Planning for Human–Robot Interaction in Socially Situated Tasks

The Impact of Representing Time and Intention

Frank Broz · Illah Nourbakhsh · Reid Simmons

Published online: 12 April 2013

Abstract This article presents the results of a study on the effects of representing time and intention in models of socially situated tasks on the quality of policies for robot behavior. The ability to reason about how others' observable actions relate to their unobservable intentions is an important part of interpreting and responding to social behavior. It is also often necessary to observe the timing of actions in order to disambiguate others' intentions. Therefore, our proposed approach is to model these interactions as time-indexed partially observable Markov decision processes (POMDPs). The intentions of humans are represented as hidden state in the POMDP models, and the time-dependence of actions by both humans and the robot are explicitly modelled. Our hypothesis is that planning for these interactions with a model that represents time dependent action outcomes and uncertainty about others' intentions will achieve better results than simpler models that make fixed assumptions about people's intentionality or abstract away time-dependent effects. A driving interaction governed by social conventions and involving ambiguity in the other driver's intent was used as the scenario with which to test this hypothesis. A robot car controlled by policies from time-dependent POMDP models or by policies from two less expressive model variants performed this interaction in a driving simulator with human drivers. The time-dependent POMDP policies achieved better results than those of the models without explicit time representation or human intention as hidden state, both according to the reward obtained and to people's subjective impressions of how socially appropriate and natural the robot's behavior was. These results demonstrate both the relative superiority of these representation choices and the effectiveness of this approach to planning for socially situated tasks.

Keywords Planning · Social interaction · Intention · Time · POMDP · Human subject experiment

1 Motivation

Humans are social animals. It is hard to imagine thinking that a person has achieved basic competency at a task if they can perform its actions only when the environment is cleared of other people. This obvious statement has deep implications for the design of robots whose purpose is to carry out tasks in everyday human environments. Almost any task performed in a populated environment will involve some amount of interaction with people acting according to their own intentions. In order to be useful, such robots must have a way to create policies that respect human behavioral norms.

An autonomous robot must pursue its own goals in a shared environment where people are acting according to existing guidelines for behavior while pursuing goals that may differ from or even in conflict with its own. A robot needs its own models of social behavior to be able to respond

F. Broz (✉)
School of Computing and Mathematics, Plymouth University,
Plymouth, PL4 8AA, UK
e-mail: frank.broz@plymouth.ac.uk

I. Nourbakhsh · R. Simmons
Robotics Institute, Carnegie Mellon University, Pittsburgh,
15213, USA

I. Nourbakhsh
e-mail: illah@ri.cmu.edu

R. Simmons
e-mail: reids@ri.cmu.edu

intelligently in these ambiguous situations. In most cooperative and many adversarial domains, the goals of other agents are assumed to be known. In a domain where the agents are pursuing one of a number of possible goals, the intent of another agent may not immediately be clear. By observing a human's actions during a social task, their intention can be inferred, which will improve a robot's ability to coordinate its actions and achieve its own intended goals. In order to preserve the patterns of behavior that allow people to coordinate their actions to achieve their individual goals, a robot must act in a way that conforms to the rest of the existing system. This leads to a criteria for successful behavior that is broader and more holistic than the more common notion associated only with a single agent's goal achievement. Behaviors that are myopically successful at reaching a robot's immediate goal but violate norms for socially acceptable behavior may damage the stability of the entire system. People, losing trust that social guidelines will be followed, may also act without consideration for others, creating conflicts that reduce the opportunities for success for all the agents involved. Consider, for example, the problem of a robot that rides an elevator with people. If the robot enters the elevator without determining whether people inside it want to exit and waiting for them to do so, this breech of etiquette could lead to confusion and delays.

The mental models people have of social behavior include the role of time in the way an interaction evolves. When interactions take place in dynamic, changing environments, people may reason about time in order to coordinate their behavior with the occurrence of significant environmental events. But people are also generally concerned about the amount of time that an interaction takes in ways that influence their behavior. In the elevator riding example, whether people are willing to hold the elevator for others depends on unspoken norms relating to how long it will take a person to get on the elevator and how long that person is likely to have to wait for the next one. Because humans are constantly reasoning about time, it is important that a robot that interacts with them be able to reason about it as well. Unfortunately, in many computational models used for action selection, the explicit representation of time is abstracted away for the sake of computational efficiency. Approaches based on these models may be unable to perform satisfactorily in many social tasks.

1.1 Planning for Social Interaction

The planning paradigm on which this work is based is decision-theoretic planning using probabilistic graphical models. Probabilistic graphical models are typically used for planning in situations where there is uncertainty about the state of the world, such as the outcomes of actions an agent may take. This uncertainty is represented by probability distributions. Decision-theoretic planning creates a policy by finding the actions that maximize the reward that can be obtained according to a given model. Decision-theoretic approaches have proven themselves useful in a variety of real world planning domains, particularly in robotics, because of the flexibility of defining agents' goals in terms of a utility function and the ability to model the uncertainty of action outcomes and observations that so often arises in physical domains [22, 29].

The particular model used in this research, the partially observable Markov decision process (POMDP), is capable of representing situations in which the entire state of the world is not directly observable and making use of this hidden state information while planning [16]. The probabilistic relationship between the underlying state and aspects of the state that can be observed by the agent is included in the model. Rather than a mapping from states to actions, a POMDP policy maps beliefs (probability distributions over the states) to actions. This belief distribution is maintained during execution according to the observations that the agent receives while acting. This model is of particular interest for social interaction domains, because even if an agent were able to sense its physical surroundings perfectly, the mental states of the people with which it interacts remain unobservable.

This research is also concerned with the importance of representing time's role in social interactions. This is a challenge because most decision theoretic models for control make simplifying assumptions to represent action outcomes as dependent only on the current state for the sake of computational efficiency. Semi-Markov models (which allow more sophisticated representations of the relationship between time, state, and action) exist, but are much less widely used [26]. Their application is less straightforward than Markov models, particularly in the case of POMDPs, where finding solutions to problems of a realistic size requires the use of sophisticated heuristic algorithms that may not readily translate to semi-Markov extensions. For this reason, it is desirable to represent time-dependent action outcomes in a Markov model that can be used with existing state-of-the-art POMDP solvers.

One might question whether it is necessary to use planning to arrive at a policy for action. Shouldn't domain experts be able to hand-code policies for interaction, especially since social behaviors tend to follow regular patterns? While social interaction is regular in a way that makes it easy to describe people's roles at a high level, human behavior is still highly variable within these roles. Hand-coded policies require a great deal of tweaking to get reasonable performance from and may be difficult to reuse. A model description expressed in terms of intuitive notions such as the effects of actions and the intentions motivating behavior is used to create policies that respond reasonably to a wide range of behavior. These model descriptions are more compact and easier to specify than the policies they produce.

An approach to developing policies for interaction is to have an agent learn them directly from interacting with people using reinforcement learning methods [35]. But learning typically requires large amounts of data. It may take prohibitively long, or be prohibitively expensive, to collect enough data from humans to reach an adequate level of performance. Also, unless the interaction is one that already involves one participant taking on a teaching role, having the agent learn the interaction online may fundamentally change the way that people perform their part in the interaction versus how they would if they were interacting with a human peer. Another approach to policy learning is to learn directly from human interaction demonstrations [9]. The data requirements may be hard to meet for this technique if the tasks to be learned are difficult, as there could be a lack of examples of successful interactions even in a large dataset. By using a model for planning that is based on both expert knowledge and available data, policies for action can be found that perform well without an online learning phase and with relatively few interaction examples.

Our proposed model design includes information about the relationship between people's unobservable intentions and their observable actions as well as a rich representation of time-dependence. Representing hidden state and time-dependent action outcomes are design decisions that may dramatically increase the complexity of acquiring policies for action when compared to less expressive representations. In order to justify this increase in model complexity, it is necessary to demonstrate that it does actually have a significant positive impact on the quality of the policies obtained. In this article, the proposed representation is evaluated in comparison to simpler alternative representations in a realistic example domain.

Statement of Hypothesis: Planning for social interaction with models that represent *time* dependent action outcomes and allow for planning over uncertainty in others' *intentions* will achieve better results compared to similar models that make fixed assumptions about people's intentionality or abstract away time-dependence, demonstrated in the following ways:

- The policies will obtain higher global reward considering the outcomes for all interaction participants.
- The policies produced will be subjectively more acceptable to people who interact with the robot.

2 Background

2.1 Social Behavior and Rationality

There has also been research on social interaction that attempts to explain social behavior as arising from people's ability for rational decision-making. Drawing on ideas from decision theory and game theory, Lewis defined social conventions as a way for a population to select which strategy they will all play in a coordination game with multiple equilibria [18]. Each member of a population has a reasonable expectation about which strategy the other members will play (the strategy that they all played to achieve the equilibrium last time), allowing them to coordinate future behavior. Lewis also introduced the concept of common knowledge, which was later formally defined by Aumann [2]. Common knowledge is information known to all players, that all players know the other players also know [15]. Referring to common knowledge can short-circuit the potentially infinite recursion of trying to reason about what you know about what I know about what you know… and so on. While this research is relevant in that it conceives of social behaviors as coordinated action arising from rational decision-making, a classical game-theoretic approach relying on finding equilibria is difficult to apply to realistic problem domains. The social interactions discussed in this research would, if represented as games, be sequential stochastic games with imperfect information. Practical approaches to solving games of this type are in their very early stages [12]. This type of game theoretic modeling becomes even less applicable when one considers the time-dependence of these domains, as issues of time representation are largely unaddressed in most work on game theory.

In some human social interactions, certain players may choose a strategy that appears to be dominated in terms of the immediate payoff to the agent. This altruistic behavior in game theory is often explained by using mechanisms such as reputation to enforce cooperation [21]. There are real-life situations, however, that are clearly not governed by these mechanisms and in which seemingly altruistic behavior still occurs (any number of anonymous interactions that follow rules of etiquette, for example). In evolutionary game theory, altruistic behavior can be explained as rational if it increases the possibility of passing on some of the agents' genes through relatives who benefit even if it does not benefit that particular agent [19]. But human altruism is often based on social factors other than blood relation. Some researchers in the social sciences see this as a failure of game theory to accurately describe and predict human behavior. An article by Coleman in Behavioral and Brain Sciences proposing a "psychological game theory" and its rebuttals provide a perspective on the differing views of psychologists, biologists, philosophers, and game theorists on the ability of game theory to explain human social behavior [11]. It is important for the purposes of this research to note that there are cases in which the strategies of interacting agents seem to be in what we would intuitively think of as an equilibrium, but which would be difficult or impossible to show to be a game theoretic equilibrium using the types of world models currently used in game theory. In such cases, existing game-theoretic models are of limited utility if the goal is to accurately model people's behavior. This research sidesteps the issue of finding equilibria in social interaction domains or modeling how such equilibria arise. Instead, the focus is on modeling the social behaviors observed in a way that allows the agent to find a socially appropriate rational response. How the people arrive at the intentions that they may have for an interaction (which may include preferring altruistic goals), is outside the scope of this work.

2.2 Social Behavior and Intention

Research in cognitive psychology has addressed the relationship between human behavior and people's goals and intentions. Recognizing goal-directed intentional behavior in others seems to be a fundamental human social ability, and one that develops very early in life. In a study conducted on 18-month old children, the subjects watched an adult try and always fail to perform a variety of simple puzzle manipulation tasks [20]. The children were later able to produce the intended behavior themselves without ever witnessing a successful task completion. When the same failure behaviors were demonstrated by a simple non-humanoid manipulator, the children did not go on to exhibit the intended behavior themselves, suggesting that they did not think of the manipulator's actions as failed attempts at a goal. Intentions have been conceptualized as part of a hierarchical action production framework linking high-level desires to concrete plans for action [5]. Baldwin and Baird provide an overview on research in cognitive science that explores how and under what conditions people interpret the actions of others as intentional, using the studies discussed to support a hypothesis that people use generative models of intention-guided action to interpret human behavior and detect people's intentions [3]. Some researchers have gone so far as to state that it is people's ability to hold shared intentions (rather than just recognizing the intentions of others) that separates adult humans from children and primates and enables the richness of social behavior that human society enjoys [37]. These studies show that humans use the concept of intention to reason about and interpret the actions of others. Furthermore, they suggest that intentions are conceptualized as something distinct from (and most likely hierarchically above) immediate goals, as intentions can be recognized even when the goal most closely associated with an intention's desired outcome has not been achieved.

Intentional action has also become a subject of research in neuroscience, with researchers attempting to understand where and how we process information about other people and how it relates to how we process information about ourselves. Research indicates that not only do people recognize goal-oriented action in others, but that they reason about the goals of others in a way that is similar to the way that we reason about our own goals. Frith and Frith present experimental results that indicate that the same structures in the brain used to monitor our own behavior and represent our own goals are also used to form a model of others' mental states during social interaction [14]. In a fMRI study where subjects watched an animated human figure perform grasping actions, the reaction to goal-directed and non-goal-directed actions could be distinguished in parts of the brain known to be related to processing the actions of others [23]. Another fMRI study observed differences in brain activation when watching chasing behaviors between instances where the pursuer merely followed the target versus when it seemed to predict the target's path during pursuit [27]. These studies suggest that the human brain has mental apparatus specially devoted to recognizing intentional, goal-directed behavior. People seem to organize their understanding of the actions of others based on their intentions.

There is further experimental evidence that there is something special about how people process social interaction beyond how they process other intentional human behavior. A FMRI study where people observed others engaged in independent or social tasks suggests that there are areas of the brain concerned primarily with reasoning about intentions of people in social interactions (as opposed to more generally about intentions of isolated actors) [39].

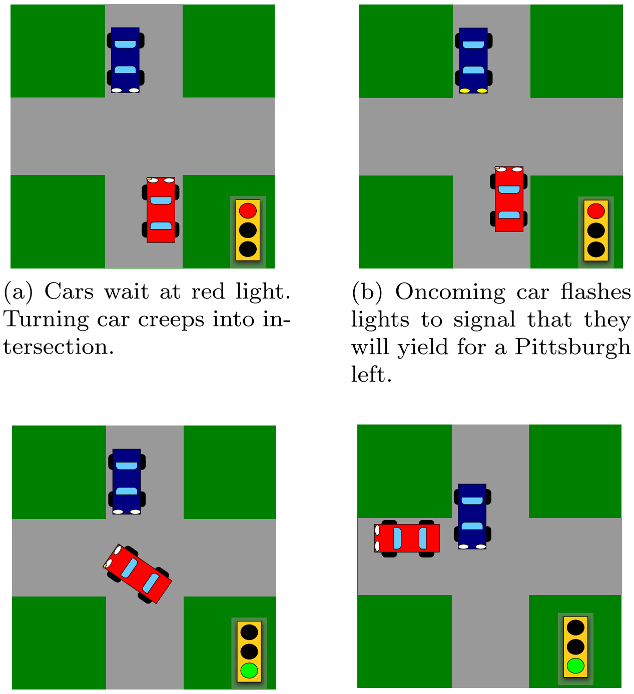
2.3 Social Behavior and Time

People are sensitive to the passage of time in social interaction. Studies indicate that even infants show lower levels of attentiveness to video feeds of their mothers if their mother's responses are delayed by a second [34]. In fact, the idea of social interaction as something that unfolds through reciprocal actions over time is so deeply ingrained that the duration and structure of an interaction with another person can influence whether or not we conceive of it as a social interaction. In one series of experiments, human subjects played coordination games where each made their decisions either simultaneously or pseudo-sequentially (meaning that they made decisions one after another but couldn't observe earlier decision makers' choices) [1]. When playing pseudo-sequentially, people were more likely to cooperate than when playing the simultaneous version of the game. The authors explained this difference in behavior as the effect of people conceptualizing the game as a social interaction when the decisions were made over time and as a game when all decisions were made simultaneously. Knowledge of timing is an important part of the knowledge about an interaction task both for predicting others' future actions and for taking appropriate actions oneself. Analysis of turn-taking behaviors in conversation have shown that the length of pauses and overlaps in these interactions is task-dependent, with different patterns for problem-solving tasks than for social chatting [4]. This suggests that in conversational tasks, the amount of time that passes before a response can provide information about the cognitive load placed on the individual, which may provide evidence to an agent about whether its statement or question to the human was expected.

A survey of studies on shared mental representations of social tasks reveals that the task models capture the way that interactions are situated in their environment, allowing people to predict other's actions based on the occurrence of external events rather than just on their previous actions [28]. It is commonly understood that people use domain knowledge about the timing of events in a dynamic environment to anticipate their occurrence. If people also use these events to predict the future actions of others, it is clear that representations that allow for reasoning about time are important for capturing the intricacies of social interaction.

2.4 Planning Socially Situated Behavior in Robotics

There has been recent work in robotics on planning with representations that either represent a person's intention as an unobservable part of the state or represent the role of time in human action, much of it focussed specifically on navigation. Cirillo and colleagues design a planning algorithm that treats the human's agenda as partially observable information and represents human action as having duration (but not as being time-dependent) [10]. Behavior that is acceptable to people is enforced through interaction constraints that specify whether or not the human's goal is to be achieved. Tipaldi and Arras create a time-dependent probabilistic model of where people are likely to be in an environment throughout the day and use it to plan to maximize the chance of encountering a person [36]. This model, while it captures the time-dependent nature of human behavior in the aggregate, doesn't represent the intentional behavior of individual humans. Foka and Trahanias's POMDP-based navigation algorithm makes use of beliefs about people's future paths [13]. In these POMDP models, the representation of human intention is restricted to which of the possible goal locations in the environment a person might be navigating towards. The work in this article differs from that previously mentioned both in using a framework that captures a more general notion of human intention and time-dependent action and in seeking to experimentally evaluate the impact of modeling interaction in this manner.


(a) Cars wait at red light. Turning car creeps into intersection.

(b) Oncoming car flashes lights to signal that they will yield for a Pittsburgh left.

(c) Turning car takes left immediately after light turns green.

(d) Other car goes straight through the intersection once the turn is completed.

Fig. 1 Example execution of the Pittsburgh left

3 Problem Domain

Driving in traffic is a domain in which self-interested agents are following a set of guidelines for interaction. The intentions of the other drivers are not directly observable. Exogenous events in the environment, such as the changing of traffic lights, provide external temporal cues that people use to coordinate the interaction. In some cases, driving interactions are governed by social rules in addition to traffic laws. One example is the "Pittsburgh left." The Pittsburgh left is a technically illegal, yet locally common driving maneuver that seems to have developed in response to Pittsburgh's many narrow, two-lane intersections that have traffic lights without turn signals [40]. A car making a left turn takes the turn immediately after the light becomes green (as if there were a left turn signal) if the oncoming car opposite it chooses to yield. This action allows the cars behind the turning car to pass through the intersection, rather than having to wait for the next light as they would if the turning car waited until all oncoming cars had passed. See Fig. 1 for an example of how a Pittsburgh left may be carried out.

Whether a Pittsburgh left will be taken is negotiated between the two cars using a variety of nonverbal cues. Either car may creep towards the intersection while the light is red in order to either suggest an interest in taking a Pittsburgh left or to indicate to the turning car that they will not allow one to be taken. The car going straight may flash their headlights to communicate their willingness to yield. These cues are not always present, however, so drivers must often determine whether a car is yielding by reasoning about the timing of their motion in relation to the changing of the traffic lights. Hesitation after the lights change may indicate the intention to yield, or may just be a delayed reaction. The ambiguity or potential absence of cues suggesting the other driver's intention and the time pressure from the cycle of the traffic lights make the task of coordinating behavior during this interaction difficult.

This domain has characteristics in common with the previously mentioned robot elevator riding task. However, during face-to-face interactions (such as entering or exiting elevators), people use gaze as a powerful social cue to indicate their intentions. Mobile robots are disadvantaged in that they are currently unable to reliably detect a person's gaze direction in natural environments. While people often also use gaze as a cue when driving, they are able to perform the Pittsburgh left interaction without access to the other driver's gaze using only the motions of the cars. This was confirmed by having human pairs perform the interaction in the same simulated driving environment used for the human-robot interaction experiment to be described in Sect. 5.1 [8].

4 Models

In order to evaluate the impact of representing time-dependence in action outcomes and reasoning over beliefs about human intention, different types of models with different representational capabilities were produced. The model variants created for comparison were time-dependent POMDP models based on a time-indexed state space, MDP models with a time-indexed state space, and POMDP models without time in their state space. In the Pittsburgh left domain models, the human's intention is the sole unobservable part of the state space. The rest of the physical characteristics of the underlying state are relayed to the agent directly through the observations or can be inferred from the action effects. Because of this characteristic, it is simple to produce a closely related MDP model by eliminating the human's intention from the state space. Similarly, the time-index is a single state variable which can be removed from the POMDP models in order to eliminate the representation of the time-dependent action effects. By removing these representational properties, we are able to study their effects on the quality of the policies these models produce.

All of the model variants evaluated were produced from the same model description, expressed using a simple domain independent language designed for this purpose. The model description uses domain knowledge encoded as probabilistic action rules to produce a complete model of the interaction. The algorithm by which the model description is transformed into a time-dependent POMDP model is described in a prior publication by Broz and colleagues [8]. The production of the non-time-dependent POMDP and the MDP models uses the same algorithm, but with certain state variables omitted from the state space of the resulting models. Two time-dependent POMDP models, one for each robot intention, and an equivalent pair of non-time-dependent POMDP models were created. Four MDP models were created that assume a fixed intention for the human driver, one for each combination of human and robot driver intentions.

The models produced contain all of the actions that the person interacting with the robot might choose, as well as all possible changes in the environment that may occur. The model description also provides a probability distribution over these action outcomes for all of the robot's possible actions. The relationship between the visible actions of the human and their unobservable intentions are encoded in the model as hidden state. In this modelling approach, an intention is assumed to be fixed for the duration of an episode of interaction, reflecting an agent's preferences about how they would like the interaction to play out even if they do not achieve their preferred goal during execution. The model's reward structure assigns values to states based on their desirability as goals, given the robot's intentions, and also penalizes states that correspond to a violation of social guidelines. The accuracy of the models were improved for the most frequently visited parts of their state space by updating the probability distributions using data collected from humans performing the task.

The models for the Pittsburgh left driving task model the interaction with the robot performing the role of the driver turning left and a human performing the role of the driver going straight. The continuous values that make up the state of the interaction were coarsely discretized to produce a finite state space of tractable size. The robot's action space is made up of 4 discrete actions. The observation space for the POMDP models consist of 724 possible observations and is made up of a subset of the state variables. In this virtual environment, the robot was given direct access to the state of the game engine rather than modeling sensor error. Because the observations arise directly from the state values, they are aliased rather than noisy. The models used in this experiment were created from a model description file containing 86 action rules. The model description file format, the variables that define the models, and a sample action rule are given in Appendix. Full details of the model description language and the model description file used for the Pittsburgh left interaction can be found in the technical report of Broz's PhD thesis [6].

The robot's own intentions are expressed as a preference ordering over possible goals. In order for these goals to reflect the social nature of the task, they involve outcomes and events for both the human and the robot. The robot prefers states in which both it and the human achieve desired outcomes. When their goals are in conflict, the robot may prioritize its own goals over those of the humans (depending on the domain), but it should still prefer better outcomes for the human over poor outcomes. The rewards assigned for

Table 1 Reward states and values in the Pittsburgh left task model description for each robot intention

Intention	Event					
	Turning Car First	Straight Car First	T. Only	S. Only	Run red	Collide
Pittsburgh Left	$15_T + 20_S = 35$	$15_T + 10_S = 25$	15_T	10_S	-5_T	-100_T
Regular Left	$15_T + 10_S = 25$	$25_T + 10_S = 35$	15_T	10_S	-5_T	-100_T

goals in the Pittsburgh left interaction are chosen so that the combinations of outcomes for the human and the robot have a value ordering that matches the desirability of these outcomes given the robot's intention. When the robot exits the intersection, it receives a one-time reward. When the human exits the intersection, the robot receives another one-time reward with a value determined by the robot's intention. If the robot's intention is to take a regular left turn rather than a Pittsburgh left, the robot receives a greater reward if the human exits the intersection first rather than second, and vice versa if the robot intends to take a Pittsburgh left. When only one of the cars is able to cross the intersection before the light turns red, the robot receives a reward for that car's success only. If a collision occurs, the robot receives a one-time reward with a large negative value. If the robot is positioned in the intersection while the traffic light is red, it receives a reward with a small negative value. Reward in these models is discounted by a factor of 0.99. This discount factor was chosen to give preference to earlier goal achievement while still maintaining the ordering of the goals' values regardless of when they were achieved. The rewards given for events involving the turning car (T) driven by the robot and the straight (S) car driven by the human are shown in Table 1.

4.1 POMDP Models

4.1.1 Time-Dependent POMDP Models

A detailed description of how the time-dependent POMDP models are produced from the action rules in the model description file and human task performance data is given in a prior publication [8]. The model description file produces two POMDP models, one for each possible robot intention. The reward structures of the models differ based on the intention. These POMDP models have a time-index in their state space. The time-indexed POMDP models, with 72,024 states each, were too large and complex to solve within a reasonable amount of time. Time-state aggregation, a reward-based state-aggregation method that operates on the time dimension of the state space, was developed to produce smaller, practically solvable time-dependent POMDP models. A property of this algorithm is that the threshold parameter which controls the amount of aggregation can be adjusted, trading off the number of resulting states with the similarity to the original time-indexed model. Details of the time-state aggregation algorithm and simulation results for policies it produced for an earlier version of the Pittsburgh left POMDP models are described in other previous work by Broz and colleagues [7].

For the time-dependent POMDP models, a time-state aggregation threshold of 4.0 was used, reducing the state sizes by a little less than half (see Table 2). While it is most likely possible that a larger threshold could have produced smaller models with similar performance, this amount of aggregation was deemed sufficient to produce a solution within an achievable time frame while being conservative with respect to the amount of aggregation. This decision was made in order to give the aggregated time-dependent models the fairest comparison possible to the time-indexed models used for the MDP model variant.

4.1.2 Non-Time-Dependent POMDP Models

The POMDP models produced without a time index as part of their state space are referred to as non-time-dependent POMDP models because they cannot accurately represent time-dependence in action outcomes. All MDP and POMDP models have some sort of time-based structure by virtue of the fact that they are sequential models. They are capable of representing the order in which states are encountered, but represent the effects of time on these state transitions with limited accuracy. The amount of time spent in a state can be represented only using a self-transition. The resulting time in state will have a geometric distribution, which may be a poor match for the actual distribution of the time spent in the state. If the state transition probabilities change over the time spent in the state, these changes cannot be represented by the model. The state transition probabilities are restricted to being an average over the different time-dependent probabilities for that state.

The non-time-dependent models are produced from the same model description file as the time-indexed models using the same procedure, with one major exception. The time index variable is not a part of the resulting state space. The action rules are still applied in a time-dependent manner during model construction (if a rule has time-dependent preconditions it will still be applied only at the appropriate timesteps), but the new states resulting from the rule's application will not have a state variable indicating the timestep at which they were created. The resulting models represent the

Table 2 Model and policy information for experiment POMDP models

Model			Policy		
Variant	Intention	States	Alpha Vectors	Convergence Time (s)	Reward
time POMDP	Pgh Left	37982	879	154592	19.6 [19.1, 20.1]
time POMDP	Reg Left	37430	2152	219484	19.0 [18.3, 19.5]
no time POMDP	Pgh Left	4398	281	1485	8.5 [7.9, 9.0]
no time POMDP	Reg Left	4398	635	3924	15.0 [14.1, 16.0]

time-dependent aspects of the domain description as accurately as they can without an explicit representation of time in the state space.

The resulting models have fewer states and more dense state transition matrices than their equivalent time-dependent models. To see this, consider a time-indexed model. There can be no more states with a particular time-index than there are total states in the non-time-dependent model. These states will only transition to states with the successive time index. In contrast, a state in the non-time-dependent model has transitions to all successor states, even if some of those states may only truly be successors at particular times. Therefore, while containing many more states, the time-indexed models have a more sparse state transition structure.

4.1.3 POMDP Policies

Though solving for even approximately optimal POMDP policies is computationally expensive, there are a number of POMDP solution algorithms that can achieve good performance on reasonably large models [24, 25, 30]. Heuristic search value iteration (HSVI) is an approximate POMDP solution algorithm that provides provable bounds on the reward obtained by the polices it produces with respect to the optimal policy [32]. The algorithm stores a compact representation of the upper and lower bounds of the value function over the belief state.

All of the POMDP models were solved using the HSVI algorithm with a convergence threshold of 5.0. The size of the models, the size of their resulting policies, and their convergence times are shown in Table 2. The original time-indexed POMDPs were not compared because they did not reach convergence within a week, demonstrating the benefit of the time-state aggregation algorithm for producing models of tractable size. While the time-dependent POMDPs took considerably longer to converge to a solution than the non-time-dependent models, the rewards obtained (and, as will be shown later, the manner in which the policies perform the task) demonstrate that non-time-dependent models, which are smaller and will typically be faster to find solutions for, will often produce policies with unacceptable performance in domains with time-dependent aspects. Reward results for simulation of 1000 trials of the POMDP policies acting on observations and state transitions produced by the time-indexed POMDP are also shown for each policy, with 95 % bootstrap confidence intervals. The non-time-dependent policies perform noticeably worse, indicating that an explicit representation of time is necessary to capture aspects of the problem that enable successful performance of the task.

It is worth noting that while the number of alpha vectors making up a policy is much larger for the time-dependent POMDPS, the number of states that the controller will have a non-zero belief of occupying at a given timestep during execution may not actually be significantly more than for the non-time-dependent model. The time-indexed POMDP would have a belief over the same number as or potentially fewer states (if some states can't occur at a timestep) than the non-time-dependent model. The complexity of the belief state of the time-dependent model will depend both on the structure of the original time-indexed model and the amount of time-state aggregation applied to it. For these particular models, the time-dependent policies typically had non-zero beliefs of three to four times as many states as the non-time-dependent policies during execution, though they transiently had up to over twenty times more. Both policies could be executed in real time at the action execution rate for the problem domain.

4.2 MDP Models

There are a number of different ways to make the state fully observable to produce MDPs that corresponds to the POMDP models. The method used was chosen to allow for the most fair and reasonable comparison. One option would be to simply eliminate the human's intention both from the state space and from the rules that produce the state structure. This would create an MDP that combined the actions of the human participants' different intentions into one aggregate model. More than not being able to reason about people's intentions, this model would not represent people's behavior as intentional at all. The model would suggest that because it is common to see humans flash their headlights (when they intend to yield) and accelerate quickly into the intersection (when they don't intend to yield), it would also be likely to see a person to flash their headlights and then immediately accelerate into the intersection. A model of this form is a poor representation of people's behavior.

An alternative MDP representation is to still represent people's behavior in terms of their intentions, but to create a separate MDP model that describes people's behavior for each possible intention. A policy produced by solving one of these models effectively assumes that the person it is interacting with holds this intention and acts accordingly. It is not an uncommon occurrence in human-human social interaction for someone to incorrectly infer the intention of the person they are interacting with, and people are accustomed to recognizing and compensating for these situations. By using a policy for interaction based on an assumption about the human's intention (that may or may not be correct), the onus of figuring out the other's intention and responding appropriately is shifted entirely onto the human participant.

The MDP models are created by first creating the states and state transition structure for the time-indexed POMDP. The resulting state space is then partitioned according to the values of the variable for the human's intention. The state transition structures for the separate state spaces are taken directly from the state transition structure of the POMDP model. Because the states of the MDP models are fully observable, the observation matrices are discarded.

4.2.1 MDP Policies

The MDP were solved using focused real-time dynamic programming (FRTDP), a heuristic search-based approximation algorithm that is similar to HSVI in that it maintains upper and lower bounds on the value function. FRTDP differs in that it uses a different search strategy, choosing states based on cached priority information in order to avoid revisiting states that do not improve. FRTDP is typically faster to converge than HSVI, at the expense of having far greater memory requirements. A description of FRTDP applied to MDP models is given in [33].

The time indexed MDP models were solved using the FRTDP algorithm with a convergence threshold of 0.1. A different algorithm from the one used to solve for the POMDP policies (which cannot practically be used for the POMDP models) and tighter convergence threshold were used to realistically reflect the advantages of using a simpler, less computationally expensive representation. The number of states in each model, their policy convergence times, and their solution bounds are given in Table 3.

5 Experiment

5.1 Pittsburgh Left Driving Task

People performed the Pittsburgh left task in a driving simulator with a robot car as their partner, interacting with controllers created from each of the three model variants. The simulated driving environment developed was based on TORCS, an open-source driving game [38]. A custom game level was designed to simulate driving in a suburban environment, including a 4-way intersection with functioning traffic lights. The game engine was extended to allow for control of the environment, automation of the experiment, and data collection. Participants controlled their cars in the driving simulator using an off-the-shelf steering wheel and pedal game controller. In addition to steering and braking, buttons on the controller's steering wheel allowed subjects to flash the car's headlights. A screenshot of the driving simulator and a human operating the simulator controls are shown in Fig. 2. The driving simulator served as a common environment in which humans and a robot car could perform an identical task. Humans had previously performed the Pittsburgh left interaction with other humans in an earlier experiment, allowing human task performance data to be collected and providing a baseline to which the robot's performance can be compared [8].

Each person engaging in the experiment interacted with each model variant's controllers multiple times, always performing the role of the car driving straight through the intersection with the robot car always performing the role of the left turner. They experienced all possible combinations of human and robot intentions, including the cases in which

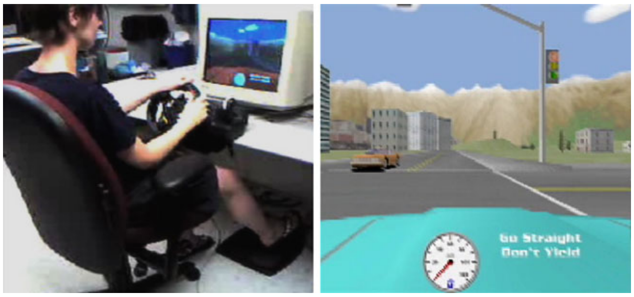


Fig. 2 The driving setup and POV in the simulator

Table 3 Model and policy information for experiment MDP models

Model			Policy	
Robot Intention	Human Intention	States	Convergence Time (s)	Reward Bound
Pgh Left	Yield	36152	862	21.5–21.6
Pgh Left	No Yield	35876	684	19.5–19.6
Reg Left	Yield	36152	1128	19.4–19.5
Reg Left	No Yield	35876	670	21.9–22.0

the human's and robot's intentions were in conflict (when the robot intended to take a Pittsburgh left and the human did not intend to yield and when the robot intended to take a regular left and the human also intended to yield).

An episode of interaction in the simulator lasted for one cycle of the traffic light. Throughout each episode, a message displayed on the lower right corner of the simulator's screen informed a participant whether their intention for that episode should be to yield or not to yield to the other car. Each episode of interaction lasted thirty seconds and began with each subject's car several carlengths from the intersection, with the traffic light on red. During each episode, the traffic light remained red for the first fifteen seconds (enough time for both cars to approach and wait at the intersection), then changed to green for eleven seconds, then to yellow for the final four seconds of the trial.

5.2 Experiment Design

The experiment had a $3 \times 2 \times 2$ factorial incomplete repeated-measures design, with the model variant, the human's intention, and the robot's intention as the independent variables. The dependent variable was the reward achieved by the robot during an interaction (as specified by the reward structure for the models). Additionally, a paper survey was given to the participants after driving with the controller for each model variant in order to measure their impressions of the interaction.

5.3 Experiment Procedure

Prior to beginning, participants were briefed on use of the steering wheel controller and the experiment's overall structure. They were told that they would be driving with three different drivers during the experiment, that the behavior of these drivers may differ from one another, and that they were to give their impressions of their interaction with each driver by filling out a short survey after completing each set of trials. They were told that they would always drive straight through the intersection, and the other driver would always be turning left. They were also informed that they would be assigned an intention to either yield or not yield to the turning driver and instructed to act as if they had independently arrived at that intention. They were informed that it was possible that their assigned intention might conflict with the other driver's, and that if that seemed to be the case they should behave as they would in a real driving situation. They were also instructed to respect the traffic light and to try to avoid collisions with the other car.

After being briefed, drivers engaged in a short training session of four trials to familiarize themselves with the controls. During the training, the other car stayed stationary at its start position. The experiment began once the driver asked the experiment administrator any questions they had and confirmed that they were ready.

The experiment was divided into three sets of eight episodes each. During each set, the driver interacted with a robot car controlled by policies from one of the three model variants tested: the time-dependent POMDP models, the time-indexed MDP models, and the non-time dependent POMDP models. The order in which the human driver encountered each policy group was randomly selected at the start of the experiment.

The robot controller's and the human driver's intentions were assigned to them at the beginning of each episode. Each possible combination of intentions was experienced twice, in a random order determined at the beginning of the set of episodes. For the POMDP controllers, the policy corresponding to the robot's intention for that trial was used for execution. Because the MDP models were separate for each combination of robot and human intention, the human intention reported to the policy executor was selected in a random order such that the intention was reported correctly once and incorrectly once for each combination of intentions. After a set of episodes was completed, the subject had two minutes to fill out the survey before beginning the next set.

Experiment subjects were recruited through advertisements placed on the campuses of local universities. Participants were pre-screened in order to ensure that they drove in Pittsburgh on a regular basis and were familiar with the Pittsburgh left. Each participant was paid five dollars upon completion of the experiment. Thirty-five participants performed the experiment over the course of two weeks. Of these participants, five were eliminated from consideration, four for software problems that occurred during the course of the experiment that may have invalidated their results and one for failure to follow directions. Data from the remaining thirty subjects were used for analysis.

5.4 Controller Implementation

During execution, the robot car received action commands from a policy execution program. Within the simulator, variables of interest were discretized using the boundaries specified during model design. These discrete values were used to construct the observations and states passed to the policy execution software. At the start of each episode, the software was passed a message announcing its beginning, along with the policy that should be used for execution. A new message was passed every half second, consisting of an observation (or the state including the assumed intention for the human in the MDP case) and the true state of the simulator for logging purposes. The policy execution software passed the action chosen by the policy back to the driving simulator. In the driving simulator, a simple PID controller was used to translate the model actions into low level controls to the steering, throttle, and brake.

Table 4 Test results for the ANOVA

Source	Sum of Squares	df	F	p value
Model	33837.42	11	18.2124	<0.0001*
Error	119548.07	708		
C. Total	153375.48	719		
Group	2341.040	2	6.9322	0.0010*
Robot Intention	3608.049	1	21.3680	<0.0001*
Human Intention	76.048	1	0.4504	0.5024
Group × Robot Intention	4819.241	2	14.2705	<0.0001*
Group × Human Intention	3216.261	2	9.5238	<0.0001*
Human Intention × Robot Intention	12826.437	1	75.9621	<0.0001*
Group × Human Intention × Robot Intention	6940.339	2	20.5514	<0.0001*

The MDP policies were represented as state-action lookup tables. The POMDP policies, were represented as alpha vectors, which meant that belief tracking had to be performed during execution in order to select an action. Policy execution was performed using the ZMDP software [31]. An intermittent problem occurred when executing the POMDP policies that caused the policy executor to not return an action selection for several seconds at a time, losing synchronization with the driving simulator. The source of this error could not be determined, but it occurred rarely and only within the last few timesteps of the trial. It was observed to occur with both POMDP model variants, so it was determined not to be caused by any model characteristic that was restricted to either variant. The decision was made to cut off the control after the 55th timestep (out of 60), or the last 2.5 seconds of interaction. After this point, the low level controller drove according to the last high-level action it had received from the policy executor. While it was not necessary to do this for the MDP policy, the same cutoff was applied so that the policies would be compared fairly.

6 Analysis

Statistical analysis was performed on the reward results and the responses to the paper survey. The comparisons made were planned contrasts of non-orthogonal data because the time-dependent POMDP results were compared to both the MDP results and the non-time-dependent POMDP results. The sequential Dunn-Sidak adjustment for k comparisons was used to adjust the p value required to reject the null hypothesis for the statistical tests making pairwise comparisons between the time-dependent POMDPs and the other models [17]. The smallest of the p values must be smaller than the smallest adjusted p value threshold, or no other results with larger p values can be accepted as significant. This criteria holds for all tests up to k. The adjusted p values for $\alpha = 0.05$ and $k = 2$ are $p_1 = 0.025$ and $p_2 = 0.05$.

The polices for the time-dependent POMDP model performed better overall than the policies for both the time-indexed MDP and non-time-dependent POMDP models. The rewards obtained were the most consistently high across the combination of human and robot intentions. The survey results confirm that the behavior of the time-dependent POMDP policies was preferred by the human interaction partners.

6.1 Reward Results

An analysis of variance (ANOVA) was conducted to compare the mean reward obtained by the groups of policies for each model variant for all combinations of intentions for both human and robot. Because the hypothesis of this experiment concerns the performance of the time-dependent POMDP model versus the other models, the post-hoc tests used are planned contrast t-tests of the time-dependent POMDP with each of the other model variants. In cases where the elements of the interaction cannot be expressed in terms of these planned contrasts, Tukey's HSD (Honestly Significant Differences), a post-hoc test that partitions the means into groups based on a more conservative test of statistical significance, is used instead.

The F-test results for all main effects and interactions of the ANOVA are given in Table 4. It should not be very surprising that virtually all of the effects were significant, given that different combinations of intentions result in very different interactions between humans performing the interaction. These results also support the hypothesis that differences in representation lead to different behavior by the policies. The ANOVA results for the reward will be examined and discussed in this section to identify the sources of variation among the groups of policies of the different model variants. The reasons for these differences in reward will be more closely investigated by examining the outcomes and events observed during the experiment in Sect. 6.2.

Table 5 Mean rewards for Groups of policies by model variant with planned contrast posthoc t-tests

Group	Mean	t Ratio	p value
time POMDP	15.873		
full time MDP	11.937	$t = 3.319$	$p = 0.001^*$
no time POMDP	12.171	$t = 3.121$	$p = 0.002^*$

Table 6 Mean rewards for the intention of the robot and human with significance results

Robot Intention			
Pgh left	Reg left	t Ratio	p value
11.088	15.565	4.623	<0.0001*
Human Intention			
No yield	Yield	t Ratio	p value
13.002	13.650	0.671	0.5024

6.1.1 *Main Effects*

The primary purpose of measuring the reward obtained by the policies in this experiment was to test the hypothesis that modeling human intention as hidden state and representing time-dependent action outcomes would result in better performing policies than those of models that lacked these representations. This hypothesis was supported by the results for the "Group" main effect, as shown in Table 5. On average, the polices for the time-dependent POMDP model outperformed the policies for both the time-indexed MDP and non-time-dependent POMDP models. The posthoc t-tests confirm that these differences were statistically significant. This is the key result for the reward-based portion of the experiment analysis. Further discussion of the results will examine the similarities and differences among the policies for the model variants and the conditions under which they succeeded or failed to achieve good performance.

The mean rewards for the other two main effects, "Robot Intention" and "Human Intention", shown in Table 6, give insight into the aspects of the task domain that proved most difficult for all model variants. The mean reward for trials during which the robot intended to take the Pittsburgh left was lower than for trials in which the robot intended to take a regular left turn. Attempting the Pittsburgh left often results in more complicated interactions (with more risk of collision), so this result is not surprising. The human's intention, however, did not have a significant effect on the mean reward when considered in isolation. An examination of higher level interaction effects will yield more insight into the role humans' intentions played in the rewards obtained.

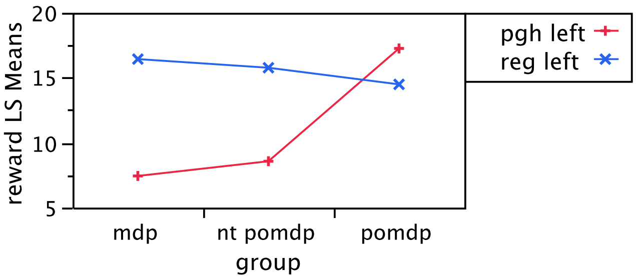


Fig. 3 Interaction of model variant policy groups with robot intention for mean reward

Table 7 Tukey's HSD posthoc test results for the Group × Robot Intention interaction

Group	Robot Intention	Mean		
POMDP	Pgh left	17.257	A	
MDP	Reg left	16.436	A	
NT POMDP	Reg left	15.770	A	
POMDP	Reg left	14.490	A	
NT POMDP	Pgh left	8.571		B
MDP	Pgh left	7.437		B

6.1.2 *Two-Way Interactions*

The interaction between the robot's intention and the model variant that produced its control polices is shown in Fig. 3. This interaction shows one of the major differences in performance between the model variants. The groups of means judged to be significantly different from one another given a Tukey's HSD posthoc test are presented in Table 7 (each letter in the table represents a group of values that the test determined to be significantly different from the values outside of that group). Both the MDP and the non-time dependent POMDP model produced policies that performed significantly worse than the time-dependent POMDP model when the robot had the intention to take the Pittsburgh left. These results suggest that only the time-dependent POMDP model was capable of representing the interaction in a way that produced high quality policies for both possible robot intentions.

The interaction between the human's intention and the model variant is shown in Fig. 4. This interaction reveals that there were differences between the POMDP and the MDP models in how successfully they responded to the humans' intentions. Though the time-dependent POMDP performed relatively well whether the human intended to yield to them or not, its performance was significantly better when the human did not intend to yield. The MDP model performed significantly worse in the case where the human intended to yield than the case where they did not, the reverse of the relationship between intentions seen for the time-dependent POMDP case. The non-time-dependent POMDP performed

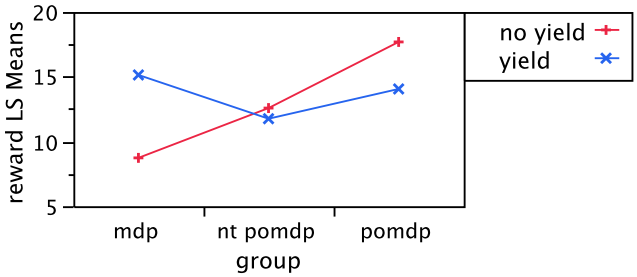


Fig. 4 Interaction of model variant policy groups with human intention for mean reward

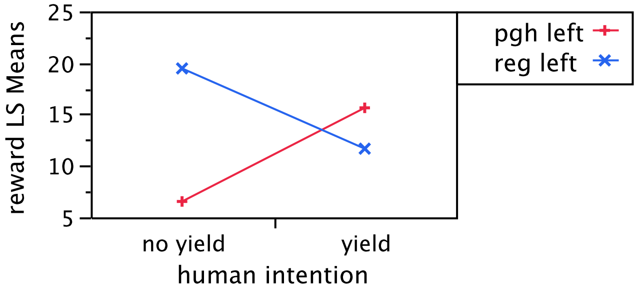


Fig. 5 Interaction of human intention with robot intention for mean reward

Table 8 Tukey's HSD posthoc test results for the Group × Human Intention interaction

Group	Human Intention	Mean			
POMDP	No yield	17.690	A		
MDP	Yield	15.139	A	B	
POMDP	Yield	14.058	A	B	
NT POMDP	No Yield	12.581		B	C
NT POMDP	Yield	11.760		B	C
MDP	No yield	8.735		B	C

similarly regardless of the human's intention. The groups of means judged to be significantly different from one another given a Tukey's HSD posthoc test are presented in Table 8.

The interaction between the human's intention and the robot's intention is shown in Fig. 5. Tukey's HSD found all of the means to be statistically significantly different from one another. This interaction shows that conflicting intentions resulted in lower rewards. It is not surprising that it was more difficult for the robot to achieve its most desired outcome in those cases. The performance of each group of policies for each of these combinations will be explored in the next section.

6.1.3 Three-Way Interaction

The rewards for each combination of intentions grouped by model variant, are shown in Fig. 6. At this level of analysis, the source of the differences in performance between the groups of policies begins to become clear. For certain combinations of intentions, the performance of the model variants were very similar, while for others they were radically different.

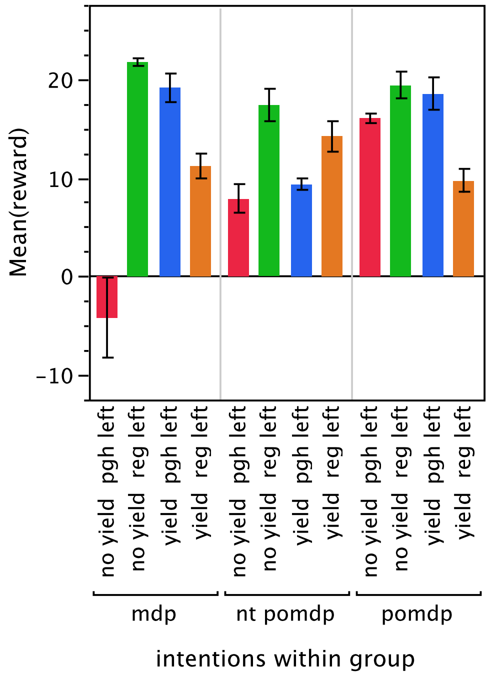


Fig. 6 Rewards with standard error for robot and human intentions, grouped by model variant

The mean reward for each group, with t-tests for significance of the differences, are reported for each combination of intentions in Table 9. These differences, when they occur, are due to what representation is necessary in order to perform that particular interaction successfully. For the case where the human intended to go first through the intersection and robot intended to take a regular left (Regular Left, No Yield), all of the policies performed similarly well. This is most likely because this is a relatively uncomplicated interaction where all that is typically required in order to coordinate behavior with the human is to wait until they have cleared the intersection.

In the case where both the human and the robot intended to yield to one another (Regular Left, Yield), all of the model variants also performed similarly. They also all performed worse than in the previous case. This suggests that none of the model variants were as successful as they could have been at negotiating this interaction, which indicates that this failure may have been caused by an aspect of the model or effect of policy execution that was common to all representations. This possibility will be explored in more detail in Sect. 6.2.

For the case in which the robot intended to take a Pittsburgh left and the human intended to yield (Pittsburgh Left, Yield), the policy for the time-dependent POMDP out-

Table 9 Reward and posthoc t-test results for each combination of robot and human intentions

Group	Mean	t Ratio	p value
Regular left, No Yield			
POMDP	19.344		
MDP	21.708	−0.997	0.319
NT POMDP	17.331	0.848	0.397
Regular left, Yield			
POMDP	9.637		
MDP	11.164	−0.644	0.520
NT POMDP	14.208	−1.927	0.0544
Pittsburgh left, Yield			
POMDP	18.478		
MDP	19.113	−0.268	0.789
NT POMDP	9.311	3.864	0.0001*
Pittsburgh left, No Yield			
POMDP	16.035		
MDP	−4.239	8.545	<0.0001*
NT POMDP	7.832	3.458	<0.0001*

performed the policy for the non-time-dependent POMDP by a significant amount. The performance of the MDP policies were similar to that of the time-dependent POMDP. This result indicates that *time representation* was critical to the success of this particular interaction. The MDP policies seemed to perform well in this case whether or not their assumption about the intention of the human participant was correct.

The case where the robot intended to take the Pittsburgh left and the human did not intend to yield (Pittsburgh left, No Yield) is the most interesting one in terms of the differences between the groups. The time-dependent POMDP performed considerably better than either of the other models. This result suggests that both *time representation* and *reasoning about people's intentions* played a role in the policy's success. While both of the other model variants performed worse, the performance of the MDP models' policies were far worse than any models for any other combination of interactions. Though this difference in representation did not have an impact in the other cases, in this case it seems to have produced a catastrophic failure in interaction. This will be considered further in the next section.

6.1.4 MDP Model Results

Rather than planning for both of the human's possible intentions in one policy as in the POMDP models, the MDP policies were created from separate models for each possible human intention. Which policy to use with respect to the human's intention was randomly selected at the beginning of the trial, causing the robot to effectively "assume" that the human's behavior would correspond to the selected intention and act accordingly. In order to understand the MDP model variant's failure in the (Pittsburgh left, No Yield) case, the trials in which the robot's assumption matched the human's intention and the trials in which they were a mismatch must be considered separately. The reward for the other cases were also compared in order to determine whether the correctness of the robot's assumption had an impact on performance. These comparisons are presented in Table 10.

Table 10 Reward for the MDP model policies, divided into trials in which the model's intention did and did not match the human's intention, with t-tests of the mean differences

Match	Mismatch	t Ratio	p value
Pgh left, No Yield			
−26.992	15.670	−11.89	<0.0001*
Pgh left, Yield			
18.707	19.404	−0.192	0.848
Reg left, No Yield			
22.486	21.226	0.342	0.733
Reg left, Yield			
11.205	11.133	0.020	0.984

The only statistically significant difference between the reward obtained for trials in which the policy's model held the correct assumption about the human's intention and the trials in which there was a mismatch between the assumed intention and the true intention occurred in the (Pittsburgh left, No Yield) case. Surprisingly, the policies with the correct assumption performed worse. This difference in performance is caused by an unexpected consequence of using human task performance data in the model. For interactions where one person intended to take the Pittsburgh left and the other person did not intend to yield, there were numerous examples in the data set used to adjust the model parameters where both of the cars drove into the intersection and narrowly missed colliding with each other because one of the cars stopped in time. The robot's controller issues new action commands only twice a second, and the overall ability of the robot to steer and modulate its speed are at a coarse resolution compared to the human's continuous control. The addition of human data to the model gave the robot an overly optimistic prediction of its ability to take last-minute maneuvers to avoid collision. For the model where the robot intended to take the Pittsburgh left and the human driver intended to yield, the vast majority of the examples of this combination of intentions in the human task performance data set had the driver going straight wait before the intersection while the turning car made the left. When this model (of the incorrect intention) was used to interact with a human driver that did not intend to yield, the human's actions of moving into the intersection took the controller into a part of the state space where little human data was observed, so

Table 11 Summary of trial outcomes, grouped by the combination of human and robot intentions

Model	Outcome					Event	
	T First	S First	T Only	S Only	Neither	Lights (S)	Collisions
Pittsburgh left, Yield (non-conflicting intentions)							
human	92 %	4 %	2 %	1 %	1 %	39 %	1 %
time POMDP	88 %	5 %	0 %	5 %	2 %	35 %	3 %
full time MDP	85 %	10 %	2 %	3 %	0 %	35 %	2 %
no time POMDP	80 %	0 %	0 %	20 %	0 %	2 %	0 %
Regular left, No Yield (non-conflicting intentions)							
human	2 %	97 %	0 %	1 %	0 %	6 %	1 %
time POMDP	2 %	88 %	3 %	6 %	0 %	2 %	2 %
full time MDP	0 %	97 %	2 %	2 %	0 %	0 %	0 %
no time POMDP	10 %	77 %	0 %	12 %	2 %	2 %	0 %
Pittsburgh left, No Yield (conflicting intentions)							
human	59 %	40 %	0 %	1 %	0 %	11 %	10 %
time POMDP	13 %	78 %	0 %	8 %	0 %	0 %	0 %
full time MDP	25 %	72 %	2 %	2 %	0 %	5 %	32 %
no time POMDP	88 %	0 %	0 %	12 %	0 %	0 %	2 %
Regular left, Yield (conflicting intentions)							
human	51 %	29 %	11 %	2 %	7 %	74 %	0 %
time POMDP	16 %	27 %	12 %	16 %	28 %	51 %	0 %
full time MDP	0 %	45 %	5 %	28 %	22 %	43 %	0 %
no time POMDP	48 %	35 %	6 %	6 %	3 %	50 %	0 %

it had little impact on the model parameters for those states. This meant that the robot's actions were determined by the predicted chance of collision according to the prior model, which was pessimistic about the chance of collision relative to the human data. This resulted in more cautious behavior (the robot stopped when the human entered the intersection) that successfully avoided collisions. Model inaccuracy is always going to be a potential issue when doing model-based planning in any realistic domain. The policies for the POMDP model, because they planned over a belief space in which either intention for the human driver was possible, were more robust to these model inaccuracies that arose from representing a fixed assumption about the human's intention in the model.

6.1.5 *Summary of Reward Results*

Overall, the policies for the time-dependent POMDP model variants outperformed the other model representations. The non-time-dependent POMDP policies performed significantly worse in both of the cases involving the Pittsburgh left intention, suggesting that it was not possible to produce quality policies for this intention without representing time-dependent action outcomes. The MDP policies performed significantly worse than the time-dependent POMDP policies only in the case where the robot intended to take the Pittsburgh left and the human did not intend to yield. The POMDP policies' ability to reason over beliefs about intentions made them more robust to minor model inaccuracies.

The time-dependent POMDP policies achieved results as good or better than the other model variants for all combinations of interactions. Additionally, it should not be assumed that policies that achieve similar rewards are subjectively similar to the humans interacting with them. There were observable differences in the driving styles of the model variants. These differences will be discussed in the next section, and their consequences will be seen in the discussion of the survey results.

6.2 Outcome and Event Results

In order to obtain a clearer picture of the differences in performance between the polices for the model variants, the frequency of possible trial outcomes and significant events are shown in Table 11. The outcome frequencies for data collection during which humans performed the same interaction in the driving simulator with other humans (involving sixteen pairs of people) are also given for comparison. In order to better understand the ways the policies of the model variants differed from one another, as well as from human behavior, the results for each combination of driver intentions will be considered separately.

6.2.1 Pittsburgh Left, Yield

For the trials in which the robot intended to take the Pittsburgh left and the human driver intended to yield, the combination of driver intentions is non-conflicting. In this case, all the models achieved outcome results relatively similar to those of human drivers. For the non-time-dependent POMDP case, notice that the human driver used their lights far less frequently than with any other model variant (or with humans). This is because the robot typically ran the red light, making signalling the driver to take the Pittsburgh left unnecessary. The non-time-dependent POMDP model was unable to accurately represent the true risk of collision between the robot and the human driver because of the lack of time representation in its state space (people are likely to start moving immediately once the light changes if they don't intend to yield and wait otherwise). In order to avoid this exaggerated risk of collision, it frequently chose to run the red light. While people occasionally jump the red by a very small amount of time (typically only after the perpendicular traffic light has also turned red), the non-time-dependent model is unable to represent this distinction. As a result, it frequently ran the red light long before it changed, resulting in very unnatural-appearing behavior. Also note the relatively high incidence of outcomes where only the driver going straight crossed the intersection.

While both the full time MDP and the time-dependent POMDP performed this interaction successfully, the way in which they achieved their outcomes differed in their responsiveness to the human participant. The MDP driver was more aggressive, usually immediately taking the left when the light turned. The POMDP driver was more cautious, and would often pause when the light changed. If the human flashed their lights, the POMDP controller would begin the turn, having had its belief converge to the case that the person intended to yield after that observation. If the person did not flash their lights, the POMDP would wait slightly longer to converge on the belief that they intended to yield. While both of these behaviors are within the range of what is normally exhibited and socially correct for the Pittsburgh left, the POMDP controller responds to behavior by the human driver in the way that is commonly understood for this interaction.

6.2.2 Regular Left, No Yield

For the trials in which the robot intended to turn left after the human driver and the human driver did not intend to yield, all the models performed relatively similarly to the human drivers. All of the models were able to coordinate their behavior relatively successfully because the human's and the robot's intentions were not in conflict. Additionally, because the event of the human driver crossing through the intersection first could be observed and then responded to (rather than needing to predict likely future actions by the human), the less expressive representation of the non-time-dependent POMDP also performed this interaction in an appropriate manner in the majority of cases.

6.2.3 Pittsburgh Left, No Yield

For trials in which the robot intended to take the Pittsburgh left and the human driver did not intend to yield, the intentions of the human driver and the robot were in conflict. The time-dependent MDP and POMDP policies yielded to the human driver far more frequently than a human turner yielded to an oncoming car. But high number of collisions for the MDP policies suggest that this is more appropriate behavior given the controllers' capabilities. Once again, the non-time-dependent POMDP policy succeeded in going first in the majority of cases by running the red light.

6.2.4 Regular Left, Yield

The intentions of the human driver and the robot were also in conflict for the trials in which the robot intended to turn left after the human driver and the human driver intended to yield. All of the policy groups had a large number of interactions where both of the cars failed to make it through the intersection before the red light. While humans were better at performing this interaction, it is worth noting that this case was the one that was most difficult for them as well. As in the other case where the robot and the human's goals were in conflict, both the time-dependent POMDP and MDP models' policies took the Pittsburgh left less frequently than human drivers. In fact, the MDP policies seemed to be quite inflexible, taking the left before the other car only in a very small number of cases. This seems to support the idea that the prior model may have overestimated the probability that the straight driver would eventually go first, allowing the turning car to achieve its most preferred outcome. The time-dependent POMDP policy chose to make the turn before the straight car more frequently than the MDP policy. This is another example of how the POMDP policies were more responsive to the behavior of the human driver. The non-time-dependent POMDP was most willing to take the left before the other car, most likely because of its inability to accurately estimate how much time remained to act. Because the model overestimated the possibility that the episode could end soon from many states, it to chose to achieve a less desirable outcome that it had direct control over rather than wait for the other car to cross the intersection first.

6.2.5 Summary of Outcome Results

The distributions of outcomes reached during interaction by each of the model variants show that the differences between their rewards were caused by noticeable differences in the way they engaged in the interaction. All of the policy groups were conservative drivers overall when compared to human drivers performing the same task, with a few notable exceptions. In the non-time dependent POMDP case, the desire to avoid collisions, combined with a model that lacked the time representation to accurately represent the interaction, caused the policy to choose to take penalized actions to run the red light in cases when the robot's intention was to make a Pittsburgh left. The behavior of the MDP policies was less responsive to the other driver than the other model variants. The time-dependent POMDP model produced behavior that was the most socially correct (not running the red light or hitting other drivers) and the most responsive to the behavior of the human driver.

6.3 Survey Results

The paper survey administered consisted of four statements about the experiment participant's driving behavior and that of their driving partner. The responses to the survey were a level of agreement with the statements presented, expressed on a 5-point Likert scale ranging from "strongly disagree" to "strongly agree". The responses were treated as ordinal data and the median was chosen as the statistic for measurement. Ninety-five percent bootstrap confidence intervals were also reported. Planned comparisons were made between the time-dependent POMDP trials and the time-indexed MDP trials and between the time-dependent POMDP trials and the non-time-dependent POMDP trials, as for the reward results. Because of the potential variability in people's subjective responses, a matched pairs test was used to evaluate the relative preferences for one model over another for the group of experimental subjects. The Wilcoxon signed rank test , a non-parametric test analogous to the paired t-test, was used to determine statistical significance in the cases where the median responses differed.

6.3.1 Human Driver Control

The first survey statement given was, "I felt that I was able to control the car in the simulation to do what I wanted it to do." This question was given as a control to ensure that difficulties with low level control of the car were unlikely to have a significant impact on people's performance of the task. The medians for each model variant are presented in Table 12. There was no statistically significant difference found between the medians. Any effect that was not due to chance may be explained by the human driver often having easier trials in terms not having to take any driving maneuvers to coordinate behavior because the other car had run the red light in the interactions with the non-time-dependent POMDP controller.

Table 12 Median values for the response to survey statement 1 with 95 % bootstrap confidence intervals and results of matched pairs Wilcoxon signed rank test

Model	Response	Test Result
time POMDP	4 [4, 5]	
full time MDP	4 [4, 5]	
no time POMDP	5 [4, 5]	$T = -12.5$, $p = 0.2734$

Table 13 Median values for the response to survey statement 2 with 95 % bootstrap confidence intervals and results of matched pairs Wilcoxon signed rank test

Model	Response	Test Result
time POMDP	4 [3, 4]	
full time MDP	2 [2, 4]	$T = 50.5$, $p = 0.041^*$
no time POMDP	2 [1, 2]	$T = 130.5$, $p < 0.001^*$

6.3.2 Perceived Naturalness of Controller

The second survey statement was, "The actions of the other car seemed natural to me." For this statement, there were significant differences found between the responses for the time-dependent POMDPs and both of the other two model variants, as summarized in Table 13. The time-indexed MDP and the non-time dependent POMDP both received a median response of 2 ("somewhat disagree") while the median response for the time-dependent POMDP was 4 ("somewhat agree"). For the time-indexed MDP responses, there was a bimodal split in the data. While some people agreed the behavior was natural, a greater proportion disagreed. This result may be due to some people finding the "aggressive" driving style of MDP policies realistic. The poor responses for the non-time-indexed POMDP were probably due to the fact that a person running a red light when making a left turn is an extremely rare occurrence.

6.3.3 Similarity of Interaction to "Real World"

The third survey statement was, "The interactions we engaged in were similar to the way I interact with other drivers taking the Pittsburgh left in real life." As shown in Table 14, the median response for all of the controllers was 4 ("somewhat agree"). Because the medians did not differ, no statistical tests were performed. However, the confidence intervals for the medians suggest that there were potentially meaningful differences between the distributions of responses. The non-time-dependent POMDP controllers had a distribution with more weight on the left ("disagree") side of the scale than the other controllers. It seems that there was less consensus among people as to how closely those interactions

Table 14 Median values for the response to survey statement 3 with 95 % bootstrap confidence intervals and results of matched pairs Wilcoxon signed rank test

Model	Response
time POMDP	4 [4, 4]
full time MDP	4 [3.5, 4]
no time POMDP	4 [2, 4]

Table 15 Median values for the response to survey statement 4 with 95 % bootstrap confidence intervals and results of matched pairs Wilcoxon signed rank test

Model	Response	Test Result
time POMDP	4 [3, 4]	
full time MDP	2.5 [2, 4]	$T = 60.0$, $p = 0.019$*
no time POMDP	2 [1.5, 3]	$T = 102.0$, $p = 0.001$*

mimicked real world interactions. This might be due to differences in subjects' opinions on how negatively they perceived the red light running behavior and how strongly it influenced their opinion versus other cases where the behavior of those controllers were more similar to the other models. Despite having the same median, people's opinion of these controllers was less consistently positive than their opinion of the other model variants, which most people thought resulted in realistic interactions. This might seem surprising, considering that the MDP sometimes caused a collision. But as noted before, this did not happen to every driver, and when it did it was typically one interaction out of eight. People might have viewed the collision as an anomaly or as an event for which they shared responsibility. These results indicate that people found the MDP controllers' relatively aggressive driving style to be similar to driving behavior that they'd experienced in the real world, which is not surprising, given that drivers vary a great deal in their level of aggressiveness.

6.3.4 Socially Appropriate Interaction

The final question of the survey, and the one most significant to assessing the relative performance of the controllers for social interaction, was, "I felt that the other driver followed proper driving etiquette for taking and yielding to the Pittsburgh left." The median response for the time-dependent POMDP controllers was 4 ("somewhat agree"). Both other model variants were assessed less favorably, with the non-time-dependent POMDP receiving a median response of 2 ("somewhat disagree") and the time-indexed MDP a median response of 2.5 (just between "somewhat disagree" and "no opinion"). Table 15 shows that these differences were found to be statistically significant.

The distribution for the non-time-dependent POMDP is strongly skewed to the left ("disagree"). This strong negative response is most likely due to the fact that this controller ran the red light during multiple trials of its interaction with every experiment subject. No matter how it may have behaved in other cases, people's impressions were undoubtedly effected by its violation of this traffic law.

The reactions to the behavior of the time-indexed MDP controller were more mixed. The responses had a bimodal distribution, with the majority of the weight split between "somewhat disagree" and "somewhat agree". One might wonder if these modes correspond to the drivers that were and were not hit by a car controlled by the MDP. But a closer examination of the data reveals a more complex picture. A collision occurred for 17 out of the 30 drivers during a trial with the MDP controller. But 6 of these 17 gave the controller a neutral to positive response of 3 or 4. Of the 13 drivers that did not experience a collision, 4 gave the controller a negative score of 1 or 2. While the collisions probably made a strong impact on people's opinion of whether the MDP controllers' behavior was socially appropriate, it was not the only factor. The MDP controller generally appeared less responsive to the actions of the other driver than the POMDP controllers. When taking the Pittsburgh left, it would start moving immediately in most cases, as opposed to the POMDP controllers, which would often hesitate until the human driver flashed their lights or did not move for several timesteps.

7 Conclusion

The benefit of explicitly representing time and planning over beliefs about intention in social interaction tasks was demonstrated through interaction with people in a realistic problem domain. Analysis of the rewards achieved by the model variants supports the experimental hypothesis that the time-dependent POMDP model performs better than either the time-indexed MDP or the non-time-dependent POMDP model. Model inaccuracies caused the MDP policy to have collisions with human drivers noticeably more frequently than the other model variants, and its behavior was also less responsive to the actions of the human participants overall. The non-time-dependent POMDP was unable to accurately represent the interactions in which the robot intended to take the Pittsburgh left, and its policies chose to run the red light rather than risk collisions with the human driver. The time-dependent POMDP policies exhibited the most consistently successful behavior across all combinations of interactions and were responsive to the actions of human drivers, most likely contributing to their preference for that model variant. According to the survey responses, the controllers from time-dependent POMDP model variant were judged as producing more natural actions than either the time-indexed

MDP or the non-time-dependent POMDP model variants. More importantly, they were also found to be better at following the proper social protocol for the Pittsburgh left. In light of these results, the time-dependent POMDP clearly outperformed the other model variants in terms of people's subjective impressions of their behavior. This work highlights the importance of making the correct representational choices in designing effective interaction policies for autonomous robots performing social tasks.

Appendix: Model Description

The overall structure of the domain description file is given in Table 16. The state space is time-indexed, so the number of timesteps in an episode must be specified. The time index is represented by a special state variable named "Time" that always has the value of the current timestep. This variable may be referred to in the preconditions of action or reward rules but may not have its value changed by an action rule. The rest of the state space is defined by a list of state variable names with the corresponding range of integer values that each variable make take.

The action space is defined in the same manner. Additionally, three special action variables are defined, "Human," "Env," and "Side Effects". These are used to specify the action rules that describe the actions of the human agent and the environment. Because these actions are automatically taken in response to the actions of the robot at each timestep rather than being chosen by the robot, their range of action values are undefined.

A.1 States and Observations

The state space for the POMDP models is defined as follows:

- human intention (2)—go first, yield
- robot position (5)—the region occupied by the robot
- human position (5)—the region occupied by the human
- robot angle (2)—whether car body is turned more than 45 degrees
- robot speed (3)—stopped, rolling/creeping, fast
- human speed (3)—stopped, rolling/creeping, fast
- human headlights (2)—whether the human has flashed their headlights
- collision (3)—no collision, collision at this timestep, collision has occurred
- traffic light (3)—the color of the traffic light
- time (60)—the current timestep

Table 16 Structure of the model description file

TIMESTEPS
<VAL>
number of timesteps
corresponds to special state variable named "Time"

STATES
<STATEVAR DEF LIST>
list of state variable names and their ranges of values

OBSERVATIONS
<STATEVAR LIST>
subset of state variables which are directly observable

ACTIONS
<ACTVAR DEF LIST>
list of action variable names and ranges of values
inc. special vars "Human", "Side Effect", and "Env"

<ACTION RULE LIST>
rules that describe the effects of actions on the state

<REWARD RULE LIST>
rules that describe the reward value of certain states

START
<STARTVAL LIST>
state variable values for the set of possible start states

The intersection and the surrounding roads are partitioned into rectangular regions relative to each agent's starting position that are significant to the interaction. These regions are: more than a car length from the start of the intersection, a car length from the start of the intersection, the half of the intersection closest to its start, the half of the intersection closest to its end, and the road on the other side of the intersection. There is an additional position value that corresponds to the event of an agent entering the region of the road beyond the intersection for the first timestep. This value is used to assign the one-time reward for an agent crossing the intersection.

Observations are created by defining a subset of the state variables that are directly observable. The observable state variables for this domain are as follows:

- robot position
- human position
- human headlights
- human speed
- collision
- traffic light

Because the robot's current speed and angle should always be known based on the effects of the set speed and turn actions, these variables are not included in the observation space.

A.2 Actions and Action Rules

Four actions are available to the robot to control its motion. These are high level actions that must be converted into control commands for the car in the driving simulator by a low level controller.

- set speed (3)—stop, go slow, go fast
- turn (1)—turn car body to the left

It is assumed that these actions can produce their immediate effect as a change in state within the timestep they are applied. For example, if the car is going fast and the robot chooses the stop action, the car will be stopped by the next timestep. Other action effects, such as whether moving at a certain speed will cause the robot to move into the next region, are probabilistic.

Further probabilistic action outcomes are created by sequentially applying the action rules for the special action variables to the states resulting from the robot's action rules. The action rules for the "Human" action describe the possible behavior of the human. The action rules for the "Env" action have preconditions based on the time variable and control the changing of the traffic light. The action rules for the "Side Effects" action detect combinations of state variables indicating that the robot and human have driven into the same region in a way that may result in a collision.

Action rules are defined by the preconditions that determine whether or not the rule is applicable in a state, the possible effects of that action on the state variable values, and a weighting factor that specifies the relative likelihood of those outcomes. This action description language provides a simple and flexible way to express the state transition structure in terms of small subsets of relevant variables and their values. For each action rule, the action variable and variable value that the rule describes is specified. In order to uniquely identify each action rule an additional rule name is also given. This is necessary because the same combination of action variable and value may have different effects and different weights depending on the preconditions. Multiple action rules may be used to specify the full range of possible outcomes and their relative weights for a single action. The possible changes to state variables caused by an action rule are specified as a list of action effects. Each action effect may effect multiple state variables simultaneously. All of the effects for an action rule are given equal weight by default. The weight, if specified, is an integer that defines a ratio of how much less likely the action effects listed are than the default. For example, a weight of 2 would mean that a rule's action effects are half as likely.

An example action rule describing the behavior of the human agent is shown in Table 17. The special "Human" action variable specifies that this action rule describes the behavior of the uncontrollable human agent. Because the robot cannot choose the actions taken by this other agent, the action value is 0, with "flash_no_yield" providing a unique identifier for this action rule. This action rule has a single possible effect, to turn on the headlights of the human's car. The condition list describes the state variable values that determine when this action rule may take effect. It only makes sense to apply this action in states where the headlights are off. The human may turn on the lights at any time before they move into the intersection. The goal condition restricts this rule to cases where the human has the intention not to yield to the turning driver. The velocity 0 condition reflects the fact that people were only observed to turn on their headlights when their car was not moving. This action effect is $\frac{1}{100}$ as likely to occur as a default outcome because a car that doesn't intend to yield flashing its headlights is a rare event.

Table 17 An example action rule describing a light flashing behavior by the human

```
RULE
Human 0 flash_no_yield
EFFECTS
Light_S ABS 1
CONDITIONS
Light_S 0
Pos_S 0 1
Goal_S 1
Vel_S 0
WEIGHTS
100
```

References

1. Abele S, Bless H, Ehrhart KM (2004) Social information processing in strategic decision-making: why timing matters. Organ Behav Hum Decis Process 93(1):28–46
2. Aumann RJ (1976) Agreeing to disagree. Ann Stat 4:1236–1239
3. Baldwin DA, Baird JA (2001) Discerning intentions in dynamic human action. Trends Cogn Sci 5(4):171–178
4. Bosch Lt, Oostdijk N, Ruiter JPd (2004) Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues. In: Proc of text, speech, and dialogue, 7th international conference (TSD), pp 563–570
5. Bratman ME (1990) What is intention? In: Cohen PR, Morgan J, Pollack ME (eds) Intentions in communication. MIT Press, Cambridge, pp 15–31
6. Broz F (2008) Planning for human–robot interaction: representing time and human intention. Tech. Rep. CMU-RI-TR-08-49, Carnegie Mellon U. Robotics Institute
7. Broz F, Nourbakhsh IR, Simmons RG (2008) Planning for human–robot interaction using time-state aggregated pomdps. In: Proc of the 23rd AAAI conference on artificial intelligence (AAAI), pp 1339–1344
8. Broz F, Nourbakhsh IR, Simmons RG (2011) Designing pomdp models of socially situated tasks. In: Proc of the 20th IEEE international symposium on robot and human interactive communication (Ro-man)
9. Chernova S, DePalma N, Breazeal C (2011) Crowdsourcing real world human-robot dialog and teamwork through online multiplayer games. AI Mag 32(4):100–111
10. Cirillo M, Karlsson L, Saffiotti A (2009) A human-aware robot task planner. In: Proc of the 11th international conference on automated planning and scheduling (ICAPS), Thessaloniki, Greece
11. Colman AM (2003) Cooperation, psychological game theory, and limitations of rationality in social interaction. Behav Brain Sci 26:139–153

12. Emery-Montemerlo R, Gordon G, Schneider J, Thrun S (2004) Approximate solutions for partially observable stochastic games with common payoffs. In: Proc of the 3rd international joint conference on autonomous agents and multi-agent systems (AAMAS)
13. Foka A, Trahanias P (2010) Probabilistic autonomous robot navigation in dynamic environments with human motion prediction. Int J Soc Robot 2:79–94
14. Frith U, Frith C (2001) The biological basis of social interaction. Curr Dir Psychol Sci 10(5):151–155
15. Geanakoplos J (1992) Common knowledge. In: Proc of the 4th conference on theoretical aspects of reasoning about knowledge (TARK), pp 254–315
16. Kaelbling LP, Littman ML, Cassandra AR (1998) Planning and acting in partially observable stochastic domains. Artif Intell 101(1–2):99–134
17. Kirk RE (1982) Experimental design, 2nd edn. Brooks/Cole, Pacific Grove
18. Lewis DK (1969) Convention: a philosophical study. Harvard University Press, Cambridge
19. Maynard Smith J (1998) The origin of altruism. Nature 393:639
20. Meltzoff AN (1995) Understanding the intentions of others: reenactment of intended actions by 18-month-old children. Dev Psychol 31(5):838–850
21. Mui L, Mohtashemi M, Halberstadt A (2002) Notions of reputation in multi-agents systems: a review. In: Proc of the 1st international joint conference on autonomous agents and multiagent systems (AAMAS), pp 280–287
22. Nourbakhsh I, Powers R, Birchfield S (1995) Dervish an office navigating robot. AI Mag 16(2):53–60
23. Pelphrey KA, Morris JP, Mccarthy G (2004) Grasping the intentions of others: the perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. J Cogn Neurosci 16(10):1706–1716
24. Pineau J, Gordon G, Thrun S (2003) Point-based value iteration: an anytime algorithm for pomdps. In: Proc of the 18th international joint conference on artificial intelligence (IJCAI)
25. Poupart P, Boutilier C (2004) VDCBPI: an approximate scalable algorithm for large scale pomdps. In: Proc of advances in neural information processing systems (NIPS), vol 17, pp 1081–1088
26. Puterman LM (1994) Markov decision processes. Wiley, New York
27. Schultz J, Imamizu H, Kawato M, Frith CD (2004) Activation of the human superior temporal gyrus during observation of goal attribution by intentional objects. J Cogn Neurosci 16(10):1695–1705
28. Sebanz N, Bekkering H, Knoblich G (2006) Joint action: bodies and minds moving together. Trends Cogn Sci 10(2):70–76
29. Simmons R, Koenig S (1995) Probabilistic robot navigation in partially observable environments. In: Proc of the 14th international joint conference on artificial intelligence, pp 1080–1087
30. Smith T (2007) Probabilistic planning for robotic exploration. PhD thesis, The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA
31. Smith T (2007) ZMDP software for POMDP and MDP planning. http://www.contrib.andrew.cmu.edu/trey/zmdp/
32. Smith T, Simmons R (2004) Heuristic search value iteration for pomdps. In: Proc of the 20th conference on uncertainty in artificial intelligence (UAI)
33. Smith T, Simmons RG (2006) Focused real-time dynamic programming for MDPs: squeezing more out of a heuristic. In: Proc of the 21st national conference on artificial intelligence (AAAI)
34. Striano T, Henning A, Stahl D (2006) Sensitivity to interpersonal timing at 3 and 6 months of age. Interact Stud 7(2):251–271
35. Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. IEEE Trans Neural Netw 9(5):1054
36. Tipaldi GD, Arras K (2011) Planning problems for social robots. In: Proc of the 21st international conference on automated planning and scheduling (ICAPS)
37. Tomasello M, Carpenter M, Call J, Behne T, Moll H (2005) Understanding and sharing intentions: the origins of social cognition. Behav Brain Sci 28:675–735
38. TORCS (2013) The open racing car simulator. http://torcs.sourceforge.net/
39. Walter H, Adenzato M, Ciaramidaro A, Enrici I, Pia L, Bara BG (2004) Understanding intentions in social interaction: the role of the anterior paracingulate cortex. J Cogn Neurosci 16(10):1854–1863
40. Wereschagin M (2006) Pittsburgh left seen by many as a local right. Pittsburgh Tribune-Review, 14 June 2006

Frank Broz is a research fellow in the Adaptive Behaviour and Cognition lab at Plymouth University. His research interests are in artificial intelligence and human–robot interaction, more specifically in modelling human behavior in order to support natural interaction between humans and autonomous agents. He currently works on modelling mutual gaze in conversational pairs to create socially appropriate robot gaze controllers and on multimodal feedback strategies for assistive robots. He received his B.S. in computer science from Carnegie Mellon University and his Ph.D. in robotics from the CMU Robotics Institute. He organized the AAAI spring symposium “It’s All in the Timing: Representing and Reasoning About Time in Interactive Behavior” in 2010 and the ICDL-EpiRob special session on “Social Gaze: From Human-Human to Human–Robot Interaction” in 2011.

Illah Nourbakhsh is Professor of Robotics, director of the Community Robotics, Education and Technology Empowerment (CREATE) lab and head of the Robotics Masters Program in The Robotics Institute at Carnegie Mellon University. His current research projects explore community-based robotics, including educational and social robotics and ways to use robotic technology to empower individuals and communities. He is co-principal investigator of the Global Connection Project, a joint initiative of Carnegie Mellon, NASA, the National Geographic Society and Google Inc. that developed the gigapixel imaging technology known as GigaPan. He is leading projects to apply GigaPan technology to scientific and educational efforts and to use GigaPan to help students communicate with peers internationally. Recent CREATE Lab projects include ChargeCar, a community-based effort to convert gas cars into electric commuter vehicles; HearMe, a project that uses technology to help children communicate their ideas and experiences; and Robot250, a 2008 project to teach Pittsburgh-area citizens how to design and build robots to address community concerns or express ideas. Other major projects include messaging systems for child car centers to improve home-school consistency; the Robot Diaries program for creative art and robotics fusion in middle schools; the Finch programmable mobile robot, and community-empowering air and water quality sensors. His past research has included protein structure prediction under the GENOME project, software reuse, interleaving planning and execution and planning and scheduling algorithms, as well as mobile robot navigation. While on leave from Carnegie Mellon in 2004, he served as Robotics Group lead at NASA/Ames Research Center. He was a founder and chief scientist of Blue Pumpkin Software, Inc., which was acquired by Witness Systems, Inc. Nourbakhsh earned his bachelor.s, master.s and Ph.D. in computer science at Stanford University and has been a faculty member of Carnegie Mellon since 1997. The National Academy of Sciences in 2009 named him a Kavli Fellow. He is co-author of the second edition an MIT Press textbook, Introduction to Autonomous Mobile Robots.

Reid Simmons is a Research Professor and the Associate Director for Education in the Robotics Institute at Carnegie Mellon University. He earned his Ph.D. from MIT in 1988 in the field of Artificial Intelligence. Since coming to Carnegie Mellon in 1988, Dr. Simmons' research has focused on developing self-reliant robots that can autonomously operate over extended periods of time while interacting in unstructured environments. Currently, the research focuses on the areas of coordination of multiple heterogeneous robots, human-robot social interaction (both conversational and spatial interaction), and robust error detection and recovery. Over the years, he has published about 200 papers and has been involved in the development of over a dozen autonomous robots.