A Unifying Bayesian Formulation of Measures of Interpretability in Human-AI Interaction

Sarath Sreedharan[1] , **Anagha Kulkarni**[1] , **David E. Smith** , **Subbarao Kambhampati**[1]
[1]Arizona State University
ssreedh3@asu.edu, anaghak@asu.edu, david.smith@psresearch.xyz, rao@asu.edu

Abstract

Existing approaches for generating human-aware agent behaviors have considered different measures of interpretability in isolation. Further, these measures have been studied under differing assumptions, thus precluding the possibility of designing a single framework that captures these measures under the same assumptions. In this paper, we present a unifying Bayesian framework that models a human observer's evolving beliefs about an agent and thereby define the problem of *Generalized Human-Aware Planning*. We will show that the definitions of interpretability measures like explicability, legibility and predictability from the prior literature fall out as special cases of our general framework. Through this framework, we also bring a previously ignored fact to light that the human-robot interactions are in effect open-world problems, with respect to the human's beliefs about the agent. The human may hold beliefs unknown to the agent and may also form new hypotheses about the agent when presented with novel or unexpected behaviors.

1 Introduction

A crucial aspect of the design of human-aware AI systems is the synthesis of interpretable behavior [Gunning and Aha, 2019; Langley *et al.*, 2017]. Existing works in this direction [Chakraborti *et al.*, 2019a] explore behaviors that instigate a desired change in the human's mental state or conform with her current mental state so as to not require explicit communication. Three distinct notions of interpretability can be seen in prior work: *legibility* – the agent signaling its objectives through behavior (c.f. [Dragan *et al.*, 2013; Dragan, 2017; Kulkarni *et al.*, 2019a; 2019b; MacNally *et al.*, 2018; Dragan and Srinivasa, 2013; Miura and Zilberstein, 2020]); *explicability* – agent behavior that conforms with the human's expectation (c.f. [Zhang *et al.*, 2017; Kulkarni *et al.*, 2019c; 2020; Chakraborti *et al.*, 2019b]); and *predictability* – agent behavior that is easier to anticipate (c.f. [Fisac *et al.*, 2020; Dragan, 2017; Dragan *et al.*, 2015; 2013]). These notions of interpretability can each improve human-AI collaborations along different dimensions. If you know your agent's objectives (legibility) and can anticipate its future behavior (predictability), you can plan around it or even exploit it; while in conforming to your expectations (explicability), the agent can avoid surprising you, which would adversely affect the fluency of collaboration.

In our earlier work, Chakraborti *et al.* [2019a], we focused on providing an overarching taxonomy and establishing equivalences between works done in this space. On the other hand, in this paper, we will introduce a single unifying reasoning framework that can account for all these measures. Two major roadblocks to such a unification are: 1) the measures are defined under competing assumptions and 2) different frameworks are used to reason about the human's beliefs (which is central to defining these measures). In this work, we present a single Bayesian framework that captures the human's reasoning over a distribution of models she ascribes to the agent. We use it to define the *Generalized Human-Aware Planning* problem. We will show how measures studied in prior literature can be seen as special cases of our unifying framework. Furthermore, our Bayesian formulation of explicability reveals an important dimension of the human-aware planning problem that to the best of our knowledge has not been explicitly studied before – relationship between explicability and open-world beliefs. That is, by identifying inexplicable behavior, the human identifies that her belief about the agent was incorrect and that the agent's behavior is stemming from an unknown model. Our unifying framework accommodates this by considering an additional hypothesis that the human's belief about the agent may be wrong. The summary of our contributions is as follows:

1. We formulate a Bayesian framework to capture a human observer's reasoning about the agent in terms of a distribution over models and within it define the Generalized Human-Aware Planning Problem.

2. We show that this single unifying framework can be specialized to existing interpretability measures under the original assumptions made by those works.

    - The ability to model an "unknown" model is critical to the unification of these competing measures.

    - The unification further generalizes these measures.

2 Background

In this paper, we will be agnostic to specific planning formulations or representations when discussing the agent's model. Instead, we will use the term "model" in a general sense to not only include information about agent actions and transition functions, but also their reward/cost function, goals and initial state. We will assume that the model can be parameterized and use $\theta_i(\mathcal{M})$ to characterize the value of a parameter θ_i for the model $\mathcal{M}$.

Since we are interested in cases where a human is observing an agent acting in the world, we will mainly focus on agent behavioral traces (instead of plans or policies). A behavior trace τ in this context will consist of a sequence of state, action pairs. The likelihood of the sequence given a model will take the form $P_\ell : \mathbf{M} \times \mathcal{T} \rightarrow [0, 1]$, where $\mathbf{M}$ is the space of possible models and $\mathcal{T}$ is the set of behavioral traces that the agent can generate. While we will try to be agnostic to likelihood functions, a fairly common approach [Fisac *et al.*, 2018; Baker *et al.*, 2007] is a noisy rational model based on the Boltzmann distribution: $P_\ell(M, \tau) \propto e^{-\beta \times C(\tau)}$. Where $C(\tau)$ is the cost of the behavior and $\beta \in \mathbb{R}^+$ is a parameter that reflects level of perceived determinism in the agent's choice of plans [Baker *et al.*, 2009]. Note that in our case, a likelihood function captures both the human's expectations about the agent's computational capabilities and their own cognitive limitations. Thus noisy-rational models like the one mentioned above are particularly useful in our scenario. For example, by setting a low β value we could possibly capture the fact that the observer may not be able to correctly differentiate between strategies of relatively similar costs.

For the human-aware scenario, we are dealing with two different models [Dragan, 2017; Chakraborti, 2018; Reddy *et al.*, 2018]: the model that is driving the agent behavior (denoted $\mathcal{M}^R$) and the human's belief $\mathcal{M}_h^R$ about it. We make no assumptions about whether these two models are represented using equivalent representational schemes or use the same likelihood functions. *This setup assumes that while the human may have expectations about the agent's model, she may have no expectation about its ability to model her. Thus she isn't actively expecting the agent to mold its behavior to what she thinks the agent knows about her, thereby avoiding additional nesting of beliefs.*

Running Example

In our running example, we will consider a robotic office assistant (Figure 1), that can perform various repetitive tasks in the office, including picking up and delivering various objects to employees, emptying trash cans, and so on. Further, we will assume it can only move in three directions: down, left and right; and that it can not revisit a cell. These restrictions allow us to control the set of possible completions of a given plan prefix. You, as the floor manager, are tasked with observing the agent and making sure it is working properly. Given your previous experience, you have come to form expectations about its capabilities and its tasks: e.g. you may think that the goal of the agent is to either deliver coffee or to deliver mail to a room (represented by the door), though you know that there may be other possible goals that you have not considered. Unbeknownst to you, the agent is trying to deliver coffee and it needs to do this while keeping in mind your beliefs about it. This scenario is particularly designed to accommodate the considerations made by prior works on interpretable behaviors. Throughout this paper, we will revisit this example to show different behaviors.

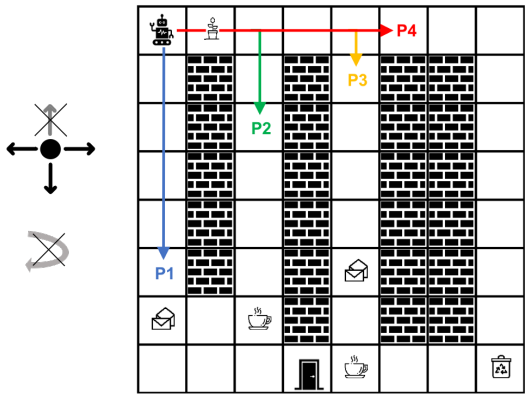

Figure 1: An illustration to show different interpretable behaviors. Here the agent only moves in three directions: down, left and right; and it does not revisit a cell.

3 A Unified Framework

The ability to anticipate and shape a human's beliefs about the agent, is a central requirement for any successful human-aware agent. So we start with a framework to capture the human's reasoning about their beliefs about the agent. In particular, we will adopt a Bayesian model of the human's reasoning process (Figure 2). This is motivated by both the popularity of such models in previous works in observer modeling and existing evidence to suggest that people do engage in Bayesian reasoning [L Griffiths *et al.*, 2008]. The node $\mathbb{M}_h^R$ represents possible models the human thinks the agent can have, τ_{pre} corresponds to the behavior prefix that they observed (in this paper we will assume full observability), and τ_{post} corresponds to possible completions of the prefix.[1]

In addition to explicit models that the human thinks are possible for the agent, we also allow for the possibility that the human may realize that she in fact doesn't know the exact agent model. That is, her previously held beliefs about the agent may not be sufficient to explain or justify the observed behavior. We incorporate this assumption by adding a special model $\mathcal{M}^0$ to the set of models in $\mathbb{M}_h^R$, that corresponds to the hypothesis that the agent model is not one of the models that the human expects.This allows for open-world reasoning since the human can form additional hypotheses about the robot and is not limited by the explicit set she originally has. This strategy of introducing a specific hypothesis that corresponds to a previously unexpected entity has been commonly used to model scenarios where there is a possibility of a novel or previously unknown event happening (c.f. [Zabell, 1992]). We represent $\mathcal{M}^0$ using a high entropy model: i.e. the like-

[1] In this paper, we focus on quantifying these measures for one shot or episodic interactions only, rather than longitudinal ones. In Section 4, we discuss more about longitudinal interactions.

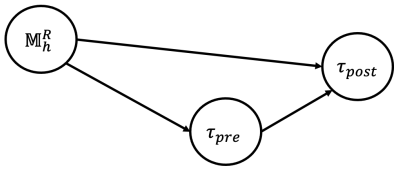

Figure 2: Graphical representation of the human's model.

lihood function of this model assigns a small but equal likelihood to any of the possible behaviors, including the ones facilitated by other models. This can be viewed as a model belonging to a random agent. We assume that the human, by default, assigns smaller priors to $\mathcal{M}^0$ than other models. We can now define the following problem:

Definition 3.1. A *Generalized Human-Aware Planning Problem* (G-HAP) is a tuple $\Pi_{\mathcal{H}} = \langle \mathcal{M}^R, \mathbb{M}_h^R, P_h^0, P_\ell, C_{\mathcal{H}} \rangle$, where P_h^0 is the human's initial prior over the models in the hypothesis set $\mathbb{M}_h^R$ and $C_{\mathcal{H}}$ is a generalized cost function that depends on the exact objective of the agent.

A solution to G-HAP consists of a behavior that is valid in $\mathcal{M}^R$ and minimizes $C_{\mathcal{H}}$. In the most general setting, $C_{\mathcal{H}}$ would be a mapping from entire behavior to a cost. Though internally $C_{\mathcal{H}}$ may be a function that takes into account each of the intermediate steps (not just in $\mathcal{M}^R$ but also the other models in $\mathbb{M}_h^R$). While the exact form would depend on the specific agent objectives, in general the cost function may need to consider (1) costs of the action in the sequence in $\mathcal{M}^R$ and their counterparts in each of the models in $\mathbb{M}_h^R$ (2) the state induced by the action in each model (3) possible completions at each intermediate step and their relation to the actual behavior and (4) the beliefs over $\mathbb{M}_h^R$ it may induce. Rather than investigate the space of all possible cost functions, we will ground the discussion by focusing on scenarios and objectives previously studied in the literature. We will see how this specialization of the framework, naturally gives rise to the specific interpretability measures. Throughout the discussion we will use the notations τ_{pre}^i and τ_{post}^i for a complete behavior τ to represent the behavior prefix that would have been observed and the behavior postfix remaining to be executed for a timestep i respectively. The overall framework presented in the paper is summarized in Figure 3. It illustrates that a human that could hold multiple hypotheses about the agent and show how the various existing measures could be extended to this more general setting. Explicability, in this case, becomes the human's confidence that they can explain the robot behavior with one of the explicit hypothesis they have regarding the robot, while legibility maps to their specific confidence that the robot model actually includes some parameter (which is, in fact, present) and predictability turns into a measure of confidence they assign to the actual future behavior the robot is going to generate. Below we will look at each of these individual measures in more detail and see how they arise from G-HAP.

3.1 Explicability

We will start by looking at cases where the agent wants to avoid behaviors that may confuse the observer about the agent

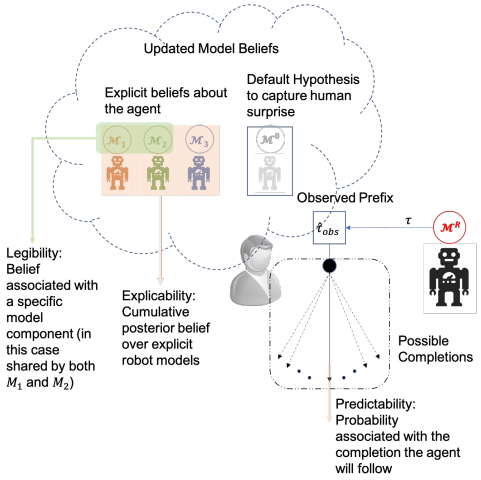

Figure 3: An overview of our unifying framework. The human holds multiple hypotheses about the agent and she uses the observed behavioral prefix to update her beliefs about the model. Each of the interpretability measure optimizes for specific inferential outcomes in this framework.

model. That is the human should be able to explain the observed behavior with the explicit models they hold. We will refer to such behaviors as *explicable behaviors*. We can capture the generation of such behavior within our framework by using a cost function that is proportional to the posterior probability associated with model $\mathcal{M}^0$, i.e.,

$$C_{\mathcal{H}}(\tau) \propto \sum_i \alpha_i P(\mathcal{M}^0 | \tau_{pre}^i) \tag{1}$$

Where $\alpha_i \geq 0$ is the weight associated with each timestep i. This means the formulation would prefer behavior with high likelihood in the explicit models for timesteps with non-zero weight. We will define the explicability score ($\mathcal{E}$) associated with a behavior prefix (τ_{pre}) to be directly proportional to one minus this probability, i.e.,

$$\mathcal{E}(\tau_{pre}^i) \propto \sum_{\mathcal{M} \in \mathbb{M}_h^R \setminus \{\mathcal{M}^0\}} P(\mathcal{M} | \tau_{pre}^i) \tag{2}$$

Likelihood functions that assign high probabilities to optimal (or low cost traces), give rise to traces like P1 and P2 in Figure 1, since they correspond to optimal plans in the explicit models considered in the example (i.e. the model for delivering coffee or delivering mail).

Reduction to Previous Explicability Definitions: Previous works generally identify a behavior to be explicable if it meets the human's expectation from the agent for the given task [Zhang *et al.*, 2017]. In the binary form this is usually taken to mean that a plan is explicable if it is one of the plans that the human expects from the agent [Chakraborti *et al.*, 2019b]. In the more general continuous form, this expectation is taken to be proportional to the distance between the observed trace and the closest expected behavior [Kulkarni *et al.*, 2019c; Zhang *et al.*, 2017]:

$$\tau_{\mathcal{E}}^* = argmin_\tau \delta(\tau, \tau_{\mathcal{M}_h^R}^E) \tag{3}$$

where δ is some distance function between two plans and $\tau^E_{\mathcal{M}^R_h}$ is the closest expected behavior for the model $\mathcal{M}^R$. While there is no consensus on the distance function or expected behavior, a reasonable possibility for the expected set is the set of optimal plans [Chakraborti *et al.*, 2017] and the distance can be the cost difference [Kulkarni *et al.*, 2020].

To see how our framework subsumes earlier works, lets start by plugging in the two assumptions made by the original works, namely (1) the human only has one explicit model about the agent (i.e. $\mathbb{M}^R_h = \{\mathcal{M}^R_h, \mathcal{M}^0\}$) and (2) the explicability is measured over the entire plan (i.e. $\alpha_i = 0$ for all i other than the last step). Thus the cost function is dependent only on the explicability of the entire behavior

$$\mathcal{E}(\tau) \propto P(\mathcal{M}^R_h|\tau) \propto P(\tau|\mathcal{M}^R_h) * P(\mathcal{M}^R_h) \tag{4}$$

Since the observed prefix is the entire plan, we can directly use the likelihood function

$$\mathcal{E}(\tau) \propto P_\ell(\mathcal{M}^R_h, \tau) * P(\mathcal{M}^R_h) \tag{5}$$

Let us consider two plausible likelihood models. First, for a normative model where the agent is expected to be optimal, $P_\ell(\mathcal{M}^R_h, \tau_{pre})$ assigns high but equal probability to all the optimal plans and 0 probabilities for the others. This is the original binary explicability formulation used by [Chakraborti *et al.*, 2019b; 2019a].

Another possible likelihood function is a noisy rational model [Fisac *et al.*, 2020] given by:

$$P_\ell(\mathcal{M}^R_h, \tau) \propto e^{-\beta \times C(\tau)} \propto e^{\beta \times C(\tau^*) - C(\tau)} \tag{6}$$

where τ^* is an optimal behavior in $\mathcal{M}^R_h$, $C(\tau) \geq C(\tau^*) \geq 0$ for $\mathcal{M}^R_h$. This maps the formulation to the distance based definition as in [Kulkarni *et al.*, 2020] where a cost-based distance is defined. We can also recover the earlier normative model by setting $\beta \to \infty$ and model $\mathcal{M}^0$ by setting $\beta = 0$.

Going back to the original motivation of explicability, it was meant to capture the human's understanding of the agent's behavior generation process (which includes both its perceived model and computational component). Earlier formulations rely on using the space of expected plans as a proxy of this process. This is further supported by the fact that the works that have looked at updating the human's perceived explicability value of a plan do so by providing information about the model and not by directly telling the human what plans to expect [Chakraborti *et al.*, 2017; 2019b; Sreedharan *et al.*, 2020a]. Thus our formulation of explicability directly in terms of the human's beliefs about the agent's model connects to the original motivation of explicability definitions.

Novel Properties of Generalized Explicability: An interesting side-effect of a probability-based explicability formulation is that, the probability of behavior and hence the explicability score can now be affected by the presence or absence of other plans. For example, consider two scenarios, one where $\mathbb{M}^R_h$ contains $\mathcal{M}_1$ and $\mathcal{M}^0$ and another where it contains $\mathcal{M}_2$ and $\mathcal{M}^0$. Now consider a behavior trace τ that is equidistant from an optimal plan in both models $\mathcal{M}_1$ and $\mathcal{M}_2$. Even though they are at the same distance, the trace may be more explicable in the first scenario than in the second, if the second scenario allows for more traces that are closer. Assuming the probability of choosing optimal plans isn't reduced, introducing new plans into the sample space better than the current trace would cause more probability to be assigned to those and thus less to the trace in question. We argue that this makes intuitive sense for explicability since the user should be more surprised in the second scenario as the agent would have ignored many more behaviors that the observer would have considered desirable. Figure 4 illustrates such an example.

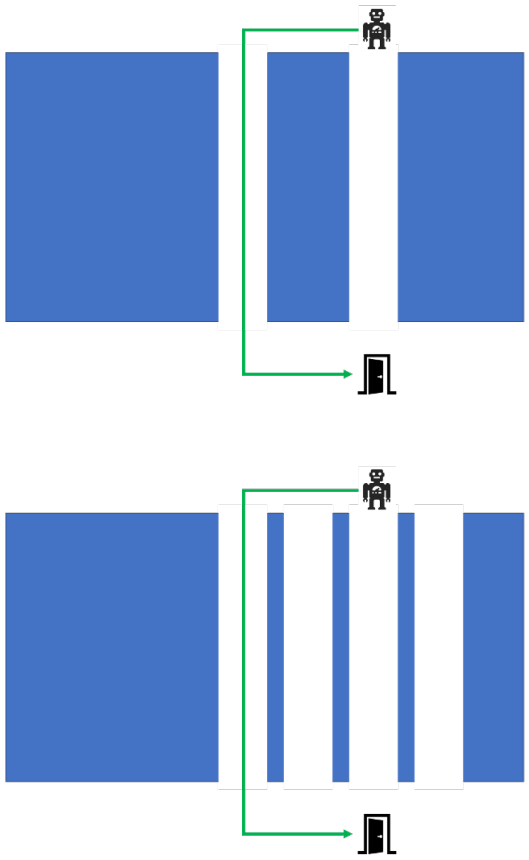

Figure 4: A possible scenario, where the introduction of new plans could cause the explicability to drop.

Property 1. *Explicability of a trace is dependent not only on the distance from the expected plans but also on the presence or absence of plans close to the expected plans.*

Here the plan remains explicable whether or not the observation leads to all the probability being assigned to a single model versus being distributed across multiple models. This means that the formulation doesn't require the human to have a single explanation for the behavior, rather it allows their belief to be distributed across multiple hypotheses. While the exact values would depend on the likelihood function, in the office robot scenario our formulation would assign high explicability scores (need not be the same) to both P1 and P2. For P1, the probability mass would be distributed across the two possible hypotheses corresponding to the two goals, while for P2 the probability mass is centered around the model corresponding to the goal to fetch coffee.

Property 2. *Explicability is agnostic to whether it is supported by multiple models or by a single one.*

Further, the explicability of a trace is now controlled by the priors on the models. E.g., a trace that is only possible in a model with low prior will not have high explicability score even if it is highly likely in that model.

3.2 Legibility

The next class of behavior is the one where the agent is trying to choose behavior that increases the agent's belief about some component (captured by the parameter θ) of the agent model. Such behavior could be especially important when the achievement of the human's desired outcome is tied directly to the model possessing a certain parameter value. An obvious example would be establishing if the end-goal itself is what the human desires, but this could also be in relation to other model parameters. Thus inducing high confidence in relation to such model parameters in the human's mind could be tied intimately with engendering trust in the human that the agent can achieve the desired objectives.

$$C_{\mathcal{H}}(\tau) \propto \sum_i \alpha_i * (1 - P(\theta = \theta(\mathcal{M}^R)|\tau^i_{pre})) \quad (7)$$

That is the cost here becomes the weighted sum of the probability associated with the target parameter having the incorrect value (i.e. different from what is true in the robot model) at each step. Keeping with the existing literature, we will refer to such behaviors as *legible behavior*, with the actual legibility score of a behavior prefix being proportional to the probability of the parameter being the true value

$$\mathcal{L}^{\theta}(\tau_{pre}) \propto P(\theta = \theta(\mathcal{M}^R)|\tau_{pre}) \quad (8)$$

$$\propto \Sigma_{\mathcal{M} \in \mathbb{M}^R_h \setminus \{\mathcal{M}^0\} \text{ Where } \theta(\mathcal{M}^R) = \theta(\mathcal{M})} P(\mathcal{M}|\tau_{pre}) \quad (9)$$

We skip $\mathcal{M}^0$ since it doesn't correspond to an explicit model in the human's mind. In the context of Figure 1, a plan prefix with high legibility score for the goal of deliver coffee would be P2 as compared to the other options illustrated. Since P1, allows for an optimal completion for both objectives and P3's completions in both models are equally bad. As we will see, while original formulations might assign P4 as a more legible option given the fact that it would assign zero probability to delivering mail, our formulation allows for the possibility that P4 may lead to more probability getting assigned to $\mathcal{M}^0$.

Reduction to Previous Legibility Definitions: Legibility was originally formalized [Dragan *et al.*, 2013] as the ability of a behavior to reveal its underlying objective. This involves a human who is considering a set of possible goals ($\mathbb{G}$) of the agent and is trying to identify the real goal by observing its behavior. Legibility is, thus, the maximization of the probability of the real goal through behavior:

$$\hat{\tau}^*_{\mathcal{L}} = argmax_{\tau_{pre}} P(G^R|\tau_{pre}) \quad (10)$$

where G^R is the agent's true goal. While originally introduced in the context of motion planning, this was later adapted to task planning by [MacNally *et al.*, 2018], and generalized to implicit communication of beliefs when the human has partial observability by [Kulkarni *et al.*, 2019a] as well as to implicit communication of any model parameter by [Miura and Zilberstein, 2020].

To keep the discussion in line with previous works, we will focus our attention on communicating end-goals (over arbitrary parameters). Some central assumptions made by earlier works is that the model only differs in terms of the end goal and the actual model is part of the set ($\mathcal{M}^R \in \mathbb{M}^R_h$). Also, the agent is expected to communicate its information as early as possible, so earlier α_i terms are given higher weights than the latter ones. They also assume that at no point would the human consider goals outside the explicit ones she had in mind. That is the possibility that she may be wrong about the original model and that the agent may be possibly trying to achieve something she didn't consider before would never cross her mind. In our framework, this would correspond to assigning a zero prior to $\mathcal{M}^0$. Thus the legibility score here would be

$$\mathcal{L}^{\theta}(\tau_{pre}) \propto P(M^R|\tau_{pre}) \quad (11)$$

A zero prior on $\mathcal{M}^0$ means the agent can create extremely circuitous routes as legible behavior provided the behavior is more likely in the agent model than others. This means that regardless of how suboptimal the plan is in the agent model (or ones with the parameter value), given its even lower probability in other models (or for other parameter values) the agent model will get assigned higher posterior probability and thus higher legibility score. For example in Figure 1, the restricted formulation would select the prefix P4 highlighted in red in order to reveal the goal of delivering coffee, eventhough that corresponds to an extremely sub-optimal plan given the set of possible plans.

Novel Properties of Generalized Legibility: A core assumption relaxed by the general formulation is that we now allow for the possibility that the human could be surprised by unexpected behavior and they may form new hypotheses about the agent. If you assume a non-zero prior for $\mathcal{M}^0$, then in cases where the agent presents an extremely suboptimal behavior they have a new hypothesis they can consider. That is they can now shift some of their belief to the fact that they may have been originally wrong about the agent model. Going back to the case of route P4 in Figure 1, given how far it is from the optimal, any completion of that prefix would have extremely low likelihood in the model for delivering coffee as opposed to $\mathcal{M}^0$ where that path is as likely as any other. This means our formulation now assigns more weight to $\mathcal{M}^0$ and thus capturing the fact that, when presented with highly unlikely behavior, the observer may question their beliefs about the agent. This brings us to the property

Property 3. *Inexplicable plans are also illegible.*

We believe allowing for such uncertainty is essential to capture more realistic human-robot interaction as it is rare for people to have absolute certainty about the agent models (and even discard the possibility that something might have just gone wrong with the agent). Also if we wish to move to a more longitudinal setting, indicating that the human no longer believes in one of the possible hypotheses in the set may not be enough, but we may need to explicitly try to identify what the newly formed hypothesis might be.

3.3 Predictability

The final case is one where the agent is interested in communicating to the human the future behavior it will be selecting. In this case, the agent would be required to choose behavior prefixes that allow the human to correctly guess the rest of the behavior the agent will follow with high confidence. This may be useful in cases where the agent may be sharing a workspace with the observer and may want to allow the observer to take into account future agent actions when coming up with their plans.

$$C_{\mathcal{H}}(\tau) \propto \sum_i \alpha_i * (1 - P(\tau^i_{post}|\hat{\tau}^i_{pre})) \tag{12}$$

This gives us *predictable behavior*. Further, $P(\tau^i_{post}|\hat{\tau}^i_{pre})$ denotes the predictability score for the prefix τ^i_{pre} (with respect to the completion τ^i_{post})

$$\begin{aligned} \mathcal{P}^{\tau^i_{post}}(\tau^i_{pre}) &\propto P(\tau^i_{post}|\tau^i_{pre}) \\ &\propto \sum_{\mathcal{M}\in\mathbb{M}^R_h} P(\tau^i_{post}|\tau^i_{pre}, \mathcal{M}) \times P(\mathcal{M}) \end{aligned} \tag{13}$$

From Figure 1, a plan prefix with high predictability would be P3. Given the prefix P3, the completion of going down the corridor has the highest likelihood for both the goals. So after marginalizing across all possible models that completion will have high probability and therefore the prefix has high predictability.

Reduction to Previous Predictability Definitions: We need to incorporate two main assumptions into the framework to reduce it to existing definitions of predictability: (1) the human observer only has a single explicit model about the agent and this is equal to the actual agent model $\mathbb{M}^R_h = \{\mathcal{M}^R, \mathcal{M}^0\}$ and (2) the user will not form new hypothesis about the agent regardless of how unexpected the behavior is (i.e. the $\mathcal{M}^0$ prior is zero). Thus we get:

$$\mathcal{P}^{\tau}(\tau_{pre}) \propto P(\tau' = \tau|\tau_{pre}, \mathcal{M}^R) \tag{14}$$

This directly maps to the predictability measure as defined in earlier works [Fisac *et al.*, 2020]. Previous works have also looked at the possibility of generating k step predictable plans, i.e., plans that try to guarantee predictability only after k steps. This allows for the system to choose unlikely prefixes for cases where the agent is only required to achieve required levels of predictability after the first k steps. We can capture such optimization preferences by setting α_i to zero for all but $i = k$. Going back to the example, prefix P3 optimizes for predictability for $k = 5$.

Generalized Predictability: Our generalization introduces two new aspects to the predictability formulation. The fact that the human now considers potential models and we also introduce the new hypothesis $\mathcal{M}^0$. However, the formulation marginalizes out the model and thus, effectively, for a given prefix, the human observer has to consider all the possible completions of the prefix in each of the individual models. Thus even if the trace is perfectly predictable in an individual model, the fact that the human has uncertainty over the models means the prefix may not be predictable. On the other hand, the fact that $\mathcal{M}^0$ assigns equal probability to all the possible completions would mean that the introduction of this new hypothesis would have less of an influence on the resulting predictability score.

3.4 Deception and Interpretability

The interpretability measures being discussed involve leveraging reasoning processes at the human's end to allow them to reach specific conclusions. At least for legibility and predictability, the behavior is said to exhibit a particular interpretability property only when the conclusion lines up with the ground truth at the agent's end. But as far as the human is concerned, they would not be able to distinguish between cases where the behavior is driving them to true conclusions or not. This means that the mechanisms used for interpretability could be easily leveraged to perform behaviors that may be adversarial [Chakraborti *et al.*, 2019a]. Two common classes of such behaviors are deception and obfuscation. Deceptive behavior corresponds to behavior meant to convince the user of incorrect information about the agent model or its future plans [Masters and Sardina, 2017]:

$$\mathcal{D}^{\mathbb{M}^R_h}(\tau_{pre}) \propto -1 * P(\mathcal{M}^R|\tau_{pre}) \tag{15}$$

Adversarial behaviors meant to confuse the user are either inexplicable plans that increase the posterior on $\mathcal{M}^0$ or, plans that actively obfuscate [Keren *et al.*, 2016; Kulkarni *et al.*, 2019a]:

$$\mathcal{O}^{\mathbb{M}^R_h}(\tau_{pre}) \propto H(\mathbb{M}^R_h|\tau_{pre}) \tag{16}$$

This is proportional to the conditional entropy of the model distribution given the observed behavior.

With explicability, the question of deceptive behavior becomes interesting, since explicable plan generation is relevant when the actual agent model may not be part of the human's expected set of models (else the agent could just follow its optimal behavior). By choosing to generate plans that align with a non-true model, explicability can be seen as deceptive behavior as it is reinforcing incorrect notions about the agent's model. Such plans would have a high deceptive score per the formulation above (since $P(\mathcal{M}^R|\tau) = 0$). One can argue that explicable behaviors are white lies in such scenarios as the goal here is just to ease the interaction and the behavior is not driven by any malicious intent. One could even further restrict the explicability formulation to a version that only lies by omission by restricting the agent to just behavior optimal in the original agent model. The agent chooses from this set the one that best aligns with the human's expectation. It is a lie by omission in the sense that while the agent has not explicitly been deceptive, by choosing behavior that aligns with the human's expectations, it is maintaining the human's incorrect beliefs.

4 Implications of the Framework

Below we briefly discuss several implications of our unifying framework.

Legibility, Explicability: These notions are related to the human's desire to recognize the model [Aineto *et al.*, 2019]). Our formulation shows that outside limited cases, legibility, and explicability are closely connected. Earlier works have been separating these measures by assuming away either legibility, like in existing explicability works with the human's hypothesis consisting of a single model [Zhang *et al.*, 2017; Kulkarni *et al.*, 2019c], or by assuming away explicability by assigning zero prior on $\mathcal{M}^0$ for legibility [Dragan *et al.*, 2013; Dragan and Srinivasa, 2013; MacNally *et al.*, 2018; Kulkarni *et al.*, 2019a; Miura and Zilberstein, 2020]. Interestingly, in cases where the human is aware that the agent is trying to be legible or more generally they know the agent is trying to model the observer, the human may be more open to suboptimal behavior from the agent as they might attribute it to the agent trying to communicate. However, this does not eliminate $\mathcal{M}^0$ but instead introduces a new level of nesting for reasoning. This comes with all the known complexities and pitfalls of reasoning with nested beliefs [Fagin *et al.*, 2003]. Though studying a limited amount of additional nesting could be important especially in cases where the agent plans to leverage communication. Since communication strategies make the most sense when the human is expecting the agent to model them.

Longitudinal Interactions: Our formulation currently looks at interpretability metrics for one-off interactions only. In cases where a human interacts with the agent for a long period, we can expect the user to start with a uniform distribution over models and a low probability for $\mathcal{M}^0$. In order to take a more long-term view of the human's interaction with the same agent (say, over a time horizon), legibility and predictability measures can be handled by directly carrying over the posterior from each interaction to the next one. However, for explicability more care needs to be taken. For example, [Kulkarni *et al.*, 2020] hypothesize a possible discounting of inexplicable behavior. The paper argues that after the first inexplicability, a human would be less surprised when similar inexplicable behavior is again presented to her. Part of this discounting can be explained by the human forming new hypothesis that explain the unexpected behavior and using that to analyze future agent behavior. As mentioned earlier, going to a longitudinal setting may require introducing new mechanisms to identify such newly formed hypothesis.

Planning and Environment Design: One of the next logical steps would be to facilitate the generation of plans that maximize the measures described in the paper. In particular, we could build on the work done in [Sreedharan *et al.*, 2020a] to encode the human's belief about the task into the planning space. Though in this case, given the possible multiple hypotheses held by the human, we would have to consider a belief space formulation, where the state includes information about the various models and each one is associated with a probability. Now each action of the robot has an effect on the likelihood of each hypothesis. Also unlike the earlier formulations, we can't just have a cost associated with each action or even one that is state-dependent. In fact for measures like predictability, not only does the plan cost for an intermediate step depend on the current state but also the eventual path that will be followed by the robot. As such these costs can only be computed at each goal node, where a cost will be assigned to each step of the path to the goal. These formulations could also be used to design the environment to facilitate easy generation of behavior that naturally aligns with these objectives (similar to [Kulkarni *et al.*, 2020]).

Goal/Plan/Model Recognition and Interpretability: This paper focuses on scenarios where the agent is acting in the world, with the knowledge that it is being observed. Though there could well be scenarios where the agent may be the observer and trying to reason about the human's model. In these settings, the agent may be engaged in similar reasoning to what is expected of the human in this paper. Chief among them is the case of model recognition [Aineto *et al.*, 2019] and it's more popular special case of goal recognition [Baker *et al.*, 2007; Ramírez and Geffner, 2009]. In a way this could be viewed as the inverse of the legibility as studied in the paper and is also associated with explicability. Though in most cases these papers assume away the possibility of the agent being surprised by assuming that the candidate hypothesis set contains the target model/goal that generated the behavior. But as the community starts shifting to more open-world cases or allow for the possibility of novel behavior [Senator, 2019], we will need to allow for the possibility that our agents may come across truly novel and inexplicable behavior (as per previous beliefs) and enable them to detect and update their beliefs from such behaviors. The next related class of abductive reasoning problems that have been studied in the literature is that of plan recognition [Kautz and Allen, 1986], wherein the agent tries to identify the full plan/behavior from some observations. One could consider this to be the inverse of the predictability problem studied in the paper.

Generalized Collaborative Behavior: One of the goals of the generalized human-aware planning problem is to establish the fact that specific interpretability behaviors could naturally be generated by an agent capable of reasoning about the human's belief and the impact of its actions on these beliefs. Though studying individual measures are still helpful not only in creating more specialized algorithms for generating them but also for understanding general strategies the agent may engage in. In this vein, we could further generalize most of the interpretability strategies the agent could engage in, to two broad categories, namely, a *model-communication* strategy or a *model-following* strategy. Model-communication involves molding the human's expectation through implicit or explicit communication, to allow the agent to achieve their objectives. On the other hand, the model following strategy involves taking the current understanding of the human and generating behavior that conforms to current human expectations. One could see the agent engaging cyclically in model-communication and model-following behaviors or possibly a combination of the two (for example involving actions that may have epistemic side-effects), wherein the agent may choose to mold the user's expectations to a point where their behavior may be better received by the observer. Legible behavior and explanations [Sreedharan *et al.*, 2020a] could be understood as specific instances of model communication strategies, while explicability can be seen as an ex-

ample of a model following behavior. Predictability is a bit harder to place, as at any point the agent is following the most likely plan as per the human's beliefs. Though the agent may have previously engaged in behavior meant to limit its future behaviors (including using techniques like projection [Chakraborti *et al.*, 2018]). One could definitely argue that these earlier efforts are in fact communicative in so far as they are trying to inform the human about the agent's intentions.

5 Conclusion

Works on interpretable behavior generation have generally focused on studying and defining individual human-aware measures under limited settings. By placing and studying human-aware planning in a more general context we are able to not only see connections between these works that were previously ignored but also help formalize new collaboration strategies. In terms of future work, one plausible direction would be performing a validation of the novel properties of this framework through user studies. While we have started investigating in this direction with preliminary studies (c.f. [Sreedharan *et al.*, 2020b]), further work is required to validate it in more general settings. Another avenue for future work would be to extend the current setting to more involved scenarios. While there are multiple ways we could extend the current setting, two that may actually be worth considering are the ones with partial observability as well as settings where humans play the role of an actor (as against just an observer). It has already been shown by Kulkarni *et al.* [2019a] that partially observability is a factor that can be directly exploited to generate observed sequences which optimize for certain interpretability/deception measures, in fact for users with different observation models [Kulkarni *et al.*, 2019b] shows that the agent can choose a single observation sequence that generates different effects in their respective mental models. Another factor worth considering is whether the human is an actor in the world, in such cases, the robot's ability to be interpretable, could not only affect the choice of human's actions but could also have safety implications. For instance [Kulkarni, 2021], present a work where interpretability measures are explored in the context of a human actor.

Acknowledgements

We would like to thank Dr. Tathagata Chakraborti for his significant contributions to an earlier version of this paper and for extensive discussions on the topic. This research is supported in part by ONR grants N00014-16-1-2892, N00014-18-1- 2442, N00014-18-1-2840, N00014-9-1-2119, AFOSR grant FA9550-18-1-0067, DARPA SAIL-ON grant W911NF-19- 2-0006, NASA grant NNX17AD06G, and a JP Morgan AI Faculty Research grant.

References

[Aineto *et al.*, 2019] Diego Aineto, Sergio Jiménez, Eva Onaindia, and Miquel Ramírez. Model Recognition as Planning. In *ICAPS*, 2019.

[Baker *et al.*, 2007] Chris L Baker, Joshua B Tenenbaum, and Rebecca R Saxe. Goal Inference as Inverse Planning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2007.

[Baker *et al.*, 2009] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action Understanding as Inverse Planning. *Cognition*, 2009.

[Chakraborti *et al.*, 2017] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy. In *IJCAI*, 2017.

[Chakraborti *et al.*, 2018] Tathagata Chakraborti, Sarath Sreedharan, Anagha Kulkarni, and Subbarao Kambhampati. Projection-Aware Task Planning and Execution for Human-in-the-Loop Operation of Robots in a Mixed-Reality Workspace. In *IROS*, 2018.

[Chakraborti *et al.*, 2019a] Tathagata Chakraborti, Anagha Kulkarni, Sarath Sreedharan, David E Smith, and Subbarao Kambhampati. Explicability? Legibility? Predictability? Transparency? Privacy? Security? The Emerging Landscape of Interpretable Agent Behavior. In *ICAPS*, 2019.

[Chakraborti *et al.*, 2019b] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. Balancing Explanations and Explicability in Human-Aware Planning. In *IJCAI*, 2019.

[Chakraborti, 2018] Tathagata Chakraborti. *Foundations of Human-Aware Planning – A Tale of Three Models*. PhD thesis, ASU, 2018.

[Dragan and Srinivasa, 2013] Anca Dragan and Siddhartha Srinivasa. Generating Legible Motion. In *RSS*, 2013.

[Dragan *et al.*, 2013] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. Legibility and predictability of robot motion. In *HRI*, 2013.

[Dragan *et al.*, 2015] Anca D Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S Srinivasa. Effects of Robot Motion on Human-Robot Collaboration. In *HRI*, 2015.

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

[Fagin *et al.*, 2003] Ronald Fagin, Yoram Moses, Joseph Y Halpern, and Moshe Y Vardi. *Reasoning About Knowledge*. MIT press, 2003.

[Fisac *et al.*, 2018] Jaime F Fisac, Andrea Bajcsy, Sylvia L Herbert, David Fridovich-Keil, Steven Wang, Claire J Tomlin, and Anca D Dragan. Probabilistically Safe Robot Planning with Confidence-Based Human Predictions. In *RSS*, 2018.

[Fisac *et al.*, 2020] Jaime F Fisac, Chang Liu, Jessica B Hamrick, Shankar Sastry, J Karl Hedrick, Thomas L Griffiths, and Anca D Dragan. Generating Plans that Predict Themselves. In *Algorithmic Foundations of Robotics*. Springer, 2020.

[Gunning and Aha, 2019] David Gunning and David W Aha. DARPA's Explainable Artificial Intelligence Program. *AI Magazine*, 2019.

[Kautz and Allen, 1986] Henry A Kautz and James F Allen. Generalized plan recognition. In *AAAI*, volume 86, page 5, 1986.

[Keren *et al.*, 2016] Sarah Keren, Avigdor Gal, and Erez Karpas. Privacy Preserving Plans in Partially Observable Environments. In *IJCAI*, 2016.

[Kulkarni *et al.*, 2019a] Anagha Kulkarni, Siddharth Srivastava, and Subbarao Kambhampati. A Unified Framework for Planning in Adversarial and Cooperative Environments. In *AAAI*, 2019.

[Kulkarni *et al.*, 2019b] Anagha Kulkarni, Siddharth Srivastava, and Subbarao Kambhampati. Signaling friends and head-faking enemies simultaneously: Balancing goal obfuscation and goal legibility. *arXiv preprint arXiv:1905.10672*, 2019.

[Kulkarni *et al.*, 2019c] Anagha Kulkarni, Yantian Zha, Tathagata Chakraborti, Satya Gautam Vadlamudi, Yu Zhang, and Subbarao Kambhampati. Explicable Planning as Minimizing Distance from Expected Behavior. In *AAMAS Extended Abstract*, 2019.

[Kulkarni *et al.*, 2020] Anagha Kulkarni, Sarath Sreedharan, Sarah Keren, Tathagata Chakraborti, David Smith, and Subbarao Kambhampati. Designing environments conducive to interpretable robot behavior. *IROS*, 2020.

[Kulkarni, 2021] Anagha Kulkarni. *Synthesis of Interpretable and Obfuscatory Behaviors in Human-Aware AI Systems*. PhD thesis, 2021.

[L Griffiths *et al.*, 2008] Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. Bayesian Models of Cognition. *The Cambridge Handbook of Computational Psychology*, 2008.

[Langley *et al.*, 2017] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. Explainable Agency for Intelligent Autonomous Systems. In *IAAI*, 2017.

[MacNally *et al.*, 2018] Aleck M MacNally, Nir Lipovetzky, Miquel Ramirez, and Adrian R Pearce. Action Selection for Transparent Planning. In *AAMAS*, 2018.

[Masters and Sardina, 2017] Peta Masters and Sebastian Sardina. Deceptive path-planning. In *IJCAI*, 2017.

[Miura and Zilberstein, 2020] Shuwa Miura and Shlomo Zilberstein. Maximizing plan legibility in stochastic environments. In *AAMAS*, 2020.

[Ramírez and Geffner, 2009] Miquel Ramírez and Hector Geffner. Plan recognition as planning. In *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

[Reddy *et al.*, 2018] Sid Reddy, Anca Dragan, and Sergey Levine. Where Do You Think You're Going?: Inferring Beliefs about Dynamics from Behavior. In *NeurIPS*, 2018.

[Senator, 2019] Mr Ted Senator. Science of artificial intelligence and learning for open-world novelty (sail-on), 2019.

[Sreedharan *et al.*, 2020a] Sarath Sreedharan, Tathagata Chakraborti, Christian Muise, and Subbarao Kambhampati. Expectation-Aware Planning: A Unifying Framework for Synthesizing and Executing Self-Explaining Plans for Human-Aware Planning. In *AAAI*, 2020.

[Sreedharan *et al.*, 2020b] Sarath Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, David E Smith, and Subbarao Kambhampati. A bayesian account of measures of interpretability in human-ai interaction. *arXiv preprint arXiv:2011.10920*, 2020.

[Zabell, 1992] Sandy L Zabell. Predicting the unpredictable. *Synthese*, 90(2):205–232, 1992.

[Zhang *et al.*, 2017] Yu Zhang, Sarath Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, Hankz Hankui Zhuo, and Subbarao Kambhampati. Plan Explicability and Predictability for Robot Task Planning. In *ICRA*, 2017.