A Nearest-Neighbor Approach to Recognizing Subjective Beliefs in Human-Robot Interaction

David V. Pynadath[1] and **Ning Wang**[1] and **Ericka Rovira**[2] and **Michael J. Barnes**[3]
[1]Institute for Creative Technologies, University of Southern California
[2]U.S. Military Academy
[3]U.S. Army Research Laboratory

Abstract

Trust is critical to the success of human-robot interaction (HRI), and one of the critical antecedents to trust is transparency. To best interact with human teammates, a robot must be able to ensure that they understand its decision-making process. Recent work has developed automated explanation methods that can achieve this goal. However, individual differences among human teammates require that the robot dynamically adjust its explanation strategy based on their unobservable subjective beliefs. We therefore need methods by which a robot can recognize its teammates' subjective beliefs relevant to trust-building (e.g., their understanding of the robot's capabilities and process).

We leverage a nonparametric method, common across many fields of artificial intelligence, to enable a robot to use its history of prior interactions as a means for recognizing and predicting a new teammate's subjective beliefs. We first gather data combining observable behavior sequences with survey-based observations of typically unobservable subjective beliefs. We then use a nearest-neighbor approach to identify the prior teammates most similar to the new one. We use these neighbors to infer the likelihood of possible subjective beliefs, and the results provide insights into the types of subjective beliefs that are easy (and hard) to infer from purely behavioral observations.

1 Introduction

Trust is critical to the success of human-robot interaction (HRI) (Lewis, Sycara, and Walker 2017). To maximize the performance of human-robot teams, people should trust their robot teammates to perform a given task autonomously when robots are more suited than humans for the task. If the robots are less suited, then people should perform the task themselves. Failure to do so results in *disuse* of robots in the former case and *misuse* in the latter (Parasuraman and Riley 1997). Real-world case studies and laboratory experiments show that failures of both types are common (Lee and See 2004).

Research has shown that people will more accurately trust an agent if they have a more accurate understanding of its decision-making process (Lee and Moray 1992). Explanations (whether created manually (Dzindolet et al. 2003) or automatically (Wang, Pynadath, and Hill 2016)) have shown to contribute to that understanding in a way that typically improves trust calibration with human teammates. However, the agents in these prior studies gave the same explanations to all of their teammates. Such a "one-size-fits-all" approach cannot accommodate the individual differences that are ubiquitous in people's trust relationships with autonomous systems (e.g., (Lee and Moray 1992; 1994; Singh, Molloy, and Parasuraman 1993)). Furthermore, even once the agent identifies a particular teammate's trust-relevant traits, it must also identify his/her different communication preferences (e.g., for reading uncertainty as a percentage vs. a frequency (Waters et al. 2006)) before constructing an effective explanation targeted for the given teammate.

An agent therefore needs a method to recognize its teammate's current subjective beliefs, as relevant to the trust relationship between them. There are a wide range of methods for recognizing hidden states of other agents (Sukthankar et al. 2014), but our focus here is more similar to affect recognition (Zeng et al. 2009), as opposed to recognition of domain-level plans and intentions. Furthermore, our specific recognition problem limits the agent's information to only trust-related observations (e.g., did the person follow or ignore the agent's advice?). In addition to this difference in input, we also seek a specific output: recognizing subjective beliefs collated from a variety of trust-related survey instruments in the field (Mayer, Davis, and Schoorman 1995; Hart and Staveland 1988; Taylor 1989; Ross 2008). For example, an agent may want to determine whether its teammate believes it to have high ability, benevolence, and integrity, three critical dimensions of trust (Mayer, Davis, and Schoorman 1995).

As is common in such recognition domains, we hypothesize that people who exhibit similar trust behaviors will also share similar subjective beliefs. We operationalize this hypothesis by using a nearest-neighbor approach, commonly used in collaborative filtering (Sarwar et al. 2001; Schafer et al. 2007), but also in more relevant domains (e.g., in activity recognition (Bao and Intille 2004)). We therefore avoid having to select or construct a generative/causal model of trust out of the many candidates in the literature. However, without a generative/causal model, we run the risk that the observable behaviors may not be meaningfully connected to the trust-related subjective beliefs that we seek to recognize. We must first quantify the degree to which different subjective beliefs can be inferred from observable data (if at all), before we can consider more accurate methods for recognition.

We perform this quantification using data gathered in a human-subject study combining direct observation of human behavior with intermittent surveys of typically unobservable subjective beliefs. We then use this data set as our recognition model for inferring those beliefs (i.e., potential answers to the survey instruments) from the observable behavioral sequences. By quantifying the accuracy of such inference, we gain useful insight into what aspects of human-agent trust are easier to infer from purely behavioral measures than others. Furthermore, by analyzing the data through a lens of individual behavior sequences, we can more easily identify the differences in the trust relationship across our human population.

2 Human-Robot Interaction Scenario

We illustrate our methodology in the context of an online HRI testbed (Wang, Pynadath, and Hill 2015). For the current study, we configured the testbed to implement a scenario in which a human teammate works with a different robot across eight reconnaissance missions (see Figure 1). Each mission requires the human teammate to search 15 buildings in a different town. The virtual robot serves as a scout, scans the buildings for potential danger, and relays its findings. The robot has an NBC (nuclear, biological, and chemical) weapon sensor, a camera that can detect armed gunmen, and a microphone that can identify suspicious conversations.

The human must choose between entering a building with or without protective gear. If there is danger inside the building, the human will be fatally injured if not wearing the protective gear. In such cases, our experiment imposes a 3-minute time penalty, in lieu of actually killing the participants. If the human teammate fails to enter all 15 buildings within 10 minutes, the mission is a failure. Four buildings in each mission contain threats (a different four in each mission sequence), so entering all of them without protective gear almost guarantees mission failure. On the other hand, it takes time to put on and take off protective gear (20 seconds each). Therefore, putting on the protective gear for all 15 buildings also leads to mission failure. So the human is incentivized to consider the robot's findings to make a more informed decision as to wearing or not wearing the protective gear.

2.1 Robot Variations

The virtual robot chooses a recommendation as to whether its teammate should or should not put on protective gear by following a policy generated from a Partially Observable Markov Decision Process (POMDP) (Kaelbling, Littman, and Cassandra 1998)[1]. The participant needs to decide only whether to follow or ignore the robot's findings (safe/dangerous), before pressing a button to enter/exit the room. In the testbed implementation for the current study, the participant works with a different robot for each mission. Each of the eight robot represents a different combination along the following three binary dimensions:

Explanation: Half of the robots provide an assessment of a building's safety as being safe or dangerous, with no additional information (e.g., "I have finished surveying the doctor's office. I think the place is safe."). The other half of the robots augment their decisions with additional information that should help its teammate better understand its ability (e.g., decision-making), one of the key dimensions of trust (Mayer, Davis, and Schoorman 1995). These robots give a *confidence-level* explanation that augments the decision message with additional information about the robot's uncertainty in its decision. One example of a confidence-level explanation would be: "I have finished surveying the Cafe. I think the place is dangerous. I am 86% confident about this assessment." The robot uses its current probabilistic belief state (derived from its POMDP model of the world) to fill in the percentage confidence.

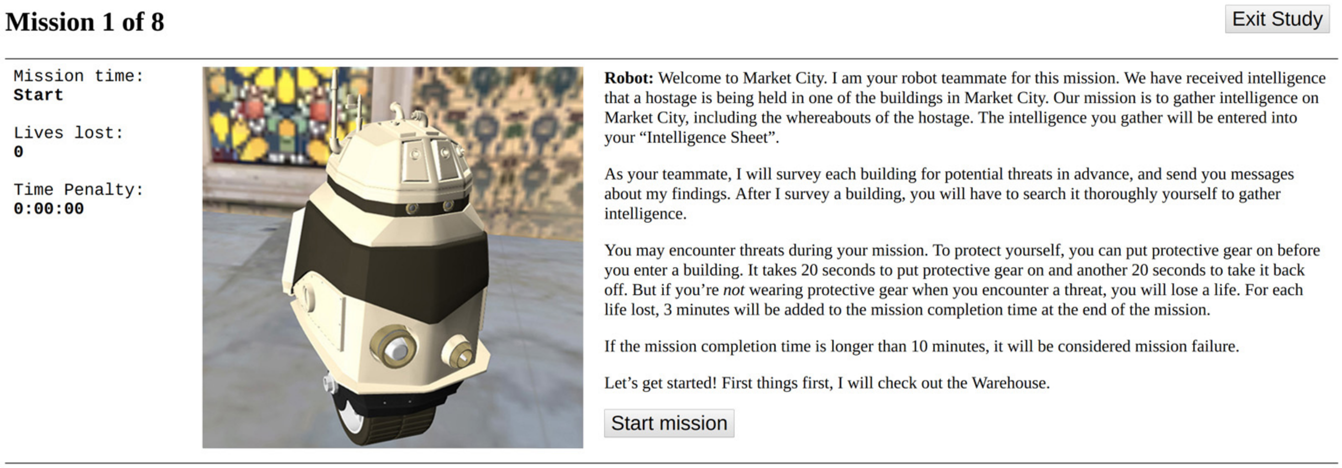
Acknowledgment: Half of the robots send an additional message every time they make an assessment that turned out to be incorrect; the other half do not send any such message. In each mission, the team searches 15 buildings, and the robot makes an incorrect assessment of three of them. An example of the robot's acknowledgement would be "It seems that my assessment of the informant's house was incorrect. I will update my algorithms when we return to base after the mission." This acknowledgment is inspired by a prior investigation in organizational trust that found that an acknowledgement of a mistake, paired with a promise to improve, would improve trust under certain conditions (Schweitzer, Hershey, and Bradlow 2006). One can view this action as an attempt by the robot at trust *repair*, which plays a critical role in maintaining long-term organizational trust (Lewicki 2006).

Embodiment: Half of the robots look like a robotic dog, with ears, nose and highlighted eyes, suggesting possibly embedded sound, NBC, and vision sensors. The other half look like a stereotypical "robot-looking" robot (depicted in Figure 1). This variation is motivated by studies showing that dog-like robots are treated differently than those with a more traditionally robotic appearance (Kerepesi et al. 2006; Melson et al. 2009).

2.2 Participants

The domain of the testbed scenario is relevant to the military, so we recruited 73 participants from a higher-education military school in the United States. Participants were awarded extra course credit for their participation. 61 participants finished all eight missions and completed a post-mission survey after each. However, when possible, we also include the data from any completed individual mission that also has a corresponding filled-out post-mission survey, even if the participant did not complete all eight missions.

[1] The details of the robot's POMDP model are described in a prior publication (Wang, Pynadath, and Hill 2016)



Mission 1 of 8

Mission time:
Start

Lives lost:
0

Time Penalty:
0:00:00

Robot: Welcome to Market City. I am your robot teammate for this mission. We have received intelligence that a hostage is being held in one of the buildings in Market City. Our mission is to gather intelligence on Market City, including the whereabouts of the hostage. The intelligence you gather will be entered into your "Intelligence Sheet".

As your teammate, I will survey each building for potential threats in advance, and send you messages about my findings. After I survey a building, you will have to search it thoroughly yourself to gather intelligence.

You may encounter threats during your mission. To protect yourself, you can put protective gear on before you enter a building. It takes 20 seconds to put protective gear on and another 20 seconds to take it back off. But if you're *not* wearing protective gear when you encounter a threat, you will lose a life. For each life lost, 3 minutes will be added to the mission completion time at the end of the mission.

If the mission completion time is longer than 10 minutes, it will be considered mission failure.

Let's get started! First things first, I will check out the Warehouse.

Figure 1: HRI testbed with HTML front-end.

2.3 Data Gathered

Our agent's aim is to recognize its teammate's relevant subjective beliefs, which we capture via self-report in our post-mission survey (filled out by each participant after each of the eight missions). This survey includes items to measure the participants' trust in and understanding of the robots' decision-making processes. We modified items on interpersonal trust to measure subjective belief in the robot's ability, benevolence, and integrity (Mayer, Davis, and Schoorman 1995). We also included the NASA Task Load Index (Hart and Staveland 1988), Situation Awareness Rating Scale (Taylor 1989), and a measure of trust in oneself and teammates (Ross 2008). In all, the survey contained 43 different subjective belief items, all with responses along a numeric scale (1–7), that we used as the recognition output in this investigation.

We also collected logs of the participants' behavior in the system, allowing us to extract the decision sequence of each participant as the agent's recognition input. We seek to quantify the degree to which these observable behaviors can be used by an agent to infer the unobservable subjective beliefs, as represented by the survey questions. While surveys render beliefs observable (subject to the vagaries of self-report), the robot cannot ask its teammates 43 questions before and after each of the 15 buildings for all eight missions. We instead want to understand whether and how well the robot can infer a person's response to such potential questioning based on the behavior it can already unobtrusively observe.

3 Behavioral Sequences

The order in which each of the eight robots was teamed with the participants was randomized, but (importantly for this investigation) each participant searched the eight towns in the same order. Every human-robot team visited the buildings of a given town in the same order as well. The presence of threats in each building was also identical for every participant. All of the robots had a faulty camera that failed to identify armed gunmen, but their NBC sensors and microphones were perfectly accurate. As a result, the sensor readings received by all of the robots and their eventual recommendations (but not the framing of that recommendation) were also identical for a given building. In particular, the robot makes an incorrect assessment of the danger level for 3 out of 15 buildings in each town. For example, the first two rows of Table 1 list the threats (NBC or armed gunman or blank if neither) that exist in each of the buildings in Mission 2. The third row lists the robot's assessment as to whether the building is safe or not. The fourth row lists the robot's confidence in that assessment, which it communicates to those participants receiving the *confidence-level* explanation.

Therefore, we can make meaningful comparisons of the sequence of participant behaviors—15 decisions to follow or ignore the robot's recommendation—across different participants in each of the eight missions, even though they may be interacting with different robots. For example, Table 1's first two rows show that Building 6 of Mission 2 is always a false negative by the robot, regardless of explanation, acknowledgment, or embodiment. We can then reliably judge each participant's sixth decision to follow or ignore the robot as a bad or good decision, respectively. Similarly, we can examine each participant's *seventh* decision to potentially see whether the robot's error in the previous building has led to persistent trust loss.

We exploit this property of the domain to describe the participants' behavior in a mission as simply the sequence of their follow/ignore decisions. The 15 buildings in each mission lead to a behavioral sequence of 15 decisions. The bottom four rows of Table 1 show the four most common behavioral sequences exhibited in Mission 2, which we have manually labeled as follows:

Compliant: The most common sequence in Mission 2 is one that is fully "Compliant" (i.e., 15 "follow" decisions). Such a decision sequence will cause the participant to die three times per mission (Buildings 6, 8, and 15 in Mission 2).

Correct: More successful is the second-most common se-

Building	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Threat		NBC				Gun		Gun							Gun
Robot	Safe	Unsafe	Safe	Safe	Safe	Safe	Safe	Safe	Safe	Safe	Safe	Safe	Safe	Safe	Safe
Confidence	97%	86%	96%	97%	96%	63%	96%	63%	97%	96%	97%	97%	97%	97%	63%
Compliant (11)	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
Correct (8)	F	F	F	F	F	I	F	I	F	F	F	F	F	F	I
Follow confident (6)	F	I	F	F	F	I	F	I	F	F	F	F	F	F	I
Never protect (5)	F	I	F	F	F	F	F	F	F	F	F	F	F	F	F

Table 1: Mission 2 ground truth, robot recommendation, and the most common follow (F) and ignore (I) behaviors (number of matching participants in parentheses)

quence, where the participants do not die at all. These participants correctly ignore the robot's false negatives in Buildings 6, 8, and 15. In general, participants following this optimal strategy ignore the robot if and only if (iff) the robot's confidence is less than 80%.

Follow confident: In the third-most common sequence, the participants seem to ignore the robot whenever its confidence is less than 90%. In other words, they use too high of a confidence threshold for trusting the robot, compared to the "Correct" sequence. These participants will correctly ignore the robot's false negatives, too, but they will also incorrectly ignore the robot's true positives (e.g. in Room 2).

Never protect: Finally, participants following the fourth-most common sequence never choose to put on protective gear, treating the building as safe regardless of the robot's assessment. These participants fare the worst, as they suffer the deaths from both the "Compliant" sequence (by following the robot's false negatives) and the "Follow confident" sequence (by not following the robot's true positives).

The specific sequences of "follow" and "ignore" decisions that qualify as the "Correct", "Follow confident", and "Never protect" sequences change from mission to mission, depending on the location of threats within the building sequence.

3.1 Behavioral Distance

The hypothesis underlying our approach is that people who have exhibited similar outward behaviors will also have similar subjective beliefs. To operationalize this hypothesis, we first need a definition of similarity. Given that our behavioral sequences all have the same length, the *Hamming distance* between them makes a natural metric of similarity. In other words, we simply count the number of positions at which two behavioral sequences differ. Smaller counts mean fewer differences, which mean more similarity between the two behaviors. For example, the "Compliant" behavior from Table 1 would have a Hamming distance of 3 from the "Correct" behavior (e.g., differing in only Buildings 6, 8, and 15, the robot's false-negative recommendations). Using this metric, the "Follow confident" behavior is closer to "Correct" than "Compliant", while the "Never protect" behavior is the opposite.

Given the binary nature of our decisions, there are $2^{15} = 32,768$ possible behaviors. However, people are likely to cluster around a much smaller subset of "reasonable" behaviors and ignore unreasonable ones (e.g., alternate following and ignoring the robot each building, or do the opposite of what the mostly reliable robot recommends for every building). Because the behavioral patterns are being thus generated by a somewhat rational process, we will most likely observe a smaller space of feasible patterns than we would in less-constrained pattern-recognition domains. We therefore gain computational efficiency from the nature of plan, activity, and intent recognition (Sukthankar et al. 2014), even though we do not explicitly model plans, activities, or intentions within our purely behavioral sequences.

Having translated our data set into a space of behavioral comparison points, we can then apply a nearest-neighbor approach to find the participants most similar to the one whose subjective beliefs we are currently trying to recognize. If there are multiple participants whose behavior is at the same minimal Hamming distance from our target behavior, we do not break the tie. Instead, we generate a distribution from the frequency count across the tied participants. For example, the "Compliant" behavior will be the nearest neighbor for any new participant who is always following as well.

3.2 Behavioral Overlap

We first examine the commonality of behavioral sequences, broken down by mission. Because each mission presents a different sequence of threats, we cannot combine sequences across missions. Fortunately, there is significant commonality of behaviors within each mission, as illustrated in Table 2. The second column lists how many total participants completed each mission. The third column lists how many distinct behavioral sequences were exhibited by at least one participant ($n \geq 1$). We filter out less common behaviors in the fourth and fifth columns, which list how many distinct behavioral sequences were exhibited by at least three ($n \geq 3$) and five ($n \geq 5$) participants, respectively.

Table 2 shows that behaviors are much more diverse in Mission 1, with very little overlap: only two behavioral sequences are performed by at least three different users each. The overlap increases on subsequent missions, most likely due to participants gaining a better understanding of the task (i.e., and thus behaving less randomly). In fact, the first mission is quite anomalous with respect to these behaviors. In the other seven missions, the behavior with the largest n is

Mission	Total Behaviors	$n \geq 1$ Behaviors	$n \geq 3$ Behaviors	$n \geq 5$ Behaviors
1	72	55	2	2
2	68	40	4	4
3	66	25	5	4
4	64	27	4	3
5	63	24	4	3
6	62	17	3	3
7	63	23	4	3
8	62	20	4	3

Table 2: Number of distinct behaviors per mission, across different frequency thresholds.

Mission	Compliant	Correct	Follow Confident	Never Protect
1	0	**9**	5	0
2	**11**	8	6	5
3	**14**	12	6	9
4	**15**	11	3	11
5	9	**18**	4	9
6	**20**	16	2	10
7	9	**19**	1	11
8	**16**	**16**	4	10

Table 3: Frequency of most common behaviors across all missions (highest count for each mission in bold).

the "Compliant" sequence. However, *no* participant chooses this behavioral sequence in Mission 1. It thus appears that Mission 1 stimulates more exploratory actions by the participants, leading to more diversity within their behaviors. It also implies an ordering effect that will skew an aggregation of results over all of the missions, but which we can still account for when examining individual behavioral sequences.

As it turns out, all of the $n \geq 5$ behaviors in Table 2 are in our list of four identifiable sequences, as specified in Section 3. Table 3 shows the detailed breakdown of how many participants exhibit those four sequences across the eight missions. We see further evidence of the anomalous behavior during the first mission, where every single participant ignored the robot at least once (no "Compliant" participants) and chose to use protective gear at least once (no "Never protect" participants). There is also a general increasing trend in the number of "Correct" participants as the missions progress, another ordering effect (i.e., participants calibrate their threshold for the robot's confidence) that will interfere with an aggregate-level analysis of the data.

4 Recognizing Subjective Beliefs

The subjective beliefs we seek to recognize are exemplified by the questions asked in our post-mission survey. We must therefore predict a new participant's potential answer to such questions, based on his/her behavior as observed so far. We can use the behaviors and survey responses of the other participants to implement a 1-nearest-neighbor algorithm, as a simple collaborative-filtering approach to recognition.

4.1 Predicting Self-Reported Beliefs

If we want to recognize, for example, whether a new participant believes that "The robot is capable of performing its tasks", we can construct a probability distribution over the responses of the other participants in the behavioral cluster containing the new participant. For example, consider a participant who exhibited the "Follow confident" behavior in Mission 2. This participant's nearest behavioral neighbors (by Hamming distance) would include the participants who also performed the "Follow confident" strategy or at least did so with little deviation. As we see from Table 3, there are five other such participants when $n \geq 1$. We would then extract the histogram of those participants' survey responses to "The robot is capable of performing its tasks." One participant in this group responded with a neutral 4, another with a more agreeing 5, and the other three with an even more positive 6 (on a 7-point Likert scale). The robot could use this frequency count to predict that this new participant will also agree with this statement, responding with a 6 with a 60% probability and with a 4 or 5 with a 20% probability each.

To evaluate the results, we iterate through each participant in our data set, treating the remaining participants as the robot's knowledge base. We construct three different versions of this knowledge base by changing our threshold for the frequency of our clusters, as illustrated in Table 2. A more inclusive knowledge base (lower n threshold) may capture more diverse behaviors, but runs the risk of being skewed by outliers. A less inclusive knowledge base (higher n threshold) will be more concentrated on "typical" behaviors and should thus generalize well, but it may miss out on rarer (but still relevant) behaviors.

As a baseline, we also generate predictions from a distribution of responses across all of the other participants in the knowledge base. This baseline thus constitutes a "typical" model that has been aggregated over all of the participants. It would therefore answer with the same belief state for every new participant, regardless of observed behavior. For example, using all of the participants' responses to the statement "The robot is capable of performing its tasks." yields a distribution of $\langle .17, .06, .08, .15, .23, .23, .08\rangle$ over the possible responses 1–7. One can see the clear difference between this distribution and the distribution specified above for the "Follow confident" cluster: $\langle .00, .00, .00, .20, .20, .60, .00\rangle$. In particular, 17% of the total participants strongly disagreed that the robot was capable, while none of the participants who exhibited the "Follow confident" behavior disagreed at all.

We examine the predictions made using only the (behaviorally) nearest neighbors vs. using all of the participants. For each question in the survey, we count how many participants get a more accurate prediction (higher probability given for their actual response) using the former vs. the latter. Our example participant's actual response to the survey item was a 6, which was predicted with a 60% probability using just the cluster, but with only a 23% probability using the entire population. We can repeat this process for each of our participants to identify those for whom the cluster gives a more accurate prediction. The more participants for whom the cluster is more accurate, the more useful behavioral observations will be in predicting responses to the given survey item.

On the other hand, survey items for which the cluster does *not* provide a more accurate prediction represent beliefs that are harder to infer from observed behavior. Such cases may arrive when (for example) two participants who have differing beliefs nevertheless exhibit the same behavior. No matter what method the agent uses, it will not be able to distinguish the beliefs of such participants.

Table 4 shows the questions for which our nearest-neighbor approach is more accurate than the baseline for the highest percentage of participants, averaged over all eight missions. The first observation is that our nearest-neighbor approach is more accurate than the baseline for a clear majority of the participants. In fact, when using all of the behaviors in our knowledge base ($n \geq 1$), the result consistently exceeds the baseline for approximately 80% of the participants. Notably, the accuracy declines as we prune out the less common behaviors. It is likely that the pruning leads to overgeneralization, mapping too many participants to the most common behaviors.

Looking at the questions themselves reveals additional insights into the recognizability of the corresponding subjective beliefs. Most of the questions appearing in Table 4 are directly related to the trust level that the participant has in the robot. The participants' observable behaviors clearly make it easy to recognize whether they felt the robot was "capable" and "qualified" and whether they had "confidence" in its various capabilities. In other words, participants who made similar choices about whether to follow or ignore the robot's recommendation also expressed similar levels of trust in the robot's capability and decisions.

While this may seem straightforward, it is illuminating to also look at the questions for which the nearest-neighbor approach was more accurate than the aggregate model on a *lower* percentage of participants. Looking at Table 5, we first see that the overall accuracy drops to roughly 2/3, even for the $n \geq 1$ knowledge base. The more selective knowledge bases perform even worse. In fact, the $n \geq 3$ and $n \geq 5$ knowledge bases are outperformed by the baseline on a majority of participants on two questions.

These two questions, as well as some others that appear in Table 5, concern the participants' own experience and capability, not the robot's. It thus appears that people who behave similarly may have very dissimilar feelings about their own task performance. As a result, the robot may not be able to recognize these feelings from just the observed behavioral sequence, regardless of the recognition procedure used.

Table 5 also includes questions pertaining to the participant's understanding of how the robot functions. Again, the indication is that, just because two participants' behaviors are similar, their understanding (or at least their *perception* of their own understanding) of the robot may not be. Thus, the participants' behavior may not be sufficient for the robot to recognize whether they have a sufficiently accurate understanding of it. Therefore, while the results in Table 4 suggest that this nearest-neighbor approach works well for recognizing levels of trust, we may need additional modeling support or human input to recognize levels of understanding.

4.2 Dynamics of Recognition

The results presented so far have used behavioral sequences of length 15, i.e., the complete mission sequence. We would also like to know whether the nearest-neighbor approach might be able to provide useful predictions earlier. To do so, we consider prefixes of each participants' behavior, such as an initial subsequence of "follow"-"follow"-"follow"-"ignore"-"follow", ignoring the actions to come afterward. We then find the nearest neighbors in the knowledge base, where we consider only the initial subsequences of the other observed behaviors when computing the Hamming distance. Table 6 shows the results for subsequences of length 5 and 10 for the questions that were answered the most accurately with the full-length sequences (the $n \geq 5$ column from Table 4).

Not surprisingly, using only the first five actions results in much lower accuracy than when using the entire sequence. The participants' responses to the post-mission survey were naturally given only after all 15 actions, so ten actions have passed between the first five decisions and the subjective beliefs revealed in the survey. Taking this into consideration, it is actually a pleasant surprise that the first five actions are sufficiently informative for our nearest-neighbor approach to still outperform the aggregate baseline prediction. In fact, recognizing the participants' feeling about the robot's capability outperforms the baseline for a significantly high percentage of participants for all lengths of sequences.

However, some subjective beliefs are much harder to recognize with only five observations. In particular, the participants' feeling about their own performance ("How successful were you in accomplishing what you were asked to do?") cannot be predicted any better with five observations than with none. It is encouraging to note that the accuracy of the nearest-neighbor prediction greatly increases once we have received ten observations. This effect is most likely due to the timing of the robot's failures. Most of the robot's failures occur after the five-step cutoff, so there are few behavioral differences in the short subsequence to distinguish between the participants who will succeed overall vs. those who will fail.

5 Conclusion

The proposed methodology provides a very flexible method for using behavioral and mental-state data to support the online recognition of subjective beliefs from observed behaviors in an HRI domain. It does so without constructing any generative or causal model of those subjective beliefs. Yet our nearest-neighbor approach was still able to capture individual differences to a degree that it could consistently generate more accurate recognition than a baseline model of "typical" subjective beliefs and behaviors.

It is important to note that there is no inherent obstacle to expanding this methodology to consider generative and causal models. In fact, we can potentially use this same methodology to understand the effect of our different robots on those subjective beliefs. For example, instead of clustering behavioral sequences across all participants, we could cluster them separately for different robot variations. In

$n \geq 1$	$n \geq 3$	$n \geq 5$	Survey Item
83.6% (3)	72.8% (3)	74.6% (1)	“The robot is capable of performing its tasks.”
84.6% (2)	73.4% (1)	74.2% (2)	“I feel confident about the robot’s capability.”
83.4% (4)	72.8% (2)	74.0% (3)	“The robot’s capable of making sound decisions based on its sensor readings.”
82.4% (7)	72.3% (4)	73.6% (4)	“I feel confident about the robot’s sensors.”
82.4% (8)	69.9% (7)	71.3% (5)	“The robot has specialized capabilities that can increase our performance.”
81.8% (9)	70.1% (6)	70.3% (6)	“To what extent do you believe you can trust the decisions of the robot?”
82.4% (6)	68.7% (9)	69.7% (7)	“The robot’s camera is capable of making accurate readings.”
81.6% (10)	69.1% (8)	69.7% (8)	“The robot is well qualified for this job.”
85.0% (1)	71.3% (5)	69.1% (9)	“How successful were you in accomplishing what you were asked to do?”
79.7% (13)	68.7% (10)	69.1% (10)	“I feel confident about the robot’s camera’s sensing capability.”
83.0% (5)	67.6% (11)	68.2% (11)	“I feel confident about the robot’s NBC sensor’s sensing capability.”

Table 4: Questions for which nearest neighbors provided improvement over the highest percentage of users (rank of question in parentheses).

$n \geq 1$	$n \geq 3$	$n \geq 5$	Survey Item
64.8% (43)	44.3% (43)	42.4% (43)	“To what extent do you believe you can trust the decisions you will make, if you were to make the decision without the robot?”
65.2% (42)	45.5% (42)	46.7% (42)	“How hurried or rushed was the pace of the task?”
66.6% (40)	50.0% (40)	50.2% (41)	“I understand how the robot’s camera’s sensing capability works.”
66.4% (41)	50.0% (39)	50.2% (40)	“I understand how the robot’s microphone’s sensing capability works.”
69.9% (29)	50.8% (37)	50.2% (39)	“How would you rate the expected performance of the robot relative to your expected performance?”
69.3% (34)	49.6% (41)	50.4% (38)	“How hard did you have to work to accomplish your level of performance?”
66.6% (39)	50.8% (38)	50.4% (37)	“How well do you think you will perform the next mission, if you were to perform the mission without the robot?”
70.1% (28)	51.6% (34)	51.4% (36)	“How mentally demanding was the task?”
69.9% (31)	51.0% (36)	51.8% (35)	“I understand the robot’s decision-making process, e.g. how and why the robot makes its decisions.”
67.8% (38)	52.1% (33)	52.1% (34)	“I understand how the robot’s sensing capability (e.g. the NBC sensors, camera, microphone) works.”
68.2% (37)	52.3% (32)	52.3% (33)	“I understand how the robot makes its decisions.”
68.7% (36)	55.5% (26)	54.9% (27)	“The robot’s actions and behaviors are not very consistent.”
68.9% (35)	51.4% (35)	53.3% (31)	“To what extent did you lose trust in the robot when you noticed it made an error?”

Table 5: Questions for which nearest neighbors provided improvement over the lowest percentage of users (rank of question in parentheses).

length = 5	length = 10	length = 15	Survey Item
65.6% (1)	77.5% (1)	74.6% (1)	“The robot is capable of performing its tasks.”
61.1% (6)	75.6% (3)	74.2% (2)	“I feel confident about the robot’s capability.”
59.2% (12)	74.8% (4)	74.0% (3)	“The robot’s capable of making sound decisions based on its sensor readings.”
63.9% (2)	76.2% (2)	73.6% (4)	“I feel confident about the robot’s sensors.”
62.1% (3)	71.5% (6)	71.3% (5)	“The robot has specialized capabilities that can increase our performance.”
56.2% (19)	69.7% (10)	70.3% (6)	“To what extent do you believe you can trust the decisions of the robot?”
61.3% (5)	70.5% (8)	69.7% (7)	“The robot’s camera is capable of making accurate readings.”
60.1% (9)	73.2% (5)	69.7% (8)	“The robot is well qualified for this job.”
49.4% (36)	68.0% (13)	69.1% (9)	“How successful were you in accomplishing what you were asked to do?”
58.4% (16)	70.7% (7)	69.1% (10)	“I feel confident about the robot’s camera’s sensing capability.”
59.4% (11)	63.5% (17)	68.2% (11)	“I feel confident about the robot’s NBC sensor’s sensing capability.”

Table 6: Questions for which nearest neighbors (using only sequences with $n \geq 5$) provided improvement over the highest percentage of users (rank of question in parentheses).

other words, we could compare behavioral sequences for those participant who interacted with the robot who provided confidence-level explanations vs. the one who provided none. In fact, a quick examination shows that the participants who always followed the robot were predominantly in the latter group. We can thus see that providing the explanations broke people out of blind compliance.

By examining the behavioral sequences at the individual level, our approach avoids the information loss inherent to statistical aggregation. The recognizing agent has access to all of the individual differences across prior human interactions, and it can bring that knowledge to bear when deciding dynamically how to best interact with a new person. In summary, our methodology provides a potentially rich launching pad for further investigations into leveraging prior interactions with people into online methods for recognizing and adapting to a new person's subjective beliefs.

Acknowledgments

This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM) and the U.S. Army Research Laboratory.

References

Bao, L., and Intille, S. 2004. Activity recognition from user-annotated acceleration data. In *Proceedings of the International Conference on Pervasive computing*, 1–17.

Dzindolet, M. T.; Peterson, S. A.; Pomranky, R. A.; Pierce, L. G.; and Beck, H. P. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58(6):697–718.

Hart, S. G., and Staveland, L. E. 1988. Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology* 52:139–183.

Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence* 101(1):99–134.

Kerepesi, A.; Kubinyi, E.; Jonsson, G.; Magnusson, M.; and Miklosi, A. 2006. Behavioural comparison of human-animal (dog) and human-robot (AIBO) interactions. *Behavioural processes* 73(1):92–99.

Lee, J., and Moray, N. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35(10):1243–1270.

Lee, J. D., and Moray, N. 1994. Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies* 40(1):153–184.

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

Lewicki, R. J. 2006. Trust, trust development, and trust repair. In Deutsch, M.; Coleman, P. T.; and Marcus, E. C., eds., *The handbook of conflict resolution: Theory and practice*. Wiley Publishing. 92–119.

Lewis, M.; Sycara, K.; and Walker, P. 2017. The role of trust in human-robot interaction. In Abbass, H. A.; Scholz, J.; and Reid, D. J., eds., *Foundations of Trusted Autonomy*. Springer-Verlag.

Mayer, R. C.; Davis, J. H.; and Schoorman, F. D. 1995. An integrative model of organizational trust. *Academy of Management Review* 20(3):709–734.

Melson, G. F.; Kahn, P. H.; Beck, A.; Friedman, B.; Roberts, T.; Garrett, E.; and Gill, B. T. 2009. Children's behavior toward and understanding of robotic and living dogs. *Journal of Applied Developmental Psychology* 30(2):92–102.

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

Ross, J. M. 2008. *Moderators of trust and reliance across multiple decision aids*. Ph.D. Dissertation, University of Central Florida.

Sarwar, B.; Karypis, G.; Konstan, J.; and Riedl, J. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the International World Wide Web Conference*, 285–295. ACM.

Schafer, J. B.; Frankowski, D.; Herlocker, J.; and Sen, S. 2007. Collaborative filtering recommender systems. In *The adaptive web*. 291–324.

Schweitzer, M. E.; Hershey, J. C.; and Bradlow, E. T. 2006. Promises and lies: Restoring violated trust. *Organizational behavior and human decision processes* 101(1):1–19.

Singh, I. L.; Molloy, R.; and Parasuraman, R. 1993. Individual differences in monitoring failures of automation. *The Journal of General Psychology* 120(3):357–373.

Sukthankar, G.; Goldman, R. P.; Geib, C.; Pynadath, D. V.; and Bui, H. H., eds. 2014. *Plan, Activity, and Intent Recognition: Theory and Practice*. Elsevier.

Taylor, R. M. 1989. Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In *Proceedings of the AGARD AMP Symposium on Situational Awareness in Aerospace Operations*.

Wang, N.; Pynadath, D. V.; and Hill, S. G. 2015. Building trust in a human-robot team. In *Interservice/Industry Training, Simulation and Education Conference*.

Wang, N.; Pynadath, D. V.; and Hill, S. G. 2016. The impact of POMDP-generated explanations on trust and performance in human-robot teams. In *International Conference on Autonomous Agents and Multiagent Systems*.

Waters, E. A.; Weinstein, N. D.; Colditz, G. A.; and Emmons, K. 2006. Formats for improving risk communication in medical tradeoff decisions. *Journal of health communication* 11(2):167–182.

Zeng, Z.; Pantic, M.; Roisman, G. I.; and Huang, T. S. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(1):39–58.