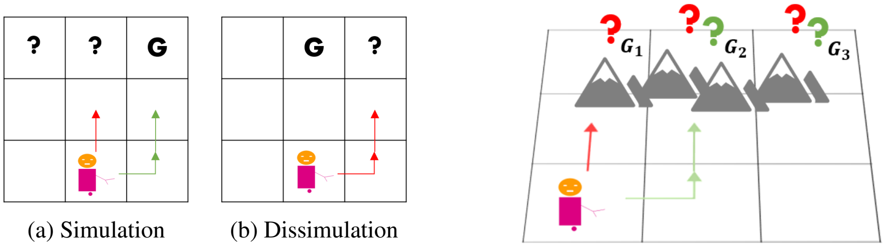
Motion versus Task Planning The distinction between motion and task planning again makes an appearance in the techniques used to approach these problems if not in the formulation of the concepts themselves. Particularly, (Masters and Sardina 2017b) arrive at a computationally efficient proxy for the likelihood of possible goals given a state and trajectory that is not necessarily available (Masters and Sardina 2017a) in the task planning setting. A similar computation used in (Kulkarni et al. 2018) only works for an incomplete computational model for the observer.

Online versus Offline The obfuscation problem is more appealing in the online setting since most of the motivation in obfuscating plans (such as in evading a pursuit or escaping surveillance) is lost after the plan is done. This is particularly the case for plan obfuscation with full observability, if not entirely true for goal obfuscation (c.f. discussion on predictability and legibility in offline settings). However, there is a rich set of problems to explore even in an offline setting (Kulkarni, Srivastava, and Kambhampati 2019; Kulkarni et al. 2018) once the observation model becomes noisy or incomplete. An offline analysis also allows for the determination of prefixes, if not entire plans, that can preserve privacy (Keren, Gal, and Karpas 2016). Many times, partially obfuscating solutions are all that is achievable.

The Observer and the Observation Model

One of the most important variables in this thread of work is the observation model. The notion of obfuscation and privacy immediately follows from settings with partial observability (Keren, Gal, and Karpas 2015; Kulkarni, Srivastava, and Kambhampati 2019; Kulkarni et al. 2018). However, as seen in (Masters and Sardina 2017b; Keren, Gal, and Karpas 2014), these concepts are still relevant to fully observable settings, especially in the online case.

The observer model also plays a critical role in this regard. Particularly, some of the surveyed works (Masters and Sardina 2017b; Keren, Gal, and Karpas 2016) have modeled a rational (more powerful) observer where $P(\mathcal{G} \mid \langle o \rangle)$ follows the distribution computed by a plan recognizer (i.e.



(a) Simulation (b) Dissimulation

Deception: Simulation / Dissimulation

(c) Obfuscation: Private / Secure Plans

Figure 2: A simple illustration of different goal obfuscation behaviors. Figures 2a and 2b show different forms of deceptive behavior (Masters and Sardina 2017b) in red – in simulation or "hiding the truth" (Figure 2a) the agent could be going to either of the three possible goals while in dissimulation or "showing the false" (Figure 2b) the likelihood of a decoy goal is strictly higher than that of the real goal. The green plan in (Figure 2a) is a truthful plan. Figure 2c illustrates the difference between privacy and security of goal obfuscating plans. Here the observer cannot observe the actions of the agent in the first row of the grid due to occlusions. The red and green plans are both 3-ambiguous (Kulkarni, Srivastava, and Kambhampati 2019) and privacy preserving (Keren, Gal, and Karpas 2016) – the former allows for $\{G_1, G_2, G_3\}$ while the latter allows for $\{G_2, G_3\}$ as possible goal sets for the agent assumed to be rational. However a secure algorithm cannot flip from the red to the green plan when rerun with G_2. This is allowed under privacy preserving (Keren, Gal, and Karpas 2016) and deceptive plans (Masters and Sardina 2017b) but not in secure plans (Kulkarni et al. 2018) – i.e. the red plan is the only secure 3-ambiguous solution.

$\chi^{\Theta} = O$), while others (Kulkarni, Srivastava, and Kambhampati 2019; Kulkarni et al. 2018) have considered an observer with a complete planner, i.e. $P(\mathcal{G} \mid \langle o \rangle) = 1$ if a plan exists that satisfies the observation, 0 otherwise (i.e. $\chi^{\Theta} = C$). Interestingly, in contrast to explicability or legibility and predictability, in the cases of goal or plan obfuscation, none of the surveyed works have scratched the surface of model differences between the agent and the observer beyond the observation and computation model.

The Many Flavors of Security, Privacy, Obfuscation, (Dis-)Simulation and Deception

The most interesting aspects of obfuscating behavior is the different notions of privacy, security and deception. We highlight the salient differences in Figure 2.

Obfuscation and Privacy These are synonymous in existing literature (Keren, Gal, and Karpas 2016; Kulkarni, Srivastava, and Kambhampati 2019) in trying to make multiple goals likely or possible given a set of observations. This is the reverse of legibility – notice that the minimizing term in Table 2 has flipped. Keren, Gal, and Karpas (2016) look at pair-wise likelihood of goals in maintaining privacy, while (Kulkarni, Srivastava, and Kambhampati 2019) provide a more general formulation for the computation of "k-ambiguous" solutions. This has implications on the security of the obfuscating algorithm, as we discuss below.

Security and Privacy The notion of security in plans (Kulkarni et al. 2018) builds on the concept of "independence of inputs" (Štolba 2017; Lindell 2005) which requires (in this specific context) that the planning algorithm return the same output or plan regardless of which obfuscated goal it is run with as the real goal so that an adversary cannot determine the real goal by rerunning the algorithm. Thus –

> *A secure plan is always private.*

This imposes an additional constraint to the privacy problem – as highlighted in Table 2, all possible goals must now lead to the same observation tokens.

Obfuscation and Deception Finally, in goal (or plan) obfuscation, the primary objective is to not reveal the true intentions, but not necessarily actively mislead. This distinction between simulation – "hiding the truth" – versus dissimulation – "showing the false" – was made in (Masters and Sardina 2017b). In the case of the latter, not only are multiple goals likely given a plan prefix but a decoy goal is also more likely than the real one. Deception, in general, can include both. It is clear from the discussion that –

> *A deceptive plan is always obfuscating, but may or may not be dissimulating.*

A more detailed discussion of this distinction can be found in (Masters and Sardina 2017b).

Discussion and Future Work

In the following discussion, we make connections to a parallel thread of work – "model reconciliation" – and outline possible directions for future work.

Communication and Model Reconciliation

Most of the discussion in this paper has revolved around communication of intentions (goals or plans) implicitly using behavioral cues. In general, predictable or legible behavior can be seen as a special case of implicit signaling behavior (Gong and Zhang 2018) when communication is undesirable. Foreshadowing certain actions (for example, through the medium of mixed reality (Chakraborti et al. 2018)) can considerably help the cause of predictability / legibility and coordination in human-agent interaction. The work on predictable (Fisac et al. 2018) or transparent (MacNally et al. 2018) plans could have similarly deployed speech, stigmergic or, in general, communication actions in the plan prefix.

Concept	Setting / Agent Perspective		Formulation / Existing Literature
Goal-Obfuscation Dissimulation or Privacy	Agent	$\Pi^A = \langle \mathcal{M}^A, \mathcal{I}^A, \mathcal{G}^A \rangle, \chi^A, \Omega$	Find: $\tilde{\pi}$
	Observer	$\Pi^\Theta = \langle \mathcal{M}^\Theta, \mathcal{I}^\Theta, \{\mathcal{G}^\Theta\} \rangle \equiv \{\Pi^\Theta_i\}, \chi^\Theta$	Subject to: $_{\exists \pi \in \{\tilde{\pi}\}} \delta(\mathcal{I}^A, \pi, \chi^A) \models \mathcal{G}^A$ and
	Target	Solve Π^A and as many Π^Θ_is	$\max \|\{g \mid g \in \{\mathcal{G}^\Theta\} \wedge _{\exists \pi \in \{\tilde{\pi}\}, \langle o \rangle \models \tilde{\pi}} \delta(\mathcal{I}^\Theta, \pi, \chi^\Theta) \models g\}\|$ This is the inverse of the legibility problem. A special case of **simulation** (Masters and Sardina 2017b) is when the real goal of the agent is explicitly hidden, i.e. $\delta(\mathcal{I}^\Theta, \pi, \chi^\Theta) \not\models \mathcal{G}^A$. **Deception** may or may not involve simulation.
		(Keren, Gal, and Karpas 2015)	$\Pi^A = \Pi^\Theta, \chi^\Theta = O, \Omega : a \times s \mapsto s$ This is a special case of *k-ambiguity* (Kulkarni, Srivastava, and Kambhampati 2019) when $k = 2$. Also, the solution is not secure as per (Kulkarni et al. 2018) – i.e. the real goal may not become the decoy, and thus reveal itself, if the algorithm is rerun with the decoy goal.
		(Keren, Gal, and Karpas 2016)	$\Pi^A = \Pi^\Theta, \chi^\Theta = O, \Omega : a \times s \mapsto o$ Same as above (but generalizes observation model).
		(Masters and Sardina 2017b)	$\Pi^A = \Pi^\Theta, \chi^\Theta = O, \Omega : a \times s \mapsto a$ *Last Deceptive Point* (LDP) introduced here (in the context of motion planning) has parallels to the notion of *equidistant states* in (Kulkarni et al. 2018). The latter deals with a general task planning setting. However, it deploys a heuristic which makes the planner incomplete.
		(Kulkarni, Srivastava, and Kambhampati 2019)	$\Pi^A = \Pi^\Theta, \chi^\Theta = C, \Omega : a \times s \mapsto o$ Instead of maximizing obfuscation, this work specifically looks for *k-ambiguous* solutions in the offline sense such that $\|\{g \mid \delta(\mathcal{I}^\Theta, \pi, \chi^\Theta) \models g\}\| \geq k$
Plan-Obfuscation	Agent	$\Pi^A = \langle \mathcal{M}^A, \mathcal{I}^A, \mathcal{G}^A \rangle, \chi^A, \Omega$	Find: $\tilde{\pi}$
	Observer	$\Pi^\Theta = \langle \mathcal{M}^\Theta, \mathcal{I}^\Theta, \mathcal{G}^\Theta \rangle, \chi^\Theta$	Subject to: $_{\exists \pi \in \{\tilde{\pi}\}} \delta(\mathcal{I}^A, \pi, \chi^A) \models \mathcal{G}^A$ and
	Target	Solve Π^A with most completions in Π^G	$\max \|\{\pi \mid \pi \in \{\tilde{\pi}\}, \langle o \rangle \models \tilde{\pi}, \delta(\mathcal{I}^\Theta, \pi, \chi^\Theta) \models \mathcal{G}^\Theta\}\|$ This is the inverse of the predictability problem.
		(Kulkarni, Srivastava, and Kambhampati 2019)	$\Pi^A = \Pi^\Theta, \chi^\Theta = C, \Omega : a \times s \mapsto o$ In addition to the cardinality of the solution set, this work looks for *l-diverse* solutions in the offline setting such that $\|\{\pi \mid \delta(\mathcal{I}^\Theta, \pi, \chi^\Theta) \models \mathcal{G}^\Theta\}\| \geq l$.
Security	Agent	$\Pi^A = \langle \mathcal{M}^A, \mathcal{I}^A, \mathcal{G}^A \rangle, \chi^A, \Omega$	Privacy + if $\langle o \rangle \models \tilde{\pi}$, then
	Observer	$\Pi^\Theta = \langle \mathcal{M}^\Theta, \mathcal{I}^\Theta, \{\mathcal{G}^\Theta\} \rangle \equiv \{\Pi^\Theta_i\}, \chi^\Theta$	$\forall g \in \{\mathcal{G}^\Theta\} : _{\exists \pi \in \{\tilde{\pi}\}, \langle o \rangle \models \tilde{\pi}} \delta(\mathcal{I}^\Theta, \pi, \chi^\Theta) \models g$
	Target	Find same solution for Π^A and as many Π^Θ_i	A privacy preserving planning algorithm is secure if and only if it emits the same observation sequence regardless of which goal it is run with.
		(Kulkarni et al. 2018)	$\Pi^A = \Pi^\Theta, \chi^A = \neg C, \chi^\Theta = C, \Omega : a \times s \mapsto o$ The approach in (Kulkarni, Srivastava, and Kambhampati 2019) can also do this with a slight modification, by generating observations that the agent wants to adhere to, with a random decoy goal.

Table 2: Summary of concepts in the adversarial setting.

Recent work (Sreedharan et al. 2018) has attempted a unified formulation through the use of explanatory actions.

During communication, the agent must be able to address the root cause of inexplicability, i.e. it must be able to explicate parts of the model that differ from the observer until they agree that its plan was, in fact, the best plan under the circumstances. This process of explanation, referred to as a process of *model reconciliation*, has been of significant interest (Chakraborti et al. 2017b; Sreedharan, Chakraborti, and Kambhampati 2018; Sreedharan, Srivastava, and Kambhampati 2018; Chakraborti et al. 2019) recently.

Particularly when the explicable plan is infeasible, such communication remains the only option for the agent to achieve common ground with the observer by, for example, expressing incapability (Raman et al. 2013; Raman and Kress-Gazit 2013; Briggs and Scheutz 2015; Kwon, Huang, and Dragan 2018), communicating misunderstandings about its capabilities (Chakraborti et al. 2017b; Chakraborti, Sreedharan, and Kambhampati 2018a) or even lying (Chakraborti and Kambhampati 2019) and augmenting new goals (Chen and Zhang 2018). The latter works are certainly more relevant from the perspective of the second part of the paper which explores obfuscation of intentions instead of revealing them. In fact, plan explicability and plan explanations form a delicate balancing act in "human-aware planning", as explored recently in (Chakraborti, Sreedharan, and Kambhampati 2018a). A concise survey of the model reconciliation process can be found in (Chakraborti, Sreedharan, and Kambhampati 2018b).

Design for Interpretability

A topic relevant to reasoning in the space of models of agents and observers is that of *goal recognition design* (GRD) (Keren, Gal, and Karpas 2014; Mirsky et al. 2019). Interestingly, even though we only brought it up in the context of goal obfuscation and privacy (Keren, Gal, and Karpas 2015; 2016), GRD is not particularly restricted to adversarial settings at all. The notion of environment design involves changing an environment to make behaviors more (or less in case of an adversarial setting) interpretable to the observer. This in contrast to changing the observer model in model reconciliation or the behavior of the agent itself as discussed in this paper. Though existing work in GRD has only looked at predictability and obfuscability issues, one could conceive of a general redesign framework that attempts to cover all the different flavors of interpretable behavior explored here. Interestingly, since environment redesign usually has more lasting effects than modifying the behavior of the agent (since they will be unable to perform certain actions or plans as a result rather than choosing not to), such a framework has to consider the longitudinal effects of changing an environment on the autonomy of the agent and on long-term interactions between the agent and the observer.

Further Generalizations

In Tables 1 and 2 we provided a general framework for describing the different aspects of the plan interpretability problem. The table also highlights gaps in the existing literature that can lead to exciting avenues of research in the future. The model considered in Tables 1 and 2, even though quite general in being able to classify the breadth of existing work on the topic, does not quite capture the full scope of plan interpretability. Below, we motivate a couple of generalizations to the framework presented in Tables 1 and 2. This was done intentionally so as not to overly generalize the overview which already captures the surveyed literature.

Observation Model with Epistemic Effects The observation model used in Tables 1 and 2 is quite general in being able to capture both partial as well as noisy sensor models. This model has been used extensively in the past (Geffner and Bonet 2013) as well as in many of the works covered in this survey; and provides a particularly elegant sensor model while formulating the planning problem for a single agent. However, when considering an observer in the loop, one should be cognizant of the effects of observations on the observer model – i.e. epistemic effects of actions. In recent work (Sreedharan et al. 2018) this has been explored in the context of implicit model updates on the part of the observer by means of "explanatory actions". One can conceive of a richer observation model that captures such epistemic effects of the agent actions on the observer.

Preference Measure on Plan or Goal Set The notion of legibility and obfuscation (Kulkarni, Srivastava, and Kambhampati 2019; Kulkarni et al. 2018; Masters and Sardina 2017b; Keren, Gal, and Karpas 2016) has largely considered the computation of a *set* of plans or goals as the desired consequence of a behavior, with additional preferences on the cardinality of that set in certain cases (e.g. predictability). Interestingly, in the solution for plan-legibility or predictability, Kulkarni, Srivastava, and Kambhampati (2019) look at "l-diverse" and "m-similar" solutions that can equally apply to the goal obfuscation and legibility cases as well. In general, the minimization or maximization term over the plan or goal sets in Tables 1 and 2 can be replaced by a function over the preferences of the observer towards the agent's achievement (execution) of any particular goal (plan) in the possible goal (plan) set, with cardinality being a special case of that function. More on this below.

An Active / Semi-Passive Observer The work surveyed here considers a passive observer. The full scope of the interpretability problem can include a more capable observer. This can be a semi-passive observer – i.e. one that can change the observation model only (in a sense reversal of the "sensor cloaking" problem explored in (Keren, Gal, and Karpas 2016)), for example, to improve observability by going to higher ground – to a fully active observer with their own goals and actions, with the ability to even assist or impede the agent from achieving its goals. This is likely to effect the relative importance of agent behaviors (e.g. is predictability more important than legibility in a collaborative setting? (Dragan et al. 2015)) and also effect the preference measure as discussed above (e.g. a surveillance scenario makes certain behaviors in the completions set more important to recognize, and hence to obfuscate, than others).

Unified Approach to Interpretable Behavior As we mentioned before, existing work has only looked at the different notions of interpretable behavior in isolation. Designing these behaviors is likely to become more challenging as we consider the effects of one or more of these behaviors simultaneously. For example, what would it mean to be explicable or predictable when there is ambiguity over the agent's goals? A legible plan given a goal might be an explicable plan for another goal. From our previous discussion regarding the fact that any of these behaviors can exist with or without the other, it will be interesting to see how they can exist simultaneously. Further, given that some of these behaviors are predicated on the notion of rationality on the agent model only (explicability) and others are not (legibility and predictability), it is unclear how the observer may be modeled once the belief of rationality has been suspended (for example, due to inexplicable but legible behavior).

Behaviors versus Plans Our discussion has mostly been confined to analysis of behaviors – i.e. one particular observed instantiation of a plan or policy. A plan – which can be seen as a set of constraints on behavior – engenders a *candidate set* of behaviors (Kambhampati, Ihrig, and Srivastava 1996) some of which may have certain interpretable properties while others may not. This means that an algorithm that can capture the "X"-ability of a plan can also do so for a particular behavior it models since in the worst case a behavior is also a plan that has a singular candidate completion. A general treatment of a plan can be very useful, for example, in decision-support where human decision-makers are deliberating over plans with the support of a planner (Sengupta et al. 2017). Unfortunately, interpretability of plans has received very little attention beyond explanation generation (Smith 2012; Fox, Long, and Magazzeni 2017; Borgo, Cashmore, and Magazzeni 2018).

Such a framework should not only be able to compute plans but also policies for communicating its information content during execution – there has been some recent work exploring the notion of disclosure policies and disclosed executions (Zhang, Shell, and O'Kane 2018a; 2018b).

Conclusion

In conclusion, we looked at a variety of interpretable behaviors of an agent that provide a rich set of directives to consider while designing agents that can account for the observer model in their decision making processes. We also saw how the ability to model and anticipate interpretability of its own behavior can be dual-use – i.e. the agent can use this to either reveal or obfuscate its intentions to the observer. We compared and contrasted existing literature that has tackled various aspects of this problem and provided a unified framework for precise specification of these (often confused) ideas. We also highlighted gaps in existing work and directions for future research. Finally, in this survey we have focused on the interpretability of behavior only, and the role of privacy and obfuscation in that context only. There is a rich body of work in the planning community that has explored these concepts in the context of *information sharing* in multi-agent planning (Brafman 2015; Štolba 2017) that can provide additional insights towards a more general of formulation of privacy preservation and obfuscation in a joint planning scenario.

Acknowledgments The majority of this work was done in Fall 2018 when all the authors were at Arizona State University. We thank the students of the ASU Human-Aware AI seminar class and members of PAIR community, including Sarah Keren, for their helpful comments in further refining some of the ideas in the paper. This research is supported in part by ONR grants N00014-16-1-2892, N00014-18-1-2442, and N00014-18-1-2840, AFOSR grant FA9550-18-1-0067 and NASA grant NNX17AD06G.

References

Borgo, R.; Cashmore, M.; and Magazzeni, D. 2018. Towards Providing Explanations for AI Planner Decisions. In *IJCAI XAI Workshop*.

Brafman, R. I. 2015. A Privacy Preserving Algorithm for Multi-Agent Planning and Search. In *IJCAI*.

Briggs, G., and Scheutz, M. 2015. "Sorry, I can't do that": Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions. In *AAAI Fall Symposium*.

Chakraborti, T., and Kambhampati, S. 2019. (When) Can AI Bots Lie? In *AIES*.

Chakraborti, T.; Kambhampati, S.; Scheutz, M.; and Zhang, Y. 2017a. AI Challenges in Human-Robot Cognitive Teaming. *arXiv:1707.04775*.

Chakraborti, T.; Sreedharan, S.; Zhang, Y.; and Kambhampati, S. 2017b. Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy. In *IJCAI*.

Chakraborti, T.; Sreedharan, S.; Kulkarni, A.; and Kambhampati, S. 2018. Projection-Aware Task Planning and Execution for Human-in-the-Loop Operation of Robots. In *IROS*.

Chakraborti, T.; Sreedharan, S.; Grover, S.; and Kambhampati, S. 2019. Plan Explanations as Model Reconciliation – An Empirical Study. In *HRI*.

Chakraborti, T.; Sreedharan, S.; and Kambhampati, S. 2018a. Balancing Explicability and Explanation in Human-Aware Planning. In *AAMAS*. Extended Abstract.

Chakraborti, T.; Sreedharan, S.; and Kambhampati, S. 2018b. Human-Aware Planning Revisited: A Tale of Three Models. In *IJCAI-ECAI XAI / ICAPS XAIP Workshops*.

Chen, Z., and Zhang, Y. 2018. Explain by Goal Augmentation: Explanation Generation as Inverse Planning. In *RSS Workshop on on Adversarial Robotics*.

Choudhury, R.; Swamy, G.; Hadfield-Menell, D.; and Dragan, A. 2019. On the Utility of Model Learning in HRI. *HRI*.

Christensen, H. I.; Batzinger, T.; Bekris, K.; Bohringer, K.; Bordogna, J.; Bradski, G.; Brock, O.; Burnstein, J.; Fuhlbrigge, T.; Eastman, R.; et al. 2009. A Roadmap for US Robotics: From Internet to Robotics. *Technical Report*.

de Graaf, M. M., and Malle, B. F. 2019. People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences. In *HRI*.

Dragan, A., and Srinivasa, S. 2013. Generating Legible Motion. In *RSS*.

Dragan, A. D.; Bauman, S.; Forlizzi, J.; and Srinivasa, S. S. 2015. Effects of Robot Motion on Human-Robot Collaboration. In *HRI*.

Dragan, A. D.; Lee, K. C.; and Srinivasa, S. S. 2013. Legibility and Predictability of Robot Motion. In *HRI*.

Dragan, A. D. 2017. Robot Planning with Mathematical Models of Human State and Action. *arXiv:1705.04226*.

Fisac, J. F.; Liu, C.; Hamrick, J. B.; Sastry, S. S.; Hedrick, J. K.; Griffiths, T. L.; and Dragan, A. D. 2018. Generating Plans that Predict Themselves. In *WAFR*.

Fox, M.; Long, D.; and Magazzeni, D. 2017. Explainable Planning. In *IJCAI XAI Workshop*.

Geffner, H., and Bonet, B. 2013. A Concise Introduction to Models and Methods for Automated Planning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*.

Gong, Z., and Zhang, Y. 2018. Behavior explanation as intention signaling in human-robot teaming. In *RO-MAN*.

Kambhampati, S.; Ihrig, L. H.; and Srivastava, B. 1996. A Candidate Set Based Analysis of Subgoal Interactions in Conjunctive Goal Planning. In *AIPS*.

Keren, S.; Gal, A.; and Karpas, E. 2014. Goal Recognition Design. In *ICAPS*.

Keren, S.; Gal, A.; and Karpas, E. 2015. Goal Recognition Design for Non-Optimal Agents. In *AAAI*.

Keren, S.; Gal, A.; and Karpas, E. 2016. Privacy Preserving Plans in Partially Observable Environments. In *IJCAI*.

Köckemann, U.; Pecora, F.; and Karlsson, L. 2014. Grandpa Hates Robots-Interaction Constraints for Planning in Inhabited Environments. In *AAAI*.

Kulkarni, A.; Klenk, M.; Rane, S.; and Soroush, H. 2018. Resource Bounded Secure Goal Obfuscation. In *AAAI Fall Symposium*.

Kulkarni, A.; Chakraborti, T.; Zha, Y.; Vadlamudi, S. G.; Zhang, Y.; and Kambhampati, S. 2019. Explicable Robot Planning as Minimizing Distance from Expected Behavior. In *AAMAS*. Extended Abstract.

Kulkarni, A.; Srivastava, S.; and Kambhampati, S. 2019. A Unified Framework for Planning in Adversarial and Cooperative Environments. In *AAAI*.

Kwon, M.; Huang, S.; and Dragan, A. 2018. Expressing Robot Incapability. In *HRI*.

Lindell, Y. 2005. Secure Multiparty Computation for Privacy Preserving Data Mining. *Encyclopedia of Data Warehousing and Mining*.

MacNally, A. M.; Lipovetzky, N.; Ramirez, M.; and Pearce, A. R. 2018. Action Selection for Transparent Planning. In *AAMAS*.

Masters, P., and Sardina, S. 2017a. Cost-based Goal Recognition for Path Planning. In *AAMAS*.

Masters, P., and Sardina, S. 2017b. Deceptive Path Planning. In *IJCAI*.

Mirsky, R.; Gal, K.; Stern, R.; and Kalech, M. 2019. Goal and plan recognition design for plan libraries. *TIST*.

Raman, V., and Kress-Gazit, H. 2013. Towards Minimal Explanations of Unsynthesizability for High-Level Robot Behaviors. In *IROS*.

Raman, V.; Lignos, C.; Finucane, C.; Lee, K. C.; Marcus, M.; and Kress-Gazit, H. 2013. Sorry Dave, I'm afraid I can't do that: Explaining Unachievable Robot Tasks Using Natural Language. In *RSS*.

Sengupta, S.; Chakraborti, T.; Sreedharan, S.; and Kambhampati, S. 2017. RADAR – A Proactive Decision Support System for Human-in-the-Loop Planning. In *AAAI Fall Symposium*.

Smith, D. E. 2012. Planning as an Iterative Process. In *AAAI*.

Sreedharan, S.; Chakraborti, T.; Muise, C.; and Kambhampati, S. 2018. Planning with Explanatory Actions: A Joint Approach to Plan Explicability and Explanations in Human-Aware Planning. *arXiv:1903.07269*.

Sreedharan, S.; Chakraborti, T.; and Kambhampati, S. 2018. Handling Model Uncertainty and Multiplicity in Explanations as Model Reconciliation. In *ICAPS*.

Sreedharan, S.; Srivastava, S.; and Kambhampati, S. 2018. Hierarchical Expertise-Level Modeling for User Specific Robot-Behavior Explanations. In *IJCAI*.

Štolba, M. 2017. Reveal or Hide: Information Sharing in Multi-Agent Planning. *Thesis*.

Zahedi, Z.; Olmo, A.; Chakraborti, T.; Sreedharan, S.; and Kambhampati, S. 2019. Towards Understanding User Preferences in Explanations as as Model Reconciliation. In *HRI*. Late Breaking Report.

Zakershahrak, M., and Zhang, Y. 2018. Interactive Plan Explicability in Human-Robot Teaming. In *RO-MAN*.

Zakershahrak, M.; Gong, Z.; and Zhang, Y. 2019. Online explanation generation for human-robot teaming. *arXiv:1903.06418*.

Zhang, Y., and Zakershahrak, M. 2019. Progressive explanation generation for human-robot teaming. *arXiv:1902.00604*.

Zhang, Y.; Sreedharan, S.; Kulkarni, A.; Chakraborti, T.; Zhuo, H. H.; and Kambhampati, S. 2016. Plan Explicability for Robot Task Planning. In *RSS Workshop on Planning for Human-Robot Interaction*.

Zhang, Y.; Sreedharan, S.; Kulkarni, A.; Chakraborti, T.; Zhuo, H. H.; and Kambhampati, S. 2017. Plan Explicability and Predictability for Robot Task Planning. In *ICRA*.

Zhang, Y.; Shell, D. A.; and O'Kane, J. M. 2018a. Finding Plans Subject to Stipulations on What Information They Divulge. In *WAFR*.

Zhang, Y.; Shell, D. A.; and O'Kane, J. M. 2018b. What Does My Knowing Your Plans Tell Me? In *IROS Workshop on Towards Intelligent Social Robots*.