Clustering Behavior to Recognize Subjective Beliefs in Human-Agent Teams

David V. Pynadath, Ning Wang
Institute for Creative Technologies,
University of Southern California
pynadath,nwang@ict.usc.edu

Ericka Rovira
U.S. Military Academy
Ericka.Rovira@usma.edu

Michael J. Barnes
U.S. Army Research Laboratory
michael.j.barnes.civ@mail.mil

ABSTRACT

Trust is critical to the success of human-agent teams, and a critical antecedents to trust is transparency. To best interact with human teammates, an agent explain itself so that they understand its decision-making process. However, individual differences among human teammates require that the agent dynamically adjust its explanation strategy based on their unobservable subjective beliefs. The agent must therefore recognize its teammates' subjective beliefs relevant to trust-building (e.g., their understanding of the agent's capabilities and process). We leverage a nonparametric method to enable an agent 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 perceptions. We then use a nearest-neighbor approach to identify the prior teammates most similar to the new one. We use these neighbors' responses to infer the likelihood of possible beliefs, as in collaborative filtering. The results provide insights into the types of beliefs that are easy (and hard) to infer from purely behavioral observations.

KEYWORDS

Human-agent teams; explainable AI; trust; affect recognition

ACM Reference Format:
David V. Pynadath, Ning Wang, Ericka Rovira, and Michael J. Barnes. 2018. Clustering Behavior to Recognize Subjective Beliefs in Human-Agent Teams. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 9 pages.

1 INTRODUCTION

Trust is critical to the success of human-agent interaction (HAI) [9, 11]. To maximize the performance of human-agent teams, people should trust their agent teammates to perform a task autonomously when they are more suited than humans for the task. When the agents are less suited, then people should perform the task themselves. Failure to do so results in *disuse* of agents in the former case and *misuse* in the latter [14]. Real-world case studies and laboratory experiments show that failures of both types are common [9].

Research has shown that people will more accurately trust an agent if they have a more accurate understanding of its decision-making process [7]. Explanations (whether created manually [3] or automatically [22]) 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 its 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., [7, 8, 19]). 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 [23]) before constructing an effective explanation targeted for the given teammate.

An agent therefore needs to recognize its teammate's current subjective beliefs, as relevant to the trust relationship between them. There is a wide range of methods for recognizing hidden states of other agents [20], even trust-relevant hidden states [2]. Our focus here is more similar to affect recognition [24], rather than 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 its 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 [4, 12, 15, 21]. 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 [12].

As is common in recognition domains, we hypothesize that people who exhibit similar behaviors when interacting with the agent will also share similar subjective beliefs. We operationalize this hypothesis by using a nearest-neighbor approach, commonly used in collaborative filtering [16, 17], but also in more relevant domains (e.g., in activity recognition [1]). 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.

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

2 HUMAN-AGENT INTERACTION SCENARIO

We illustrate our methodology in the context of an online HAI testbed[1]. 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: it 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) [5][2]. 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 [12]. 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 turns 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 [18]. 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 [10].

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 [6, 13].

[1]The details of the testbed appear in a prior publication, omitted for blind review
[2]The details of the POMDP appear in a prior publication, omitted for blind review

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.

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 [12]. We also included the NASA Task Load Index [4], Situation Awareness Rating Scale [21], and a measure of trust in oneself and teammates [15]. 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.

Figure 1: HRI testbed with HTML front-end.

3 BEHAVIORAL SEQUENCES

The order in which the eight robots were 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 accurately 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 sequence, 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

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)

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 [20], 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 a great deal of 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.

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).

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 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 did so with little deviation). As Table 3 shows, there are five other such participants when $n \geq 1$. We 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 take each participant in our data set, treating the remaining participants as the robot's knowledge base. We construct 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 risks 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 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 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 others that appear in Table 5, concern the participants' own experience and capability, not the robot's. People who behave similarly may thus have very dissimilar feelings about their own task performance. As a result, the robot

$n \geq 1$		$n \geq 3$		$n \geq 5$		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 improved over the highest percentage of users (rank 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 improved over the lowest percentage of users (rank in parentheses).

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

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 $n \geq 5$) improved over the highest percentage of users (rank in parentheses).

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.

4.3 Recognizing the Agent's Effect on its Teammate

In the analysis so far, we viewed the agent as performing *keyhole* recognition, where it observes its teammate's behavior passively and tries to interpret it. However, the reality is that the agent performs such recognition only to serve its own decision-making on how best to interact with its human teammate. The agent's decisions will then affect its teammates' beliefs and behavior, hopefully improving the overall team's performance. Fortunately, we can use behavioral sequences to recognize these effects as well. Furthermore, we can do so without using the survey and instead relying on only the objective behaviors of the person and agent.

If we give the robot the freedom to choose its explanation, acknowledgment, and embodiment (i.e., choosing which of the eight variations of Section 2.1 to be), then it needs to understand the effect of that choice on its current teammate. Table 7 shows the number of participants whose behavior falls into one of the four main clusters under each possible variation. It is clear from the different distribution that appears in the top vs. bottom half of the graph that the presence or absence of additional transparency (i.e., communicating the robot's confidence level) is the key determining factor. For robots that give no explanation, participants gravitate toward either following the robot's recommendation every time or never choosing to put on protective gear (regardless of the robot's recommendation). Indeed, once the participants realize that the robot will occasionally make a mistake, they are left to guess when those mistakes might occur when no confidence level is provided. In fact, no participants were able to be 100% correct with no explanation from the robot, although 9 came close through luck.

When the robot provides its confidence level as well, the participants most often behave correctly, although a large number of them fall into the "Follow Confidence" behavior, miscalibrating the threshold for trusting the robot's confidence. Therefore, the confidence-level explanation *should* be sufficient for the participants to follow the "Correct" strategy, yet many do not. We can examine Table 7 to identify other effects that might suggest ways in which the agent could distinguish and possible repair the decision failures its teammate is making. However, at this aggregate level of granularity, we do not see any effect of acknowledgment or embodiment when a confidence-level explanation is given.

We instead need to look at the individual participants' behaviors, focusing on only the conditions when a confidence-level explanation was provided. In particular, for each participant, we identify the conditions (if any) under which s/he followed the "Correct" strategy. For most of these participants, there is no obvious interaction between the acknowledgment, embodiment, and participant's correctness. For example, a few participants made no errors when working with the dog-shaped robot that acknowledged its errors and when working with the robot-shaped robot that did not acknowledge them, but they did make errors when working with the other two robots. In other words, there is no consistent effect of acknowledgment and embodiment for these participants.

Fortunately, there were many participants who *did* exhibit a consistent dependency between their correctness and the robot's variation (the number of participants are in parentheses):

Always Correct (11): These participants made correct decisions for all four missions with robots that offered a confidence-level explanation. For these participants, the acknowledgment and embodiment did not matter.

Correct iff no acknowledgment (3): These participants made correct decisions for both missions with a robot who offered

Explanation	Acknowledgment.	Embodiment.	Compliant		Correct		Follow Confident		Never Protect	
No	No	Robot	40	(17)	2	(0)	0	(0)	20	(13)
No	No	Dog	41	(23)	2	(0)	0	(0)	18	(12)
No	Yes	Robot	39	(27)	3	(0)	0	(0)	19	(8)
No	Yes	Dog	40	(21)	2	(0)	1	(0)	19	(10)
Yes	No	Robot	2	(2)	37	(32)	14	(7)	9	(5)
Yes	No	Dog	5	(1)	33	(25)	17	(8)	7	(6)
Yes	Yes	Robot	2	(1)	36	(29)	15	(6)	9	(4)
Yes	Yes	Dog	1	(0)	40	(22)	12	(8)	9	(6)

Table 7: Size of behavior clusters under each of the eight robot variations (number of exact matches in parentheses)

confidence-level explanations, but did not acknowledge errors with a promise to learn. These participants made incorrect decisions when the robot did offer such an acknowledgment. Embodiment had no effect.

Correct iff embodiment is dog (1): One participant made correct decisions for both missions in which the robot offered confidence-level explanations and had a dog-like appearance. This participant made incorrect decisions when the robot looked like a robot, but acknowledgment had no effect.

Correct iff embodiment is robot (6): In contrast to the previous participant, these participants made correct decisions when interacting with the robot-like robot, but made mistakes with the dog-like robot.

Never Correct, always Compliant (1): One participant followed the robot's recommendation throughout all of the missions and thus never followed the "Correct" strategy with any robot.

Never Correct, always Never protect (3): Three participants never chose to use protective gear and thus never followed the "Correct" strategy. Their behavior represents less of an issue of trust and more an issue of motivation, in that the ingame time penalty for dying did not outweigh the real-world wait time of putting on protective gear.

Never Correct, other (10): These participants never followed the "Correct" strategy, but never followed any of the other three identified strategies either.

Although these groups represent small samples, they suggest possible strategies for our agent. For example, offering the acknowledgment has no effect for most of these groups and even has a negative effect on the second one. Our robot should probably not employ such an acknowledgment, unless it can also perform the machine learning to carry through on its promise of improvement. The data also suggest that the dog-like embodiment is more likely to be detrimental than a more robot-like one. However, the agent should be alert to the possibility that a specific teammate may prefer the dog embodiment, so that if it notices errors being made despite its explanation, it should consider changing. Unfortunately, this current level of analysis does not provide much insight into how to get the participants who make mistakes in every mission to work effectively with the agent. Of course, by gathering more data, we can repeat this same methodology to potentially arrive at more robust conclusions.

5 CONCLUSION

The proposed methodology provides a very flexible method for using behavioral and belief data to support the online recognition of subjective beliefs from observed behaviors in an HAI domain. It does so without constructing any generative or causal model of those 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" beliefs and behaviors. Perhaps more importantly, the results provide a metric on the recognizability of different beliefs when given only behavioral data. The focus on individual participants' behaviors also provides insight into ordering effects and other anomalies that may be obfuscated within aggregate data.

It is important to note that there is no inherent obstacle to expanding this methodology to inform generative and causal models as well. In fact, we can potentially use this same methodology to understand the effect of our different robots on those subjective beliefs. For example, one obvious next step is to examine the groups identified in Section 4.3 to identify any significant differences in survey responses between (for example) those who always interpret the robot's explanation correctly and those who always make at least one mistake. A cursory analysis found some general trends (e.g., those who were always correct generally responded more positively to "In general, people really do care about the well-being of others."). However, more analysis is needed to establish harder evidence for potential causation.

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 teammate'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

[1] Ling Bao and Stephen Intille. 2004. Activity recognition from user-annotated acceleration data. In *Proceedings of the International Conference on Pervasive computing*. 1–17.

[2] Robin Cohen and Rino Falcone Timothy Norman (Eds.). 2016. *International Workshop on Trust in Agent Societies*.

[3] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58, 6 (2003), 697–718.

[4] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology* 52 (1988), 139–183.

[5] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence* 101, 1 (1998), 99–134.

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

[7] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270.

[8] John D Lee and Neville Moray. 1994. Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies* 40, 1 (1994), 153–184.

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

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

[11] Michael Lewis, Katia Sycara, and Phillip Walker. 2017. The role of trust in human-robot interaction. In *Foundations of Trusted Autonomy*, Hussein A. Abbass, Jason Scholz, and Darryn J Reid (Eds.). Springer-Verlag.

[12] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An integrative model of organizational trust. *Academy of Management Review* 20, 3 (1995), 709–734.

[13] Gail F. Melson, Peter H. Kahn, Alan Beck, Batya Friedman, Trace Roberts, Erik Garrett, and Brian T. Gill. 2009. Children's behavior toward and understanding of robotic and living dogs. *Journal of Applied Developmental Psychology* 30, 2 (2009), 92–102.

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

[15] Jennifer Marie Ross. 2008. *Moderators of trust and reliance across multiple decision aids*. Ph.D. Dissertation. University of Central Florida.

[16] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the International World Wide Web Conference*. ACM, 285–295.

[17] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. Collaborative filtering recommender systems. In *The adaptive web*. 291–324.

[18] Maurice E. Schweitzer, John C. Hershey, and Eric T. Bradlow. 2006. Promises and lies: Restoring violated trust. *Organizational behavior and human decision processes* 101, 1 (2006), 1–19.

[19] Indramani L Singh, Robert Molloy, and Raja Parasuraman. 1993. Individual differences in monitoring failures of automation. *The Journal of General Psychology* 120, 3 (1993), 357–373.

[20] Gita Sukthankar, Robert P. Goldman, Christopher Geib, David V. Pynadath, and Hung Hai Bui (Eds.). 2014. *Plan, Activity, and Intent Recognition: Theory and Practice*. Elsevier.

[21] R. M. Taylor. 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*.

[22] Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. The impact of POMDP-generated explanations on trust and performance in human-robot teams. In *International Conference on Autonomous Agents and Multiagent Systems*.

[23] Erika A. Waters, Neil D. Weinstein, Graham A. Colditz, and Karen Emmons. 2006. Formats for improving risk communication in medical tradeoff decisions. *Journal of health communication* 11, 2 (2006), 167–182.

[24] Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 1 (2009), 39–58.