Planning for Signaling in Low-Trust Environments

Septia Rani, Turgay Caglar, Sarath Sreedharan

Keywords: Human-AI Interaction, Trust Signaling, Preemptive Commitment, Legible Avoidance.

Summary

In human-AI interaction, a low-trust environment could emerge when users are uncertain about the agent's capabilities, intentions, or reliability. This may hinder not only effective collaboration but also the adoption of extremely beneficial systems. In this study, we present a general framework to model such environments and a strategy to improve user trust, namely, through signaling behavior. Specifically, we propose two signaling techniques for an AI agent in low-trust environments. The first strategy is called preemptive commitment, in which the agent attempts to drive itself into a state from which it cannot violate any of the normative contracts. The second strategy is called legible avoidance, which involves the agent conveying that it is not trying to violate a contract in the normative contracts by moving away from the object of interest. We evaluated the performance of these strategies across standard MDP planning benchmarks and conducted a user study to investigate how users perceive preemptive commitment and legible avoidance signaling techniques, particularly their trust level when the agent demonstrates these techniques.

Our experimental results suggest that legible avoidance behavior, while computationally more expensive to compute, is preferred over preemptive commitment and optimal behavior across multiple dimensions.

Contribution(s)

1. We formalize the notion of low-trust environments and trust violations for MDP planning, propose the notion of normative contracts, and use it as a way to define signaling behaviors.
   Context: Human-AI Interaction, Human-Robot Interaction.
2. We identify two classes of signaling behavior, namely, preemptive commitments and legible avoidance. We also develop methods to generate such behavior for a given agent's MDP and a set of possible normative contracts.
   Context: Similar notions have been used to communicate goals (cf. (Dragan et al., 2013)), but not to specify what the agent is trying to avoid.
3. We conducted computational experiments to identify trade-offs among the behavior generation methods. We also performed a detailed user study to identify how people perceive the different behavior types.
   Context: Empirical evaluation of interpretable/legible agent behavior and human trust in Human-Robot Interaction (Dragan et al., 2013; Hancock et al., 2011). We use well-established psychometric tools to measure human trust. Human trust is usually represented as a multi-dimensional measure using dimensions like predictability, reliability, faith, and overall trust (Muir, 1994).

Planning for Signaling in Low-Trust Environments

Septia Rani, Turgay Caglar, Sarath Sreedharan
{septia.rani,turgay.caglar,sarath.sreedharan}@colostate.edu

Department of Computer Science, Colorado State University

Abstract

In human-AI interaction, a low-trust environment could emerge when users are uncertain about the agent's capabilities, intentions, or reliability. This may hinder not only effective collaboration but also the adoption of extremely beneficial systems. In this study, we present a general framework to model such environments and a strategy to improve user trust, namely, through signaling behavior. Specifically, we propose two signaling techniques for an AI agent in low-trust environments. The first strategy is called preemptive commitment, in which the agent attempts to drive itself into a state from which it cannot violate any of the normative contracts. The second strategy is called legible avoidance, which involves the agent conveying that it is not trying to violate a contract in the normative contracts by moving away from the object of interest. We evaluated the performance of these strategies across standard MDP planning benchmarks and conducted a user study to investigate how users perceive preemptive commitment and legible avoidance signaling techniques, particularly their trust level when the agent demonstrates these techniques.

1 Introduction

A low-trust environment is a setting in which individuals or entities interacting within it lack confidence in others' intentions, reliability, or competence. This characterization aligns with standard definitions of trust as confidence in others' intentions and reliability (Rotter, 1967; Rousseau et al., 1998) and accounts of trustworthiness that emphasize ability or competence (Mayer et al., 1995). Individuals in a low-trust environment tend to feel anxious or threatened by others' actions and may assume that others have harmful intent in their behavior (Fermin et al., 2022), which, in turn, can lead to reduced or no collaboration. Similarly, in human-AI interaction, a low-trust environment could emerge when users are uncertain about the agent's capabilities, intentions, or reliability. Thus hindering not only collaboration but also the adoption of extremely beneficial systems.

In existing human-AI collaboration, the primary tool for engendering trust in AI systems remains explanations (Ferrario & Loi, 2022; Cheung & Ho, 2025). While explanations have been shown to be effective in communicating agents' competence and capabilities (Qin et al., 2025), they do not provide adequate tools to address low trust arising from the user's suspicion of the AI system's intent. After all, a robot claiming to mean no harm is unlikely to calm a person convinced that the robot is out to get them. Thankfully, economics and game theory have identified various tools that can be leveraged in such situations (Johnson & Mislin, 2011; Han et al., 2021). Chief among them is that of signaling behavior (Wojtowicz & DeDeo, 2025), in which an agent performs a set of costly actions to implicitly communicate its intentions.

In this paper, we will look at how one could account for the automatic generation of such behavior during the planning of agent behavior. To do so, we will start by formalizing the notion of planning in low-trust environments. A key part of the formulation is a set of implicit behavioral expectations that the human places on the robot, that they do not want the robot to violate. Consider the cleaning

robot visualized in Figure 1. There is an expectation that the robot will not steal the valuables the user placed in their safe for its programmers. We will refer to such expectations as Normative Contracts (NCs). The goal of our formulation is to generate behaviors that signal to the user that the robot is not trying to violate such contracts.

As we will see, our formulation naturally gives rise to different intuitive strategies to signal their intention not to violate NCs. The first one would involve taking an action early on that will limit their ability to violate NCs (in the given example, this may involve closing the door to the room containing the valuable items), and the other would involve prioritizing communicating their intent not to violate NCs at each step. In the example, this would involve moving away from the room entrance as early as possible, then proceeding to their eventual goal through an alternate route. We will refer to the first strategy as *Preemptive Commitment* and the other as *Legible Avoidance*.

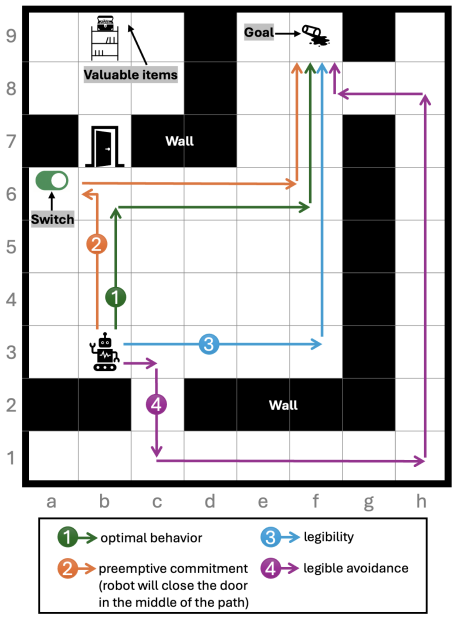

Figure 1: Robot cleaner's environment setting, with a highlight of the different paths.

Our paper will introduce planning methods for each of these two distinct strategies. We chose to focus on developing planning algorithms as the first step, as we believe this would provide a solid theoretical foundation for later developing learning methods that could generate such behaviors. We evaluate our algorithms on standard MDP benchmarks to identify their computational properties. Next, we conduct user studies to evaluate how non-expert users perceive these different strategies. To summarize, our contributions are as follows:

- We formalize the problem of planning in a low-trust environment (Sec. 4),
- We formalize and introduce a method to generate behavior for preemptive commitment (Sec. 5),
- We formalize and introduce an MILP encoding for legible avoidance (Sec. 6),
- We evaluate our proposed method using computational experiments and a user study (Sec. 7).

2 Related Works

As artificial intelligence (AI) becomes increasingly integrated into many aspects of human daily life, research on AI trust has become crucial. Human trust in AI is a key factor in the successful integration of AI into daily life (Afroogh et al., 2024). Trust guides reliance on automation, and miscalibrated trust may lead to the disuse of beneficial systems (Lee & See, 2004).

Understanding how humans develop trust in AI has been studied in the past few years. The development of trust is likely to be significantly influenced by the embodiment of AI (Glikson & Woolley, 2020). Researchers have studied AI in various forms of representation, including physical robots, virtual agents, and embedded systems within other tools. In human-robot interaction, a meta-analysis study found that the main factors contributing to the development of trust were robot performance and characteristics (Hancock et al., 2011). Additionally, researchers examined the development of human trust in AI based on the AI machine's capability level (Glikson & Woolley, 2020; Gonzalez et al., 2025).

The nature of trust can be different depending on the entities involved. Between human interactions, trust generally increases over time with repeated interactions. However, over time, human trust in

technology declines due to its errors and malfunctions (Madhavan & Wiegmann, 2007). Interestingly, when it comes to AI technology, the opposite progress can occur. Some researchers have reported that after a direct interaction with AI, a low level of trust from the first encounter may increase (Ullman & Malle, 2017). Developing or improving trust is important in many situations, as under-trust may lead to the disuse of beneficial systems (Parasuraman & Riley, 1997). We will use the mental model-based theory proposed by Zahedi et al. (2023) as the mathematical framework to capture trust. It posits that the user's trust in an agent to carry out a task or contract can be approximated by the likelihood the user assigns to the agent's carrying out the task, given their beliefs about the agent's capabilities and intent. This approach offers an alternative to subjective questionnaires (Muir, 1994) for estimating trust.

Improving trust often involves a signaling strategy, in which entities seeking to gain trust use specific signals to demonstrate their trustworthiness and reduce uncertainty in interactions. Signaling is particularly useful when two entities have access to different information (Connelly et al., 2011). A signal, in theory, can be any observable action or communication that carries information about the latent quality. While widely used, the cleanest mathematical account of signaling comes from signaling games (Lewis, 2008). In a signaling game, a sender has private information about the world and can choose a signal to convey it. The receiver observes the signal and takes actions based on their interpretation. The reliability of the signal stems from its cost, as a lower-quality sender is often unable to fake a high-cost signal. Examples of signals in the real world include a premium pricing strategy that signals the high quality of products to customers (Verma & Gupta, 2004), or a job applicant who uses their education level to signal their capabilities to a prospective employer (Fudenberg & He, 2018). Recent work has raised concerns that the widespread adoption of AI may, in fact, reduce the effectiveness of many signaling actions, as it tends to reduce the cost associated with those actions (Wojtowicz & DeDeo, 2025).

Previous works have looked at the viability of using behaviors as a way to communicate information about the agent model, including intent (cf. (Chakraborti et al., 2019)). Of particular importance is that of legible motion (Dragan et al., 2013), which looks at communicating an agent's goal by choosing a behavior that is more likely for a target goal as opposed to other potential goals being considered by the user. For example, in Figure 1, the legible path is represented in blue if the user believes the agent's goal is to either get to the door or clean the floor. Existing legibility methods expect the true agent goal to be part of the possible candidate goals they consider.

3 Background

In this paper, we focus on problems that can be represented as Markov Decision Processes (MDPs) (Puterman, 1990). MDP can be defined by a tuple $\mathcal{M} = \langle S, A, R, P, \gamma, s_0 \rangle$. In this notation scheme, S is a set of states where $s_0 \in S$ is the initial state, A is a set of actions, R is a reward function, P is the transition function such that $P(s, a, s')$ represents the probability that the environment moves to next state s' when the agent takes action a in current state s, and $\gamma \in [0, 1)$ is the discount factor.

A solution to an MDP is represented by mappings from states to actions, called policies. Given a policy π, the value function V_π captures the expected sum of discounted rewards or the expected utility obtained from executing the policy in a given state, and Q_π represents the value obtained by executing a specific action followed by the policy. An optimal policy π_* is a policy that has the highest expected utility (Russell et al., 1995).

In an MDP, a state-action trajectory or trajectory, is a potentially infinite sequence of the form $\tau = (s_0, a_0, s_1, a_1, \ldots, s_{T-1}, a_{T-1}, s_T)$. A trajectory is valid if every transition is consistent with the MDP's definition, i.e., for all t in horizon T, the action a_t is an element of the valid action set for state s_t, and the next state s_{t+1} has a nonzero probability value under the transition model $P(s_t, a_t, s_{t+1})$. If we have a valid trajectory τ, we can compute its return (usually denoted by $\mathcal{G}$), which is the discounted cumulative reward obtained along that trajectory.

In this paper, we will use the notion of occupancy frequency $x^\pi(s,a)$ that is associated with a policy π (Poupart, 2005). The occupation frequency $x^\pi(s,a)$ represents the frequency with which an action a is taken in a state s under the policy π. The value obtained under a policy can be reformulated in terms of its rewards and occupancy frequency. By solving the following linear program (LP) expression that incorporates occupancy frequency variables, we can obtain the optimal policy:

$$\max_x \sum_{s,a} x(s,a)\, r(s,a) \quad \text{s.t.} \qquad \forall s \in S,\ \sum_a x(s,a) = \delta(s,s_0) + \gamma \sum_{s',a'} x(s',a')\, P(s',a',s) \tag{1}$$

In the above equation, x is the set of variables that captures the occupancy frequency possible under a policy, and $\delta(s,s_0)$ is an indicator function that returns 1 when $s = s_0$ and zero otherwise.

4 Trust Model for MDPs

We consider a setting in which a human observes an agent acting in the environment. Our formulation will start with the definition of the original agent task, given as $\mathcal{M}^\alpha = \langle S, A, R^\alpha, P, \gamma^\alpha, s_0 \rangle$, where the terms R^α and γ^α correspond to the specific agent reward and discount factor. We will limit our attention to MDPs with finite state spaces. We will denote the minimal value the reward function can take as R^α_{min}, i.e, $R^\alpha_{min} = \min_{s,a} R^\alpha(s,a)$. Such a min bound on the reward function also directly gives us a lower bound on the optimal policy for the model.

Property 1. *For any given MDP $\mathcal{M}$ with a minimal value R_{min}, we can guarantee that the optimal value satisfies the condition: $V_*(s_0) \geqslant \frac{1}{(1-\gamma)} R_{min}$*

This property follows directly from the fact that the worst-case scenario would be one where the optimal policy always chooses actions that result in R_{min}. One of our starting assumptions is that the human is unaware of the agent's intent, which in our case translates to uncertainty about the human reward function. Our first step is to translate the notion of contract-based trust defined in a deterministic goal-directed planning setting into an MDP setting. While the earlier paper (Zahedi et al., 2023) leaves the nature of contracts vague, we will start with one of the most basic contracts there could be in such a setting, namely that the agent will not perform some disallowed action in some prespecified states. We will denote such a contract as simply $\mathcal{C}_i = (s_i, a_i)$. As with the previous paper (Zahedi et al., 2023), the trust that the agent will uphold its contract is given as

$$\mathcal{T}(C_i) = \sum_\pi \mathcal{P}(\mathcal{C}_i|\pi) * \mathcal{P}^H(\pi)$$

Where $\mathcal{P}(C_i|\pi)$ is the likelihood the human assigns that a contract will be upheld in a given policy π, and $\mathcal{P}^H(\pi)$ is their prior about what policy the agent is following (which in turn reflects their uncertainty about the agent's intent). For simplicity, we assume a uniform distribution over $\mathcal{P}^H(\pi)$, but we could replace it with any prior. For our approaches, rather than directly using the trust value, we use a measure of mistrust, which we equate with their belief that the agent will break the contract (denoted by $\neg$). We could capture this as: $\bar{\mathcal{T}}(C_i) = \sum_\pi \mathcal{P}(\neg\mathcal{C}_i|\pi) * \mathcal{P}^H(\pi)$. Intuitively, the more frequently a policy visits the state-action pair (s_i, a_i), the more likely it is to violate contract C_i, making the occupancy frequency as a natural proxy for violation probability, $\mathcal{P}(\neg\mathcal{C}_i|\pi) \propto x^\pi(s_i, a_i)$.

Now, coming back to our original problem setting. Our problem setting is one where the user has set up a number of contracts that the user expects the agent not to violate. We refer to this set of contracts as the normative contracts (NC), denoted by $\mathbb{N}$. The question the human is struggling with is whether the agent may, in fact, be trying to break one, and if so, it would need to stop the agent from acting in the environment. While within the set of possible reward functions, there could be many that could end up violating one or more of the norms, an MDP of particular interest in our case is one whose sole purpose is to violate one or more of the norms. We refer to such an MDP as *norm violating* MDP, denoted as $\mathcal{M}^{\neg\mathbb{N}}$. More formally, these two components are defined as:

Definition 1. *For a given problem setting with an agent model $\mathcal{M}^\alpha = \langle S, A, R^\alpha, P, \gamma^\alpha, s_0 \rangle$, the **normative contracts (NC)** is a set of the form $\mathbb{N} = \{\mathcal{C}_1, ..., \mathcal{C}_k\}$, where $\mathcal{C}_i = (s_i, a_i)$, and $s_i \in S$*

and $a_i \in A$. For this set, the norm violating MDP is given as $\mathcal{M}^{\neg\mathbb{N}} = \langle S, A, R^{\neg\mathbb{N}}, P, \gamma^\alpha \rangle$, where: $R^{\neg\mathbb{N}}(s,a) = \begin{cases} R^+ \text{ if} (s,a) \in \mathbb{N} \\ 0 \text{ Otherwise} \end{cases}$ *, where R^+ is some positive number.*

A policy for the *norm violating* MDP will have a positive value only if it violates at least one contract within $\mathbb{N}$. Additionally, following the construction of our reward function, an optimal policy for $\mathcal{M}^{\neg\mathbb{N}}$ is one that violates as many norms as possible.

To simplify the formulation, we assume that each contract is independent, as such, the mistrust over the entire normative contract set can be given as: $\bar{\mathcal{T}}(\mathbb{N}) = \sum_{\mathcal{C}_i \in \mathbb{N}} \mathcal{P}(\neg\mathcal{C}_i|\pi') * \mathcal{P}^{\mathcal{H}}(\pi')$.

Now that the basic definition is in place, we are ready to move on to the actual generation of signaling behavior. In particular, a signaling policy is one that the agent can adopt to reduce mistrust on average. More formally,

Definition 2. *A policy π is said to be a signaling policy for a state s_0 if $\mathbb{E}_{\tau\sim\pi(s_0)}[\bar{\mathcal{T}}(\mathbb{N}|\tau)] < \bar{\mathcal{T}}(\mathbb{N})$, where $\mathbb{E}_{\tau\sim\pi(s_0)}[\bar{\mathcal{T}}(\mathbb{N}|\tau)] = \sum_\tau \sum_{\mathcal{C}_i \in \mathbb{N}} \mathcal{P}(\neg\mathcal{C}_i|\pi') * \mathcal{P}(\tau|\pi') * \mathcal{P}^{\mathcal{H}}(\pi')$.*

Here, $\mathcal{P}(\tau|\pi')$ captures the probability that a policy π' can generate the trajectory τ. It is worth noting that not all signaling policies are equally useful. One could signal that it is not trying to violate contracts by simply doing nothing. As such, in addition to signaling that the agent is trying not to violate a norm, it must also pursue its own task. In practice, we would want the policy value to exceed some threshold ℓ, set by the agent designer. Such requirements can be enforced in the MDP solver. We will show how to incorporate them as constraints into a mixed-integer linear programming (MILP) formulation for signaling behavior generation (Section 6).

Finally, by definition, a signaling policy could take any arbitrary form. However, in the rest of the paper, we focus on two intuitive categories of signaling behaviors. Our motivation for doing so is twofold. First, our focus on well-defined behavioral patterns enables us to design specialized algorithms, thereby allowing us to compare and contrast the computational overhead of generating them. Secondly, by focusing on behavior classes with analogues in real-world human behaviors, we can conduct rigorous user-subject studies to assess the effectiveness of such signaling behavior relative to the optimal one. In the following sections, we describe the two categories of behaviors and show how they can be generated.

5 Preemptive Commitment

The first strategy we will study is one where the agent tries to drive itself to a state from which it can no longer violate any of the contracts in $\mathbb{N}$. In the running example, it involves the robot voluntarily choosing to lock the door. We can see people exhibiting similar behavior under extreme conditions. For example, cases where a person might voluntarily give up weapons or lie on the ground when asked to by an authority figure could be seen as examples of such preemptive commitments.

To generate policies that yield a preemptive commitment strategy, we define a modified version of the agent MDP whose optimal value function automatically induces such behavior. We will refer to such MDPs as the Preemptive Commitment model or $\mathcal{M}^{PC}$, and formally define it as follows.

Definition 3. *For an agent model $\mathcal{M}^\alpha$, a normative contract set $\mathbb{N}$, and a norm violation MDP $\mathcal{M}^{\neg\mathbb{N}}$, a Preemptive Commitment model is given by the tuple of the form $\mathcal{M}^{PC} = \langle S, A, R^{PC}, P, \gamma^\alpha, s_0 \rangle$, where:*

$$R^{PC}(s,a) = \begin{cases} R^\alpha(s,a) \text{ if } Q_*^{\neg\mathbb{N}}(s,a) = 0 \\ R^- \quad \text{Otherwise}. \end{cases}$$

Here, $R^- < \epsilon * (1-\gamma) * R^\alpha_{min}$, $Q_*^{\neg\mathbb{N}}(s,a)$ is the optimal Q value function for $\mathcal{M}^{\neg\mathbb{N}}$, and ϵ is given as: $\epsilon = (p_{min})^{|S|}$ where p_{min} is the lowest non-zero probability transition possible under P.

As per the definition, the agent basically operates in two modes: one where it gets a lower reward than the original reward function, when it can still violate contracts, and goes back to the original reward function, once it gets to a point from which it can no longer violate the norms. Here, the term ϵ gives a lower bound on the likelihood of the shortest path to any state that is reachable in finite steps in the MDP. To see why $Q_*^{\neg\mathbb{N}}(s,a) = 0$ presents an irreversible transition point with respect to contract violation, consider the following property.

Property 2. *Let s_1, a_1 be a state-action pair such that $Q_*^{\neg\mathbb{N}}(s_1,a_1) = 0$, now there exists no valid trajectory of the form $\tau = \langle s_1, a_1, ...\rangle$, where there exists a s_i, a_i pair in the trajectory, such that $(s_i, a_i) \in \mathbb{N}$.*

This property follows from the definition of $\mathcal{M}^{\neg\mathbb{N}}$, and the fact that it is trying to maximize norm violation. If, in fact, there was a way to violate any norm that could result from a state resulting from the execution of a in s, then $Q_*^{\neg\mathbb{N}}(s_1,a_1)$ will have a positive value. This means that $Q_*^{\neg\mathbb{N}}(s_1,a_1) = 0$ is only true in cases where, after that action execution, the agent can never violate the norm again, and hence all future state actions will have $Q_*^{\neg\mathbb{N}}$ value of zero.

Another property we would like is that the agent tries to reach such states as soon as possible.

Property 3. *Let τ_1 and τ_2 be two trajectories such that τ_1^- and τ_2^- presents the suffix before the first state where $Q_*^{\neg\mathbb{N}}$ is zero, and τ_1^+ and τ_2^+ are the corresponding prefixes that include this state, then the return of the first trajectory should be greater than the second trajectory, $\mathcal{G}(\tau_1) > \mathcal{G}(\tau_2)$, provided the following conditions are*

- *C1 - $|\tau_1^-| < |\tau_2^-|$.*
- *C2 - $\tau_1^+ = \tau_2^+$ and there exists at least one transition with non-zero reward.*
- *C3 - $\gamma^\alpha > 0$.*

This property follows from two facts. First, return from both τ_1^- and τ_2^- is zero (per the reward function). Secondly, even though the suffixes are the same, the return from τ_2^+ is more heavily discounted due to the longer length of $|\tau_2^-|$. We now see that the optimal policy for this MDP satisfies the criteria for a signaling policy, as defined in Definition 2.

Theorem 1. *For an agent model $\mathcal{M}^\alpha$ and a normative contract set $\mathbb{N}$, the optimal policy for the corresponding Preemptive Commitment model $\mathcal{M}^{PC}$ is a signaling policy for s_0, provided there exists state action pair s, a, reachable from s_0 in finite steps with non-zero probability, such that $Q_*^{\neg\mathbb{N}}(s_1,a_1) = 0$.*

Proof Sketch. The cost of any policy that does not support visiting a state with $Q_*^{\neg\mathbb{N}}(s_1,a_1) = 0$, would be equal to $\frac{1}{(1-\gamma)}R^-$. This can be lower than the lower bound on the optimal value (per Property 1 and construction of R^-). Since there exists a state with $Q_*^{\neg\mathbb{N}}(s_1,a_1) = 0$ within finite steps, there exists a policy that can generate a trajectory with return greater than or equal to $\frac{1}{(1-\gamma)}R^\alpha_{min}$ with probability of at least ϵ. Since $\epsilon * \frac{1}{(1-\gamma)}R^\alpha_{min} > \frac{1}{(1-\gamma)}R^-$, any policy with such a trajectory should have a higher value than one that does not visit the state. Since visiting the state is the only way to increase the total policy value, the optimal policy must contain trajectories that pass through such states. If the policy contains such trajectories, then it must reduce the expected mistrust, as the presence of the trajectory cannot support policies that violate norms. Hence, proving that it is a signaling policy. □

However, we can see that this method has one disadvantage, namely that we cannot guarantee the existence of such policies under all conditions.

Property 4. *There exists an agent model $\mathcal{M}^\alpha$, such that no policy exists for $\mathcal{M}^{PC}$ which supports a trajectory that passes through a state from which norms cannot be violated.*

One can prove this by constructing a deterministic MDP in which every state is reachable from every other state. In this case, the agent always has the option to violate the norm. However, one could

check whether the model supports such preemptive commitment policies by checking the optimal value, and if it is greater than $\frac{1}{(1-\gamma)}R^{-}$.

It is worth noting that we could adopt variations of the definition, where instead of driving to a state where it is impossible to violate norms, one could simply go to a state where it is just costly to violate the norms. We could do this by updating the definition of R^{PC} such that it receives the true reward only if $Q_*^{\neg\mathbb{N}}s, a$ is below some acceptable limit. However, because the selection of such limits depends strongly on the user's risk tolerance, we leave this extension as future work.

6 Legible Avoidance

Now, we move to the next strategy, which will involve the agent trying to convey that it is not trying to violate a contract in $\mathbb{N}$, by moving away from it. In the robot example, the robot could take a more circuitous route that is farther away from the object of interest. Real-world analogies could include people backing away from dangerous or cordoned-off locations.

As mentioned in Section 2, the closest existing approach is that of legible planning, where the agent is trying to signal their goal. However, signaling their goal is not necessarily the same as signaling what they plan not to do. First off, the current legibility formulation assumes that the human has access to a set of candidate goal sets, which contains the true robot goal (cf. (Dragan et al., 2013)). This is not an assumption our work makes. Additionally, in a low-trust environment, even if communication is possible, the human may not accept it. Finally, communicating your goal and communicating that you are trying to avoid some constraints could give rise to very different behavior (consider the purple and blue lines in Figure 1).

To communicate that you are simply trying to avoid violation of some norms is easy to communicate. At every point, perform the action that is least likely to be performed by an agent trying to violate a norm in $\mathbb{N}$. If we were to consider that the human observer roughly uses a Boltzmann noisy rational model (Tang et al., 2012) to evaluate agent behavior, a simple signaling policy would be: $\pi(s) = \text{argmin}_a\, Q_*^{\neg\mathbb{N}}(s, a)$.

This is a signal policy trivially. However, following this policy will not guarantee that the agent satisfies the minimum value requirement $V_\pi^\alpha(s) \geqslant \ell$. This also brings us to a central difference between legibility and legible avoidance. This is not a requirement for legibility since achieving the goal is encoded into the objective of the behavior generation mechanism. In our case, avoidance is the primary objective, and task achievement is a constraint. Specifically, we will define our optimal legible avoidance policy as follows:

Definition 4. *For an agent model $\mathcal{M}^\alpha$, a normative contract set $\mathbb{N}$, a norm violation MDP $\mathcal{M}^{\neg\mathbb{N}}$, and a minimum value threshold ℓ an optimal legible avoidance policy ($\pi_\ell^{\mathcal{L}}$) is one that is an optimal solution for the given optimization problem:*

$$\min_\pi \quad \sum_s V_\pi^{\mathcal{L}}(s),\ \textit{s.t.}\ V_\pi^\alpha(s_0) \geqslant \ell$$

where V_π^Λ is the value of the policy as evaluated in the model $\mathcal{M}^{\mathcal{L}} = \langle S, A, R^{\mathcal{L}}, P, \gamma^{\mathcal{L}}\rangle$, such that $R^{\mathcal{L}}(s, a) = -1 * Q_*^{\neg\mathbb{N}}(s, a)$ and $\gamma^{\mathcal{L}} = 0$.

Note that maximizing for $-1 * Q_*^{\neg\mathbb{N}}(s, a)$ means it is always selecting the action with the lowest Q value. Additionally, a discount factor of zero means that this is done myopically at every time step. Note that here we try to maximize the sum of values across all states, rather than just the initial one, so the agent chooses the least likely action at every state when possible. By using the sum, we account for the fact that in the new MDP, the discount factor is zero.

Theorem 2. *For an agent model $\mathcal{M}^\alpha$ and a normative contract set $\mathbb{N}$, there exists a threshold ℓ for which the legible avoidance policy ($\pi_\ell^{\mathcal{L}}$) is a signaling policy.*

Proof Sketch. One could trivially show it to be true by setting ℓ to zero. This would involve the agent always selecting actions that reduce the likelihood of violating the norm, and hence it would reduce the likelihood of norm violation. □

However, implementing this optimization problem in practice is a challenge since it requires us to simultaneously evaluate a given policy across two different MDPs with different reward functions and different discount factors. To do so, we will leverage a mixed integer linear programming (MILP) formulation that will maintain two different sets of occupancy frequencies, one for each MDP. We will express our objective in one, and the constraints in the other. We will then add constraints that will force the two occupancy frequencies to correspond to the same deterministic policy. This latter constraint requires us to use integer variables, hence moving us into the realm of MILP as opposed to the traditional linear programming formulations used in MDPs. Specifically, the encoding is given as follows:

$$
\begin{aligned}
\max_{x^{\mathcal{L}}, x^{\alpha}, y} \quad & \sum_{s \in S} \sum_{a \in A} x^{\mathcal{L}}_{s,a} \left(R^{\mathcal{L}}\right) \\
\text{s.t.} \quad & x^{\mathcal{L}}_{s,a} \geqslant 0, \;\; x^{\alpha}_{s,a} \geqslant 0 && \forall s \in S,\; a \in A, \\
& \sum_{a \in A} x^{\mathcal{L}}_{i,a} = d_0(i) && \forall i \in S, \\
& \sum_{a \in A} x^{\alpha}_{i,a} - \gamma^{\alpha} \sum_{j \in S} \sum_{a \in A} x^{\alpha}_{j,a} P(i \mid j, a) = d_1(i) && \forall i \in S, \\
& \text{Occupancy frequency calculation} \\
& \sum_{s \in S} \sum_{a \in A} x^{\alpha}_{s,a} R_{\text{pos}}(s,a) \geqslant \ell, \\
& \text{Constraint to ensure minimum value} \\
& \sum_{a \in A} y_{s,a} = 1 && \forall s \in S, \\
& y_{s,a} \in \{0,1\} && \forall s \in S,\; a \in A, \\
& x^{\mathcal{L}}_{s,a} \leqslant y_{s,a}, \;\; x^{\alpha}_{s,a} \leqslant y_{s,a} && \forall s \in S,\; a \in A. \\
& \text{Constraints to ensure the occupancy frequency is defined over the same policy}
\end{aligned}
\tag{2}
$$

Here, $x^{\mathcal{L}}$ corresponds to the occupancy frequency of the policy in the model $\mathcal{M}^{\mathcal{L}}$ and x^{α} that in the agent model $\mathcal{M}^{\alpha}$. The integer variable $y_{s,a}$ ensures that only one state action pair has positive occupancy frequency for both sets of occupancy frequencies, and it needs to be less than one. Additionally, d_0 and d_1 represent the initial state distribution set. Which, in the case of $\mathcal{M}^{\mathcal{L}}$, will be a uniform distribution across states, and in the case $\mathcal{M}^{\alpha}$, will be zero in all states except the initial state s_0. This brings us to our main theorem in this section.

Theorem 3. *The occupancy frequency that is part of the optimal solution for the MILP formulation defined in Equation 2, will correspond to an optimal legible policy.*

Here, the proof follows from the construction of the MILP. However, the reader can find a more detailed discussion of the encoding in the supplementary materials.

7 Evaluation

To evaluate our approach, we conducted experiments on a set of standard Markov Decision Process (MDP) planning benchmarks and also ran a user study.

7.1 Experiments on MDP Planning Benchmarks

Setting We evaluated three policy formulations: Optimal Behavior (standard dual occupancy LP), Preemptive Commitment ($Q^{\neg\mathbb{N}}_{*}$-based reward shaping LP), and Legible Avoidance (MILP) across four grid-world domains. Our focus on grid-world domains was primarily motivated by the fact that

in such domains, it was easier to control for factors like the viability of preemptive commitment. For two domains, we included a close-only door toggle that blocks access to the normative contract locations: *Safe Room* and *Corridor*. The remaining two domains (*Simple Grid* and *Puddle Grid*) offer no structural intervention. Each domain type is instantiated at two grid sizes: small (5×5 to 8×5) and large (9×9 to 14×8). We include a more detailed explanation about domain descriptions in the supplementary materials. For legible avoidance, we sweep the positive-goal threshold over $\ell \in \{0, 4, 8\}$. Each configuration is repeated 3 times to account for solver time variability. All optimization problems are formulated using PuLP (Mitchell et al., 2011) and solved with HiGHS (Huangfu & Hall, 2018). Experiments are run on a single machine with 18 GB RAM and an Apple M3 CPU.

Computational Cost Table 1 reports solver computation times across all configurations. The LP-based formulations for optimal behavior and preemptive commitment solve in under 50 ms even on the largest grids, with virtually identical runtimes. This confirms that constructing $\mathcal{M}^{PC}$ and solving its dual occupancy LP adds no meaningful overhead beyond the one-time computation of $Q_*^{\neg\mathbb{N}}$. Legible avoidance, which requires solving the MILP with binary action variables $y_{s,a}$ and two coupled occupancy measures $x^{\mathcal{L}}$ and x^{α}, is 1–3 orders of magnitude slower, ranging from 0.14 s (Puddle Grid 10×10, $\ell=0$) to 24.4 s (Puddle Grid 6×6, $\ell=4$). Despite this increase, all 120 runs converge to optimal solutions within practical time limits. MILP solve time does not scale monotonically with grid size: the 6×6 Puddle Grid at $\ell=4$ is harder for the branch-and-bound solver than the 10×10 instance, likely due to tighter combinatorial coupling among the binary variables that enforce a shared deterministic policy across both occupancy measures. The relationship between ℓ and solver difficulty is similarly non-monotonic: for Simple Grid 5×5, solve time increases from 2.3 s at $\ell=4$ to 16.8 s at $\ell=8$, indicating that stricter goal constraints can reduce the feasible region for $y_{s,a}$ and create harder combinatorial subproblems.

Table 1: Solver computation time (seconds) across domains, reported as mean $\pm$ std over 3 runs.

				Legible Avoidance		
Domain	**Grid**	**Optimal Behavior**	**Preemptive Commitment**	$\ell = 0$	$\ell = 4$	$\ell = 8$
Corridor	14×8	0.041 ± 0.000	0.042 ± 0.001	0.631 ± 0.003	3.055 ± 0.032	2.398 ± 0.015
	8×5	0.008 ± 0.000	0.008 ± 0.000	2.297 ± 0.019	6.781 ± 0.179	3.809 ± 0.006
Puddle Grid	10×10	0.011 ± 0.000	0.012 ± 0.000	0.145 ± 0.004	0.167 ± 0.002	0.237 ± 0.003
	6×6	0.005 ± 0.002	0.003 ± 0.000	0.137 ± 0.002	24.430 ± 0.234	0.727 ± 0.010
Safe Room	12×8	0.031 ± 0.000	0.032 ± 0.000	0.727 ± 0.019	2.082 ± 0.020	1.687 ± 0.030
	7×5	0.007 ± 0.000	0.007 ± 0.000	0.864 ± 0.005	1.947 ± 0.005	2.188 ± 0.013
Simple Grid	9×9	0.009 ± 0.000	0.010 ± 0.002	0.482 ± 0.009	0.400 ± 0.013	0.605 ± 0.004
	5×5	0.002 ± 0.000	0.002 ± 0.000	5.454 ± 0.140	2.280 ± 0.008	16.772 ± 0.145

Sensitivity to the Threshold ℓ. The positive-goal threshold ℓ governs the trade-off between the avoidance objective and task completion in the legible avoidance formulation. The relationship between ℓ and solver difficulty is non-monotonic: for Simple Grid 5×5, solver time increases from 2.3 s at $\ell=4$ to 16.8 s at $\ell=8$, suggesting that stricter goal constraints can create harder combinatorial subproblems for the MILP by reducing the feasible region for the binary variables $y_{s,a}$.

Note on Preemptive Commitment. As expected for Puddle and Simple Grid, the lack of any structural intervention meant the agent could not achieve a state in which the likelihood of violating the norm was zero.

7.2 User Study

Next, we discuss our IRB-approved user study to evaluate the different signaling behaviors.

7.2.1 Hypotheses

As part of the user study, we tested five primary hypotheses. We expected participants to report higher overall trust in AI agents exhibiting signaling behaviors, specifically preemptive commitment (H1-a) or legible avoidance (H1-b), compared to those following optimal behavior. Between

these two signaling behaviors, we expected legible avoidance to be perceived as more upfront about intent (H2), lead to higher reported clarity (H3), and result in a stronger perception of safety (H4) than preemptive commitment. Finally, we hypothesized that participants are unlikely to intervene or stop the robot's operation until it begins performing demonstrably damaging actions (H5). Each hypothesis was informed by pilot studies conducted at our institution. We assessed each hypothesis using Likert scale questionnaires, and in the case of trust, we specifically leveraged Muir questionnaire (Muir, 1994), a well-established psychometric tool to evaluate subjective perception of trust.

7.2.2 Study Design and Participants

We designed three different cases to compare the three behaviors. We crafted cases relevant to potential real-world robotics applications that are understandable to laypeople without an extensive tutorial. All three cases are in the grid world, but have different configurations, and each has a different robot: (a) robot cleaner, (b) robot delivery, and (c) robot collector, as can be seen in Figure 2. In each environment, the robot must go to the goal while avoiding the valuable item. For each robot case, we created simulations for the three different behaviors: (a) optimal behavior (condition OB), (b) legible avoidance behavior (condition LA), and (c) preemptive commitment (condition PC). The study design is a repeated-measures (within-subjects) design with three levels (condition OB, LA, and PC). Each participant saw one example from each case corresponding to a unique condition. Each participant was asked to take the role of a caretaker of some workspaces, tasked with protecting a valuable item. They are told that a robot has been deployed in the environment by an unknown person with unknown intent. They are expected to observe the robot's behavior and stop it if they believe the robot may harm the objects they are caring for.

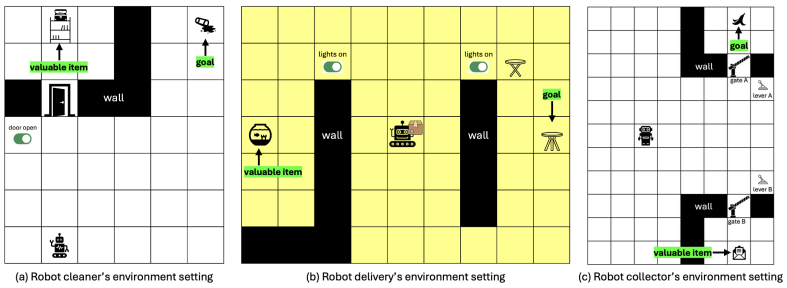

Figure 2: Robots' environment setting.

We recruited a total of 30 participants from a pool of users who identified English as their native language on the Prolific platform (Palan & Schitter, 2018). They were paid 6 USD for completing the study that takes approximately 20 minutes. Multiple attention-check questions were included throughout the study to ensure that participants were paying attention. The order in which the participants saw the conditions was randomized to ensure counterbalancing. The supplementary section includes additional information regarding the study conditions and specific questions used.

7.2.3 Results and Discussions

The main dependent measures in the user study fall into four categories: trust-related (predictability, reliability, faith, overall trust), intention perception (damage intention, hiding intentions), behavioral clarity (clarity, ability to predict), and normative compliance (rule compliance, safety perception). Figure 3 shows box plots of the results for these metrics. Compared with the other two conditions, legible avoidance behavior achieves relatively higher performance in trust, behavioral clarity, and normative compliance. Participants also reported relatively lower scores on intention perception metrics when the robot demonstrated legible avoidance, which indirectly supports trust in the robot.

One-factor repeated-measures ANOVA was performed to compare the effect of the robot's behavior according to the condition (OB, PC, LA) on each dependent variable. The statistical result is

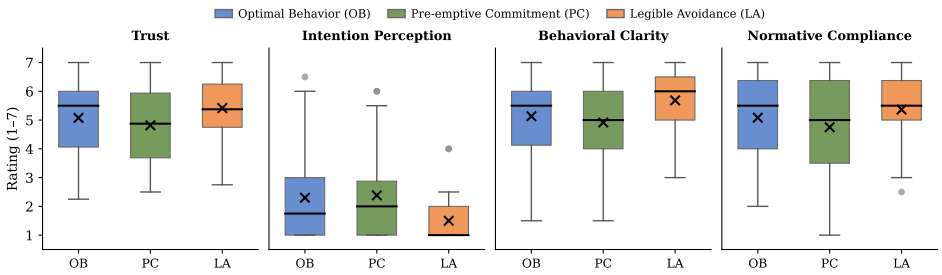

Figure 3: Box plots of conditions by dependent variables. 'X' represents the mean.

Table 2: Statistical analysis results from the user study. Values reported as mean (SD) across conditions. Post-hoc pairwise comparisons use paired t-tests with Bonferroni correction.

	Dependent	Condition					Pairwise p-values		
Category	**Variable**	**OB**	**PC**	**LA**	$F(2,58)$	p	**OB–PC**	**OB–LA**	**PC–LA**
Trust	Predictability	5.03 (1.43)	5.00 (1.14)	5.57 (1.14)	3.94	.025*	1.000	.047*	.038*
	Reliability	5.37 (1.40)	4.90 (1.56)	5.73 (1.11)	4.99	.010*	.269	.330	.028*
	Faith	5.20 (1.58)	4.93 (1.70)	5.43 (1.30)	2.32	.107	—	—	—
	Overall trust	4.70 (1.56)	4.43 (1.83)	4.93 (1.55)	2.14	.127	—	—	—
Intention perception	Damage intention	2.40 (1.96)	2.47 (1.93)	1.40 (0.72)	5.94	.004*	1.000	.011*	.011*
	Hide intentions	2.20 (1.63)	2.30 (1.66)	1.60 (1.00)	4.92	.011*	1.000	.040*	.046*
Behavioral clarity	Clarity	5.20 (1.42)	5.03 (1.54)	5.67 (1.21)	3.19	.049*	1.000	.239	.118
	Ability to predict	5.07 (1.44)	4.80 (1.49)	5.70 (1.15)	6.99	.002*	1.000	1.000	.002*
Normative compliance	Rule compliance	5.67 (1.42)	5.20 (1.85)	5.90 (1.12)	$3.70^{\dagger}$	.048*	.299	.495	.114
	Safety perception	4.50 (1.87)	4.30 (1.93)	4.83 (1.62)	$2.58^{\dagger}$	.097	—	—	—

Note. * significant at $\alpha = .05$. † Greenhouse–Geisser corrected (Mauchly's $p < .05$). — = post-hoc not conducted ($p > .05$).

shown in Table 2. From 10 dependent variables, the following variables show significant differences: predictability, reliability, damage intention, hide intentions, clarity, ability to predict, and rule compliance. We performed a post-hoc test using a paired t-test with Bonferroni correction for these variables to examine which pairs differ significantly. The result can be seen in Table 2. Additionally, Table 3 summarizes support for each hypothesis based on the statistical tests.

Table 3: Summary of statistical support for hypotheses

Hypothesis	**Dependent variable**	**Statistic**	**p**	**Interpretation**
H1 (H1a, H1b)	Overall trust (PC vs OB, LA vs OB)	F(2,58) = 2.14	0.127	Not supported
H2	Hiding intentions (LA vs PC)	t(29) = -2.57	0.046*	Supported
H3	Clarity (LA vs PC)	t(29) = 2.16	0.118	Not supported
H4	Safety perception (LA vs PC)	t(29) = 1.92	0.196	Not supported
H5	Stopping the robot	OB = 13.33% stopped PC = 26.67% stopped LA = 6.67% stopped	-	Supported

Trust. It should be noted that the current consensus in psychology is that trust is not a single-dimensional metric. In fact, the trust questionnaire we used (Muir, 1994) asks questions about the individual dimensions (predictability, reliability, faith), in addition to the overall trust, which is only intended to capture a coarse, aggregated view of the user's trust. The "faith" item and the single "overall trust" rating did not show significant effects of condition. In terms of hypotheses, H1 (H1-a and H1-b), which predicted higher "overall trust" in PC and LA than in OB, was not statistically supported, even though LA had a relatively higher mean than OB. While we did not notice a statistically significant improvement for "overall trust", we were able to establish it for two important dimensions of trust, namely "predictability" and "reliability", especially for legible avoidance. It is also worth noting that, even for "overall trust", the reported average for legible

avoidance was higher, so we might see more significant differences as we shift to larger population sizes and/or higher-stakes settings.

Intention perception. Both "damage intention" and "hiding intentions" showed significant main effects of condition. Post-hoc tests revealed that LA has significantly lower perceived intent to damage the valuable item than both OB and PC. Similarly, participants rated the robot as significantly less likely to hide its intentions in the LA condition than in PC. This finding directly supports H2, which predicted lower perceived hiding of intentions in LA than in PC. These results indicate that when the robot explicitly routes around the valuable item in a legible manner, participants infer clearer intentions of the robot.

Behavioral clarity. The "clarity" metric, which states "The robot's behavior made its goal clear to me," showed a significant main effect of condition. LA achieves the highest mean rating, whereas PC achieves the lowest. However, post-hoc comparisons between LA and PC did not reach significance after Bonferroni correction. This means that H3 is not supported. In contrast, in the "ability to predict" metric, participants reported being significantly more able to predict the robot's next action in LA than in both OB and PC.

Normative compliance. There is a mixed pattern in normative compliance categories. Rule compliance differed significantly across conditions. The highest average rating was obtained in the LA condition, and the lowest in the PC condition. Although the post-hoc comparisons did not reveal pairwise differences after Bonferroni correction, the mean value suggests that participants perceived the legible avoidance strategy as more closely adhering to the workspace rules for protecting the valuable item. Safety perception, in contrast, did not differ significantly across conditions. Thus, H4, which predicted higher safety perception in LA than in PC, was not statistically supported.

Stopping the robot. The robot's stopping rate was relatively low in OB (13.33%) and LA (6.67%), and slightly higher in PC (26.67%). This result supports H5, in which most participants did not stop the robot until it was actually performing damaging actions. In preemptive commitment behavior, the robot chose a path that brought it closer to the valuable item, and this costly motion may be the main reason participants stop the robot. However, 73.33% participants who did not stop the robot in the preemptive commitment condition suggest that they were willing to tolerate ambiguous or even mildly concerning behavior rather than preemptively interrupting the robot.

Limitations of Study Please note that our studies focused on low-stakes grid-world cases. One might expect participants to be more risk-averse when violating a norm could lead to catastrophic outcomes. As such, one could obtain different results when applying it to high-stakes scenarios.

8 Conclusions and Future Work

We focus on the use of signaling behavior to communicate agents' intent in low-trust environments. As part of this effort, we provide general definitions of a low-trust environment and signaling behavior. We also identify two specific types of signaling behavior and propose algorithms to generate them. The computational evaluation of these strategies across standard MDP planning shows that while both strategies are computationally feasible, legible avoidance has a substantially higher computational cost due to its MILP formulation (as opposed to the LP formulation used for pre-emptive commitment). Across domains and grid sizes, solve times were 1-3 orders of magnitude slower than the LP-based methods. Meanwhile, in the user study results, legible avoidance outperformed both optimal behavior and preemptive commitment across multiple dimensions. Participants consistently rated it as more predictable, reliable, and rule-compliant, and perceived it as having lower harmful intent. It is worth noting that this trade-off is sometimes observed in real-world instances of such behavior, where people trying to dispose of dangerous objects may be perceived as threats.

As the next step, we plan on developing learning methods that could generate such behaviors. Additionally, future work will evaluate real-world trade-offs (e.g., when valuable items pose physical hazards) and validate findings through physical robot deployments across varied configurations. This would bridge simulation efficiency to practical deployment readiness.

Acknowledgments

Dr. Sreedharan's research is supported by NSF through grant 2303019. This material is based in part upon work supported by Other Transaction award 1AY2AX000062 from the U.S. Advanced Research Projects Agency for Health (ARPA-H) Platform Accelerating Rural Access to Distributed Integrated Medical Care (PARADIGM) program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

Saleh Afroogh, Ali Akbari, Emmie Malone, Mohammadali Kargar, and Hananeh Alambeigi. Trust in ai: progress, challenges, and future directions. *Humanities and Social Sciences Communications*, 11(1):1568, 2024.

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 *Proceedings of the international conference on automated planning and scheduling*, volume 29, pp. 86–96, 2019.

Justin C Cheung and Shirley S Ho. The effectiveness of explainable ai on human factors in trust models. *Scientific Reports*, 15(1):23337, 2025.

Brian L Connelly, S Trevis Certo, R Duane Ireland, and Christopher R Reutzel. Signaling theory: A review and assessment. *Journal of management*, 37(1):39–67, 2011.

Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 301–308. IEEE, 2013.

Alan SR Fermin, Toko Kiyonari, Yoshie Matsumoto, Haruto Takagishi, Yang Li, Ryota Kanai, Masamichi Sakagami, Rei Akaishi, Naho Ichikawa, Masahiro Takamura, et al. The neuroanatomy of social trust predicts depression vulnerability. *Scientific reports*, 12(1):16724, 2022.

Andrea Ferrario and Michele Loi. How explainability contributes to trust in ai. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 1457–1466, 2022.

Drew Fudenberg and Kevin He. Learning and type compatibility in signaling games. *Econometrica*, 86(4):1215–1255, 2018.

Ella Glikson and Anita Williams Woolley. Human trust in artificial intelligence: Review of empirical research. *Academy of management annals*, 14(2):627–660, 2020.

Victoria Gonzalez, Laura Amo, and Sanjukta Das Smith. Building trust in ai: Exploring the impact of ai competence framing. In *Proceedings of the annual Hawaii international conference on system sciences*, pp. 526–535, 2025.

The Anh Han, Cedric Perret, and Simon T Powers. When to (or not to) trust intelligent machines: Insights from an evolutionary game theory analysis of trust in repeated games. *Cognitive Systems Research*, 68:111–124, 2021.

Peter A Hancock, Deborah R Billings, Kristin E Oleson, Jessie Y Chen, Ewart De Visser, and Raja Parasuraman. A meta-analysis of factors influencing the development of human-robot trust. 2011.

Qi Huangfu and JA Julian Hall. Parallelizing the dual revised simplex method. *Mathematical Programming Computation*, 10(1):119–142, 2018.

Noel D Johnson and Alexandra A Mislin. Trust games: A meta-analysis. *Journal of economic psychology*, 32(5):865–889, 2011.

John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

David Lewis. *Convention: A philosophical study*. John Wiley & Sons, 2008.

Poornima Madhavan and Douglas A Wiegmann. Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, 8(4):277–301, 2007.

Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

Stuart Mitchell, Michael OSullivan, and Iain Dunning. Pulp: a linear programming toolkit for python. *The University of Auckland, Auckland, New Zealand*, 65:25, 2011.

Bonnie M Muir. Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37(11):1905–1922, 1994.

Stefan Palan and Christian Schitter. Prolific. ac—a subject pool for online experiments. *Journal of behavioral and experimental finance*, 17:22–27, 2018.

Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

Pascal Poupart. *Exploiting structure to efficiently solve large scale partially observable Markov decision processes*. University of Toronto Toronto, 2005.

Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.

Jiaqi Qin, Yuanfeixue Nan, Zichao Li, and Jingbo Meng. Effectiveness of communication competence in ai conversational agents for health: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 27:e76296, 2025.

Julian B Rotter. A new scale for the measurement of interpersonal trust. *Journal of personality*, 1967.

Denise M Rousseau, Sim B Sitkin, Ronald S Burt, and Colin Camerer. Not so different after all: A cross-discipline view of trust. *Academy of management review*, 23(3):393–404, 1998.

Stuart Russell, Peter Norvig, and Artificial Intelligence. A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, 25(27):79–80, 1995.

Yichuan Tang, Ruslan Salakhutdinov, and Geoffrey Hinton. Robust boltzmann machines for recognition and denoising. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 2264–2271. IEEE, 2012.

Daniel Ullman and Bertram F Malle. Human-robot trust: Just a button press away. In *Proceedings of the companion of the 2017 ACM/IEEE international conference on human-robot interaction*, pp. 309–310, 2017.

DPS Verma and Soma Sen Gupta. Does higher price signal better quality? *Vikalpa*, 29(2):67–78, 2004.

Zachary Wojtowicz and Simon DeDeo. Undermining mental proof: How ai can make cooperation harder by making thinking easier. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1592–1600, 2025.

Zahra Zahedi, Sarath Sreedharan, and Subbarao Kambhampati. A mental model based theory of trust. *arXiv preprint arXiv:2301.12569*, 2023.

Supplementary Materials

The following content was not necessarily subject to peer review.

A MILP Formulation with Binary Action Variables and Two Occupancy Measures

Setup and Q-values

We consider a finite MDP with:

- State set S,
- Action set A,
- Transition probabilities $P(s' \mid s, a)$ for all $s, s' \in S$, $a \in A$.

Let $d_0(s)$ be the initial state distribution for the first occupancy measure; in our implementation we take it as uniform:

$$d_0(s) \;=\; \frac{1}{|S|} \quad \forall s \in S.$$

We also define a (different) initial distribution $d_1(s)$ for the second occupancy measure, concentrated on the true start state s_{start}:

$$d_1(s) \;=\; \begin{cases} 1 & \text{if } s = s_{\text{start}}, \\ 0 & \text{otherwise.} \end{cases}$$

We assume that we have already solved the norm-violating MDP $\mathcal{M}^{-\mathbb{N}}$, which produced optimal Q-values

$$Q_*^{-\mathbb{N}}(s, a) \quad \text{for all } s \in S,\ a \in A,$$

which encode how attractive each action is under the norm-violating reward. In particular, $Q_*^{-\mathbb{N}}(s, a)$ is large and positive for norm-violating actions and for actions leading towards norm violations.

In the new optimization problem, we define the legibility reward as $R^{\mathcal{L}}(s, a) = -1 * Q_*^{-\mathbb{N}}(s, a)$, so the resulting policy is discouraged from taking actions associated with norm violation.

Positive action reward

We define a reward function for the agent's original task:

$$R_{\text{pos}}(s, a) \;=\; \begin{cases} R_{\text{pos}} & \text{if } (s, a) \text{ corresponds to task completion,} \\ 0 & \text{otherwise,} \end{cases}$$

where $R_{\text{pos}} > 0$ is a constant.

Let $\gamma^{\alpha} \in (0, 1)$ be the discount factor used for the agent's occupancy measure. We also fix a threshold $\ell > 0$ which specifies how much discounted positive reward we require.

Decision variables

We introduce three groups of decision variables:

- Binary action-selection variables

$$y_{s,a} \in \{0, 1\} \quad \forall s \in S,\ a \in A.$$

Here $y_{s,a} = 1$ means that, in state s, the policy selects action a.

- Legibility occupancy measure
$$x^{\mathcal{L}}_{s,a} \geqslant 0 \quad \forall s \in S,\, a \in A.$$
This will be used in the objective, together with $R^{\mathcal{L}}(s,a)$.
- Agent task occupancy measure
$$x^{\alpha}_{s,a} \geqslant 0 \quad \forall s \in S,\, a \in A.$$
This will be used in the constraint involving R_{pos} and ℓ.

Both occupancy measures are linked to the binary variables $y_{s,a}$ by upper bound constraints of the form $x^k_{s,a} \leqslant y_{s,a}$, so an occupancy measure can be positive at (s,a) only if the policy actually selects action a in state s.

Objective

The objective maximizes the legibility reward under the first occupancy measure $x^{\mathcal{L}}$:

$$\max_{x^{\mathcal{L}}, x^{\alpha}, y} \quad \sum_{s \in S} \sum_{a \in A} x^{\mathcal{L}}_{s,a}\,(R^{\mathcal{L}}).$$

Since $R^{\mathcal{L}}(s,a) = -1 * Q^{-\mathbb{N}}_{*}(s,a)$, this objective discourages the use of norm-violating actions and other actions with high $Q^{-\mathbb{N}}_{*}$ values.

Constraints

1. Deterministic policy. We enforce that the policy is deterministic and stationary: in each state s exactly one action is chosen:

$$\sum_{a \in A} y_{s,a} \;=\; 1 \quad \forall s \in S,$$

with $y_{s,a} \in \{0,1\}$ as defined above.

2. Legibility occupancy flow (no discount). The legibility occupancy measure $x^{\mathcal{L}}$ is tied directly to the initial distribution d_0. For every state i,

$$\sum_{a \in A} x^{\mathcal{L}}_{i,a} \;=\; d_0(i) \quad \forall i \in S.$$

3. Agent task occupancy flow (discount γ^{α}). The agent task occupancy measure x^{α} satisfies the discounted flow constraints with discount factor γ^{α} and initial distribution d_1:

$$\sum_{a \in A} x^{\alpha}_{i,a} \;-\; \gamma^{\alpha} \sum_{j \in S} \sum_{a \in A} x^{\alpha}_{j,a}\, P(i \mid j, a) \;=\; d_1(i) \quad \forall i \in S.$$

Intuitively, x^{α} describes the discounted state–action visitation frequencies starting from s_{start} (via d_1) and following the policy induced by y.

4. Linkage between occupancies and action-selection variables. For both occupancy measures, we require that positive occupancy is allowed only where the policy selects the corresponding action:

$$x^{\mathcal{L}}_{s,a} \;\leqslant\; y_{s,a}, \quad x^{\alpha}_{s,a} \;\leqslant\; y_{s,a} \quad \forall s \in S,\, a \in A.$$

Thus $x^{\mathcal{L}}_{s,a}$ and $x^{\alpha}_{s,a}$ must be zero whenever $y_{s,a} = 0$.

5. Positive-action reward constraint. Finally, we require that under the agent task occupancy measure x^{α}, the total discounted positive reward must meet a minimum threshold $\ell > 0$:

$$\sum_{s \in S} \sum_{a \in A} x^{\alpha}_{s,a}\, R_{\text{pos}}(s,a) \;\geqslant\; \ell.$$

Final MILP

Collecting everything, the final mixed-integer linear program is:

$$
\begin{aligned}
\max_{x^{\mathcal{L}}, x^{\alpha}, y} \quad & \sum_{s \in S} \sum_{a \in A} x^{\mathcal{L}}_{s,a} \, (R^{\mathcal{L}}) \\
\text{s.t.} \quad & x^{\mathcal{L}}_{s,a} \geqslant 0, \;\; x^{\alpha}_{s,a} \geqslant 0 && \forall s \in S, \; a \in A, \\
& \sum_{a \in A} x^{\mathcal{L}}_{i,a} = d_0(i) && \forall i \in S, \\
& \sum_{a \in A} x^{\alpha}_{i,a} - \gamma^{\alpha} \sum_{j \in S} \sum_{a \in A} x^{\alpha}_{j,a} \, P(i \mid j, a) = d_1(i) && \forall i \in S, \\
& \sum_{s \in S} \sum_{a \in A} x^{\alpha}_{s,a} \, R_{\text{pos}}(s, a) \geqslant \ell, \\
& \sum_{a \in A} y_{s,a} = 1 && \forall s \in S, \\
& y_{s,a} \in \{0, 1\} && \forall s \in S, \; a \in A, \\
& x^{\mathcal{L}}_{s,a} \leqslant y_{s,a}, \;\; x^{\alpha}_{s,a} \leqslant y_{s,a} && \forall s \in S, \; a \in A.
\end{aligned}
\tag{3}
$$

B Domain Descriptions

We evaluate our framework on four grid-world domains, each instantiated at two grid sizes (small and large). In every domain, the agent must navigate the grid by choosing among four movement actions (North, East, South, West) plus any available local special actions. Movement is deterministic ($\epsilon = 0$). A discount factor of $\gamma = 0.9$ is used throughout. Table 4 summarizes the key characteristics.

Table 4: Summary of domain characteristics

Domain	Grid Sizes	Positive Goal	Reward	Negative Goal	Penalty	Living Reward	Cell Penalties	Toggle
Simple Grid	5×5, 9×9	COLLECT	$+10$	TRAP	-10	-1.0	—	No
Puddle Grid	6×6, 10×10	REACH_GOAL	$+10$	FALL_PIT	-10	0.0	Puddles (-1)	No
Safe Room	7×5, 12×8	GET_PACKAGE	$+10$	TOUCH_WIRE	-10	-1.0	—	Yes
Corridor	8×5, 14×8	CLEAN_AREA	$+10$	SPILL_CHEMICAL	-10	0.0	Debris (-1)	Yes

Simple Grid. The agent operates on a grid with scattered wall cells and must navigate from its start position to a collection point (COLLECT, $r = +10$) while avoiding a trap cell (TRAP, $r = -10$). A per-step living reward of -1.0 encourages the agent to reach the goal quickly. No toggle mechanism or cell-level penalties exist.

Puddle Grid. Similar to Simple Grid, the agent must reach a goal cell (REACH_GOAL, $r = +10$) while avoiding a pit (FALL_PIT, $r = -10$). The key distinction is the presence of *puddle cells* that impose a -1.0 penalty upon entry, creating regions the agent should route around. The living reward is 0.0, so there is no urgency to minimize path length beyond avoiding puddles. No toggle mechanism is available.

Safe Room. The grid is divided into two regions by a wall with a single *door* (initially open). The agent must reach a package (GET_PACKAGE, $r = +10$) located on one side while avoiding a wire hazard (TOUCH_WIRE, $r = -10$) on the other side of the door. A *close-only toggle switch* is available at a designated cell: activating it permanently closes the door, blocking passage and thereby preventing access to the hazard region. A living reward of -1.0 encourages efficient navigation.

Corridor. The agent navigates a corridor-shaped grid divided by a wall with a single door (initially open). The positive goal is to reach a cleaning station (CLEAN_AREA, $r = +10$), while the negative

outcome is a chemical spill (SPILL_CHEMICAL, $r = -10$) located beyond the door. *Debris cells* impose a -1.0 step penalty on entry, affecting route choice. A close-only toggle switch allows the agent to permanently seal the door. The living reward is 0.0. This domain combines the toggle mechanism of Safe Room with the cell-level penalties of Puddle Grid.

C User Study Questions

To assess the hypotheses, we asked participants to answer several questions and rate them on a 7-point Likert scale (see Table 5).

Table 5: Overview of questions and statements

Categories	Dependent Variables	User Study Questions/Statements	7-point Scale
Trust	Predictability	To what extent can the robot's behavior be predicted from moment to moment?	1-Low, 7-High
	Reliability	To what extent can you count on the robot to do its job?	1-Low, 7-High
	Faith	What degree of faith do you have that the robot will be able to cope with similar situations in the future?	1-Low, 7-High
	Overall trust	Overall, how much do you trust the robot?	1-Low, 7-High
Intention perception	Damage intention	To what degree did you think that the robot was trying to harm the valuable item?	1-Low, 7-High
	Hiding intentions	To what extent do you think the robot is trying to hide its intentions from you?	1-Low, 7-High
Behavioral clarity	Clarity	The robot's behavior made its goal clear to me.	1-Strongly Disagree, 7-Strongly Agree
	Ability to predict	I was able to predict the robot's next action.	1-Strongly Disagree, 7-Strongly Agree
Normative compliance	Rules compliance	The robot follows and complies with the rules given in this workspace.	1-Strongly Disagree, 7-Strongly Agree
	Safety perception	I feel it is safe to let the robot act autonomously.	1-Strongly Disagree, 7-Strongly Agree

D List of Combinations of Cases and Conditions

Table 6 provides more detailed information about the variation of scenarios in the user study.

Table 6: Combination of cases and conditions

Scenario Number	Case		
	Robot Cleaner	**Robot Delivery**	**Robot Collector**
1	Optimal behavior	Legible avoidance	Preemptive commitment
2	Optimal behavior	Preemptive commitment	Legible avoidance
3	Legible avoidance	Optimal behavior	Preemptive commitment
4	Legible avoidance	Preemptive commitment	Optimal behavior
5	Preemptive commitment	Optimal behavior	Legible avoidance
6	Preemptive commitment	Legible avoidance	Optimal behavior

E Participants' Demographics for the User Study

Table 7 provides more detailed information about the participants' demographics from the user study that we conducted.

F Additional Resources

Here are some of the additional resources, including code and an example of the survey: Planning for Signaling - Repository.

Table 7: Participants' demographics for the user study

Demographic Variable	**Category**	**Frequency**
Age	18-24 years old	2
	25-34 years old	8
	35-44 years old	8
	45-54 years old	7
	55-64 years old	3
	65+ years old	2
Gender	Man	16
	Woman	13
	Non-binary	1
Highest level of education	High school or equivalent	3
	Attended college/university	7
	Associate degree	3
	Bachelor's degree	8
	Master's degree	8
	Doctorate degree	1
Have taken Computer Science courses	Yes	7
	No	23
Have taken AI courses	Yes	3
	No	27