Field Trial Analysis of Socially Aware Robot Assistant

Socially Interactive Agents Track

Florian Pecune
Carnegie Mellon University
fpecune@andrew.cmu.edu

Jingya Chen
Tsinghua University
chenjingya13@mails.tsinghua.edu.cn

Yoichi Matsuyama
Carnegie Mellon University
yoichim@cs.cmu.edu

Justine Cassell
Carnegie Mellon University
justine@cs.cmu.edu

ABSTRACT

The Socially-Aware Robot Assistant (SARA) is an embodied conversational agent that works toward using detection of visual, vocal and verbal cues as an input to estimate the strength of its relationship (namely the level of rapport) with a user. SARA then answers to the user through similar visual, vocal and verbal behaviors with the goal of building and maintaining rapport with that user as we hypothesize that this will improve task performance and user satisfaction over time. In this paper, we report results of a field trial with a semi-automatic SARA system that took place in a large high-profile conference. Participants interacted with SARA during the whole conference, receiving recommendations about sessions to attend and/or people to meet. We analyzed these interactions to shed light on the dynamics of the rapport level between SARA and the conference attendees, and investigate how SARA's task performance would influence the evolution of rapport over time. Although we did not find evidence supporting our claim that the recommendations' outcomes would influence rapport dynamics, our findings emphasize the importance of interactional features plays in both rapport and SARA's task performance. We thus propose design guidelines for the next generation of socially aware personal assistants.

KEYWORDS

Socially-aware; personal assistant; rapport

ACM Reference Format:
Florian Pecune, Jingya Chen, Yoichi Matsuyama, and Justine Cassell. 2018. Field Trial Analysis of Socially Aware Robot Assistant. 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

As we are living in an era of artificial intelligence, we are getting more comfortable with the idea of interacting with machines in our everyday life [42]. Artifacts such as Embodied Conversational Agents (ECAs) [10] are specifically designed to facilitate these interactions, providing a natural, human-like user interface capable of reproducing the different modalities of human communication such as speech, gestures, facial expressions or gaze.

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.

While Bledsoe's dream [7] of seeing humans befriend machines is still a long way off, many efforts are being made to make these ECAs capable of building and maintaining a relationship in the long run with their users. Indeed, this social bond also known as *rapport* has been shown to increase ECA performance during tasks such as tutoring [33], health coaching [21], or museum guidance [6].

One particular role that would greatly benefit from rapport building mechanisms is that of personal assistant [2]. Indeed, companies large and small are now moving forward with the vision of intelligent virtual personal assistants such as Apple's Siri, Microsoft's Cortana or Amazon's Alexa. However, these current personal assistants only provide a vocal interface and do not yet allow multimodal input, or provide embodied output to their users. Increasingly, these same companies have begun to investigate adding social chitchat, but their approach is not grounded in human behavior. They lack the social awareness and reasoning that would allow them to sense and generate relevant social language leading to increase rapport with their user.

In this paper, we report results of a field trial with a semi-automatic socially aware personal assistant that helped participants of a large conference to find relevant sessions to attend and interesting people they should meet. Rather than simply delivering information through a textual interface, or a plain dialogue, we designed our personal assistant to build a relationship with the conference attendees through a multimodal rapport-building dialogue. Conference attendees interacted with our personal assistant during the conference, getting recommendations about sessions to attend and/or people to meet. Our main contribution here is to analyze these interactions and (1) investigate the relationship between the rapport dynamics and the task performance of our personal assistant and (2) propose design guidelines for personal assistants and socially aware ECAs in general.

2 RELATED WORK

Since 2011, the Siri's debut, major tech companies have released a number of voice-based intelligent personal assistants, such as Microsoft's Cortana, and Amazon's Alexa. One of the origins of such intelligent personal assistants might be CALO, the Cognitive Agent that Learns and Organizes, a foundation technology of Siri. CALO is able to handle major cognitive tasks, such as task and schedule management [5][25], and human communication mediation [39]. Similarly, the RADAR project developed a software-based personal assistant to help users cope with email overload as effectively as a human assistant. The system analyzes text messages received by

the user to distill out task-relevant information including new tasks elicited by a message[16]. The major purpose of these research prototypes and commercial products is to support users' cognitive tasks. They only focus on the task aspect of the interaction without taking into account the social cues delivered by their users. Some other work, however, has started to investigate the positive influence of rapport on their agents' task-performance. Rea is a real-estate virtual agent who talks about apartments to rent while building trust and rapport with her users through the use of social language and small-talk [11]. The authors found that extroverted users trusted the social version of Rea more than the version that only focused on the task itself. However, there was no evidence to suggest that users who engaged with social Rea would pay more for a recommended apartment than users who interacted with Rea's task-only counterpart. In [6], the authors deployed Tinker, a virtual museum guide designed to describe various exhibits to guests and then help them find their way out. Tinker was able to build rapport with its users through a short dialogue using predetermined strategies. The authors reported that Tinker's use of relational behavior improved users' engagement, which consequently improved the amount of information retained by the users about museum exhibits. Ellie [14] is a virtual agent designed to have engaging interactions in which users would feel comfortable sharing and disclosing information. Ellie uses non-verbal behavior and a set of dialogue policies to build rapport with its users. In terms of rapport, the authors discovered that people who interacted with a fully autonomous version of Ellie reported feelings comparable to people who had a face-to-face interview with a human semi-expert. Furthermore, most of the participants (75.8%) who interacted with autonomous Ellie agreed that they were comfortable sharing information with it. Most of these works rely on self-rated rapport score to investigate the relationship between rapport and the agent task-performance. None, however, try to capture the evolution of rapport *during* the interactions nor do they investigate whether an agent's task performance could affect the dynamics of rapport over time. We try to address these gaps by shedding the light on the following research question:

RQ: *"How does the task performance of a personal assistant affect the dynamics of rapport over the course of an interaction?"*

In this paper, we hypothesize that rapport between a personal assistant and its user is likely to increase if the former achieves high task-performance. On the other hand rapport is likely to decrease if the personal assistant fails to achieve its task.

3 SYSTEM DESIGN AND ARCHITECTURE

3.1 System Design

We deployed our Socially Aware Robot Assistant (SARA) during a large high-profile conference. The participants of this conference included CEOs, politicians, representatives from academia, NGOs, religious leaders, their spouses, as well as journalists. The conference lasted five days and was filled with discussions, lectures, workshops, and showcases. During the conference, many sessions were happening simultaneously and a number of public and private parties were organized by companies and countries. Attendees were typically busy during the days to achieve their own goals, such as making network for their business and learning about new technologies. SARA was designed to help the conference attendees by being a matchmaker: recommending sessions, professional contacts (even by breaking boundaries of their social status), restaurants, parties, and even leisure activities. Given such requirements, we defined the following essential goals of SARA:

Task Goals:

- Be consistent by recommending relevant items
- Help achieve guests' goals by providing information related to users' interests

Interpersonal Goals:

- Establish a good relationship by using conversational strategies during interactions
- Encourage attendees to disclose more intimate information

Interactional Goals:

- Ease interaction frustration by minimizing response latency

The SARA booth was centrally located in the middle of conference center's main corridor. SARA had access to the conference database of sessions, participants, demos, food vendors, and private parties. She assisted the global leaders by finding out about their interests and goals and then recommending sessions and people relevant to these desideratum. She usually asked for feedback after each recommendation, checking whether the recommendation displayed on the board behind her (see fig.1) matched attendee's interests. In the case of a positive response, SARA was able to send a session reminder or introduce the attendee to the person recommended using an online collaboration platform designed for the conference. Although attendees had access to this online collaboration platform, journalists and attendees' guests did not have such accounts. In these cases, SARA suggested to take pictures using their smart phones.

3.2 System Architecture

The system we describe in this paper was built on top of our SARA platform [24] that was demoed during a previous conference. The system's architecture is organized around a task-pipeline and a social-pipeline. The task-pipeline consists of a task-oriented Natural Language Understanding (NLU), extracting user's intention from its speech, and a Task Reasoner selecting SARA's next intention based on the NLU's output. The social-pipeline consists of three different modules. The Conversational Strategy Classifier detects user's conversational strategy based on user's multimodal cues, the Rapport Estimator relies on these conversational strategies as well as visual and acoustic features to predict the level of rapport going on during the interaction, and the Social Reasoner selects SARA's next conversational strategy based on the history of the interaction. Given the system's task and social intentions decided by the Task and Social Reasoners, a Natural Language Generator (NLG) and Nonverbal Behavior Generator interpreted these intentions into a sentence and nonverbal behavior plans rendered on SARA's character animation realizer and Text-to-Speech (TTS). The system also had access to the recommendation database, user authentication and messenger applications of the online collaboration platform system. Because exhibiting in this conference was very high-stake, we decided to

operate the system in a semi-automatic way to avoid critical recognition and decision making errors. The Task Reasoner suggested the best possible task actions, and a human operator (or Wizard of Oz) approved them each time by clicking a button on a control panel, while the other modules were operated autonomously. In this way, when SARA received unexpected open-ended questions from a user, the human operator was able to intercept the Task Reasoner outcome, and selected proper sentences from a fall-back utterance list.

3.2.1 Speech and Head Gesture Extraction. The Automatic Speech Recognizer (ASR) leverages the REST Speech API of Microsoft Cognitive Services that sends continuous audio streams from the microphone, listen to detect the voice and process speech to return recognized texts which are sent to the NLU and the Conversational Strategy Classifier. The system also uses Open Smile [15] to extract acoustic features from the audio signal. Extracted features includes fundamental frequencies (F0), loudness (SMA), jitter and shimmer. OpenFace is an open source framework that implements state-of-the-art facial behavior analysis algorithms, including: facial landmark detection, head pose tracking, eye gaze and Facial Action Unit estimation [3]. It detects 3D facial landmarks, tracks the head pose, gaze and Action Unit activations with respect to the Camera axis. Both acoustic and visual features are sent to the rapport estimator and the conversational strategy classifier modules.

3.2.2 Conversational Strategy Classifier. We have implemented a conversational strategy classifier to automatically recognize the user's conversational strategies - particular ways of talking, that contribute to building, maintaining or sometimes destroying a budding relationship. These include self-disclosure (SD), elicit self-disclosure (QE), reference to shared experience (RSD), praise (PR), violation of social norms (VSN), and back-channel (BC). By including rich contextual features drawn from verbal, visual and vocal modalities of the speaker and interlocutor in the current and previous turns, we can successfully recognize these dialogue phenomena with an accuracy of over 80% and with a Cohen's kappa κ over 60% [45]. The Conversational Strategy Classifier then sends its result to the Rapport Estimator to predict the level of rapport.

3.2.3 Rapport Estimator. We used the framework of temporal association rule learning [18] to perform a fine-grained investigation into how sequences of interlocutor behaviors signal high and low interpersonal rapport. The behaviors analyzed include visual behaviors such as eye gaze and smiles, and verbal conversational strategies, such as self-disclosure, shared experience, social norm violation, praise and back-channels. We developed a forecasting model involving two-step fusion of learned temporal associated rules. The estimation of rapport comprises two steps: in the first step, the intuition is to learn the weighted contribution (vote) of each temporal association rule in predicting the presence/absence of a certain rapport state (via seven random-forest classifiers); in the second step, the intuition is to learn the weight corresponding to each of the binary classifiers for the rapport states, in order to predict the absolute continuous value of rapport (via a linear regression model) [46]. Ground truth for the rapport state was obtained by having naive annotators rate rapport between two interactants in the teen peer-tutoring corpus [22] for every 30 second slice of an hour long interaction (those slices were randomized in order before being presented to the annotators so that ratings were of rapport states and not rapport deltas).

3.2.4 Natural Language Understanding and Task Reasoner. NLU takes uni-gram and bi-gram features and previous user's and system's intentions to classify a user's intention using logistic regression. The model was trained with a dataset collected in former conferences. NLU also extracted keywords using named entity recognition and a list of predefined keywords. The Task Reasoner was initially designed as a probabilistic finite state machine whose transitions are governed by a set of rules and learned probability distribution. The Task Reasoner takes user's intentions recognized by NLU as its inputs, and transition to a new state, based on the current state of the dialog and other contextual information (e.g., how many sessions it has recommended). During the conference, the transition probability was regularly updated to adapt for the majority of the user's behaviours.

3.2.5 Social Reasoner. The social reasoner was designed as a spreading activation model [23][30] - a behavior network consisting of activation rules that govern which conversation strategy the system should adopt next [31]. Taking as inputs the system's phase (e.g. "recommendation"), the system's intentions (e.g. "elicit_goals", "recommend_session"), the history of the user's conversational strategies, select non-verbal behaviours (e.g. head nod and smile) and the current rapport level, the activation energies are updated, the Social Reasoner selects the system's next conversational strategy, and sends both the system's intent and conversational strategies to the NLG. The activation pre-conditions of the behavior network are inspired by the analysis carried out on the peer-tutoring corpus and the personal assistant WoZ corpus.

3.2.6 NLG and BEAT. Given the system's intention (which includes the current conversational phase, the task intent, and the conversational strategy) these modules generate sentence and behavior plans. NLG selects certain syntactic templates associated with the system's intention from the sentence database. The templates are filled with content items. A generated sentence plan is sent to BEAT, a nonverbal behavior generator [13], and BEAT generates a behavior plan in the BML (Behavior Markup Language) form [20]. The BML is then sent to SmartBody [36], which renders the required non-verbal behaviours and the speech of the agent.

3.2.7 Online Collaboration Platform APIs. The SARA system was connected to the conference's online collaboration platform backend system, through user authentication API recommendation and search API, messenger API. When a user came into the booth, s/he was asked to swipe her/his ID badge on a badge reader, then the system loaded a user ID and profile in a secure way. The authenticated user ID allowed the SARA system to access to the online collaboration platform resources. Given keywords extracted by NLU, the recommendation API returned items. With the authenticated user ID and the other attendee's ID she recommended, SARA system was able to send messages to both of them through the messenger API.

Table 1: List of features annotated for each interaction

Feature Type	Feature Description	Feature Abbr.	Feature Value
Interactional Features	SARA average response time	SARA_RespTime	Seconds
	User average response time	User_RespTime	Seconds
	Balance of response time	Balance_RespTime	%
	SARA average word count per sentence	SARA_WordCount	Number
	User average word count per sentence	User_WordCount	Number
	Balance of word count	Balance_WordCount	%
	SARA number of interruptions	SARA_Interruptions	Number
	User number of interruptions	User_Interruptions	Number
	Total duration of the interaction	Total_Duration	Seconds
	Total number of turns	Total_Turns	Number
Interpersonal Features	Average rapport score	Rpt_Avg	Score from 1 to 7
	Rapport utopy	Rpt_Utopy	Score from -1 to 1
Propositional Features	Direct messages from SARA accepted	Messenger_Yes	%
	Pictures of the reco. taken proactively	Pic_Pro	%
	Pictures of the reco. taken when asked	Pic_Asked	%
	Person already known/ session signed up	Reco_Known	%
	Session already over	Bad_RecoOver	%
	Irrelevant reco. (different domain)	Bad_DiffDom	%
	Relevant reco. (but need more precise request)	Close_PrecDom	%
	Relevant reco. (but need more precise region)	Close_PrecReg	%
	Refused without explanation	Refused_NoReason	%

4 FIELD TRIAL

Participants interacted with SARA during the conference, receiving recommendations about sessions to attend and/or people to meet. After the attendees entered the booth, SARA first introduced herself and asked several questions about the attendees' current feelings and mood. Then, the attendees were asked about their occupation as well as their interests and goals for attending the conference. SARA would then cycle through several rounds of people and/or session recommendations, showing information about the recommendation on the virtual board behind her (see Fig.1). The attendees were able to request as many recommendations as desired, and were able to leave the booth anytime they wanted. Finally, SARA proposed to take a "selfie" with the attendees before saying farewell. During each interaction, attendees' video and audio were recorded using a camera and a microphone. SARA's animations, for their part, were recorded separately in a log file. Audio records were used to get text transcriptions of both attendee's and SARA's utterances using a third party transcription service. These transcriptions contained turn-taking information such as speaker ID and starting and ending timestamps for each turn. With rapport being a dyadic phenomenon, we eventually reconstructed the interactions to have both attendee and SARA present in the same video before annotating them.

Our corpus contains data from 69 of these interactions, including both attendee's and SARA's video, audio and textual speech transcription, which combined accounted for more than 5 hours of interaction (total time = 21055 seconds, mean session duration = 305.15 seconds, SD = 65.00 seconds). Out of these 69 attendees, 29 were women and 40 were men. We did not gather any information about the attendees' age or nationality.

To answer to our research question, and understand the relationship between rapport and SARA's task performance, we study three different levels of features: (1) the *interactional* features, as represented by objective low-level features characterizing interactional mechanisms, such as turn taking. (2) The *interpersonal* features represent subjective features describing the relationship between SARA and the attendee. (3) The *task* features are metrics that we used to measure and evaluate SARA's task performance. Interactional level features are automatically extracted and calculated based on the textual transcriptions of the interactions. Interpersonal and task-related features are annotated manually. Table 1 lists all the features studied in this paper.

4.1 Interactional features

Our selection of interactional level features is motivated by previous work. According to [12], turn-taking information such as interaction length, participants word count per sentence or number of turns have an influence on the relationship between two persons. Furthermore, the response time [19](i.e the duration after which one participant takes its turn) and the interruptions occurring during the interaction [9] are also good indicators of the relationship. As a first step, we only consider the number of interruptions for both SARA *(SARA_Interruptions)* and the attendee's *(User_Interruptions)*. Temporal information such as interaction length *(Total_Duration)*, number of turns *(Total_Turns)* and response time were computed using the transcriptions time-stamps. Word count and response time were calculated for both SARA *(SARA_WordCount, SARA_RespTime)* and the attendee *(User_WordCount, User_RespTime)*. We also computed the balance of each of these two features *(Balance_RespTime, Balance_WordCount)*.

4.2 Interpersonal features

We relied on prior work to measure the interpersonal relationship between SARA and the attendee, namely the rapport score [37]. All the reconstructed videos of the interactions were segmented into 30 seconds video 'slices', and randomly reordered afterwards following [1]'s guidelines to avoid any bias. First, we trained four different naive annotators based on the following protocol: the four annotators were given a general definition of rapport based on [37] and watched examples of interactions previously annotated as low,

neutral and high rapport. Then, each annotator rated the same set of 20 slices using a 7 points likert scale. After each annotation round, we calculated Inter-Rater Reliability scores and discussed with the four annotators about their agreements and disagreements. Once we reached acceptable agreement (Krippendorf's $\alpha = 0.68$), each annotator rated one fourth of the entire dataset. We then calculated the average rapport score *(Rpt_Avg)* for each session.

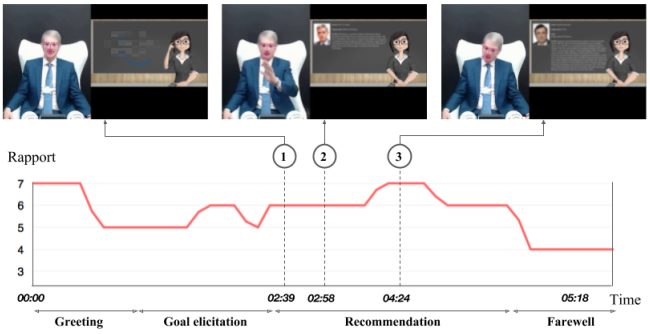
In addition to the average rapport score, and in order to capture the evolution of rapport over time, we also characterized each session through a metric called rapport utopy *(Rpt_Utopy)* [32]. For each interaction, we first built a transition probability matrix based on the thin slices annotations. This squared matrix of size n represents the probability that rapport goes from one score to another during the interaction, weighted by the difference in stepwise increase/decrease. For instance, an increase from 3 to 6 is more important than an increase from 3 to 4. The upper part of this matrix thus represents positive transitions (rapport increases from one slice to following one) while the lower part represents negative transitions (rapport decreases from one slice to the following one). We then calculate the sum Up of each value in the upper part of the matrix, as well as *Low* the sum of each value in the lower part. The final value of utopy for an interaction is computed using the following formula: $Rpt\_Utopy = (2 * (Up - Low))/(n * (n-1))$. A utopy score <0 means that rapport is likely to decrease during the interaction. A utopy score >0 means that rapport is likely to increase during the interaction.

4.3 Task features

To assess the relevance of SARA's recommendations, and to measure her task performance, we annotated the result of each recommendation, based on the attendee's responses and reactions and clustered them in four different categories: accepted, relevant, bad and rejected recommendations. For each of these annotations, we calculated different ratios for sessions, people and overall recommendation (both sessions and persons).

4.3.1 Accepted Recommendations. Our first feature is representing acceptance rate of direct message from SARA. As explained in section 3.1, SARA asked attendees whether they would accept to receive a reminder each time a session was displayed on the screen. Hence, we considered and annotated a session recommendation as accepted when the attendee accepted to receive a message as a reminder. Likewise, people displayed on the screen were considered a good match whenever the attendee accepted to be introduced via a message. The message acceptance rate (*Messenger_Yes*) for one interaction was therefore calculated by the ratio of message accepted to the number of recommendations displayed during the interaction. We also annotated as accepted recommendation each instance in which an attendee took a picture of the recommendation displayed on the virtual board, but distinguished whether the picture was taken proactively (*Pic_Pro*) or if SARA had to suggest it beforehand (*Pic_Asked*).

4.3.2 Relevant Recommendations. We considered instances in which the attendee already knew the person recommended or already signed up for the session recommended (*Reco_Known*) as relevant recommendations. In this category, we also annotated



Greeting phase

SARA: "*I am so glad you're here I am SARA. What's your name?*"
[task intent: **greeting**] [conversational strategy: **self_disclosure**]

USER: "*My name is [anonymized].*"
[task intent: **self_naming**] [conversational strategy: **self_disclosure**]

SARA: "*I'm very excited to speak with [anonym.] participants. How are you?*"
[task intent: **pleasure_coming_together**] [conversational strategy: **self_disclosure**]

USER: "*I'm having just a spectacular time I so enjoy the ability to come here and learn about something like this that I don't know anything about.*"
[task intent: **none**] [conversational strategy: **self_disclosure**]

Recommendation phase

SARA: "*I think you'd enjoy meeting this attendee on the screen are some more details.*" ··· (1)
[task intent: **outcome_person_recommendation**]

USER: "*Great, I actually met [anonymized] for breakfast this morning ... [omitted]*" ··· (2)
[task intent: **negative_confirmation**] [conversational strategy: **self_disclosure**]

SARA: "*Let's find somebody you don't know just wait while I look this up.I think you'd enjoy meeting this person on the screen are details.*"
[task intent: **outcome_person_recommendation**]

USER: "*A very good match thank you for that SARA. My wife happens to be named Sara, so I have a natural affinity to you.*" ··· (3)
[task intent: **positive_confirmation**] [conversational strategy: **self_disclosure + praise**]

Figure 1: An excerpt of an interaction.

times when attendees explicitly said the recommendation was relevant and the domain was matching their expectations but asked for a more precise or specific recommendation in that domain (*Close_PrecDom*), or asked for a person from the same domain, but from a specific geographic region (*Close_PrecReg*).

4.3.3 Bad Recommendations. The Recommendations that attendees refused by explicitly saying there were not related to what they wanted (*Bad_DiffDom*) or because the session recommended was already over (*Bad_RecoOver*) were considered bad recommendations.

4.3.4 Rejected Recommendations. The recommendations that were refused without any explicit explanation (*Refused_NoReason*) were annotated as rejected.

5 DATA ANALYSIS

To investigate the influence of SARA's task performance on the rapport dynamics over time and thus answer to our research question, we define the following hypotheses.

H1.a - The likelihood that rapport increases during an interaction *(Rpt_Utopy)* is negatively and monotonically correlated with attendee's bad recommendations rate *(Bad_RecoOver + Bad_DiffDom)*, meaning that rapport was more likely to decrease during an interaction when SARA delivered bad recommendations.

H1.b - The likelihood that rapport increases during an interaction *(Rpt_Utopy)* is positively and monotonically correlated with good

recommendations rate *(Messenger_Yes + Pic_Pro + Pic_Asked)*, meaning that rapport was more likely to increase during an interaction if SARA delivered good recommendations.

Moreover, following previous work described in section 2, we want to investigate whether rapport had an influence on the attendee's feedback when SARA delivered recommendations.

H2.a - Average rapport score *(Rpt_Avg)* is positively and monotonically correlated with good recommendations rate, meaning that attendees acceptance rate *(Messenger_Yes + Pic_Pro + Pic_Asked)* was higher when they had a good relationship with SARA.

H2.b - Average rapport score *(Rpt_Avg)* is negatively and monotonically correlated with rejected recommendations rate, meaning that attendees rejected more recommendations without giving any explanations *(Refused_NoReason)* when they had a bad relationship with SARA.

5.1 Investigating task-performance

Before investigating the relations between rapport and task-performance, we first analyzed SARA's performance. In total, SARA delivered 203 recommendations (80 sessions and 123 persons) that we regrouped into four different categories and 9 sub-categories in total as described in section 4.3. Overall, around 46% of the recommendations were considered as good recommendations by the attendees. Only 8% of the recommendations were explicitly rejected because they did not match attendees' preferences (*Bad_DiffDom*), or because there were sessions that were already over (*Bad_RecoOver*). Around 22% of recommendations were considered relevant because the attendees already knew the person recommended or already signed up to the session recommended *(Reco_Known)*, had a more precise request related to the domain (*Close_PrecDom*) or expected someone from a similar domain, but a different region (*Close_PrecReg*). Eventually, 24% of the recommendations were rejected without any explanation (*Refused_NoReason*).

One interesting result comes from the difference between persons and sessions recommendations. Indeed, attendees accepted SARA's reminder for almost half of the sessions recommended by SARA (49%) whereas they accepted to be be introduced to less than one fourth of the persons who were recommended to them (23%). This can be explained by the different nature of the messages. The attendee was the only one receiving a reminder when a session was recommended, while someone else was involved when attendees accepted to be introduced to the person recommended. The latter can be seen as a more invasive action, forcing the attendee to send a meeting request to someone else, thus threatening the attendee's public self-image also called *face* [8]. The face-threatening nature of this action can therefore explain the lower acceptance rate for persons recommendations.

Another noticeable result concerns attendee's types. We differentiate between two types of attendees: on one hand, regular conference attendees that were personally invited to the conference to participate. These attendees clearly interacted with SARA to get interesting recommendations that matched their specific interests. On the other hand, spouses, conference staff, and journalists were not necessarily as interested in attending to particular sessions. These attendees did not have clear preferences nor expectations. The results tend to confirm these observation: regular conference attendees accepted to receive a messenger reminder for 60% of the sessions recommended, versus only 37% for spouses, staff or journalists.

Multiple Spearman's correlation tests were run to determine the relationship between attendee's interactional features and SARA's task-performance. We found a moderate, negative correlation between the amount of user's interruptions (*User_Interruptions*) and good recommendation's rate which was statistically significant (ρ = -.243, p = .045), meaning that attendees were more likely to interrupt SARA when they considered the recommendation was not relevant for them. We also found a moderate, negative correlation between the balance of word count (*Balance_WordCount*) and good recommendation's rate, meaning that when SARA spoke much more than the attendee, the latter was less likely to find that the recommendations were relevant.

5.2 Investigating rapport

The average rapport score across all the participants is 4.21 (std=0.59, min=2.92, max=5.73). For rapport utopy, the average score is -0.12 (std=0.34, min=-0.87, max=0.92). In general, we notice that the range of average rapport scores is quite narrow, as shown in Fig.2. In the worst cases, rapport was slightly below average while in the best cases, rapport was slightly above average. Moreover, we did not notice a lot of variance within each interaction, meaning that rapport did not really change that much during the interactions. This can be explained by the attendees' characteristics. As already explained, most of the attendees who interacted with SARA were world leaders, CEOs of big companies or politicians, who have to maintain a certain public image -or face- when they behave in public [17]. This face management plays an important role in rapport building [35, 37], especially during the beginning of the relationship when people are used to be very polite in order to maintain their own face as well as the other person's face. Given the length of the interactions (approximately 5 minutes), we can assume that SARA did not have enough time to break through the politeness barrier, and encourage attendees to care less about managing their own face.

Moreover, the average rapport utopy score indicates that rapport was slightly more likely to decrease during the interactions. This can be explained by the nature of the interactions. A vast majority of participants interacted with SARA during the day, meaning that their time was limited. They did not necessarily had time to build rapport with SARA, and they probably wanted to test SARA's capabilities to get an interesting recommendation. Table 2 shows the evolution of rapport based on the number of recommendations delivered by SARA. We can note that rapport drops down for attendees who stayed and received more than four recommendations.

We also ran multiple Spearman's correlation tests to determine the relationship between attendee's interactional features and the rapport scores between SARA and the attendee. We found moderate, negative correlations between SARA's response time (*SARA_RespTime*) and the average rapport score (*Rpt_Avg*) (ρ = -.335, p = .005), but also between the attendee's response time (*User_RespTime*) and the average rapport score (*Rpt_Avg*) (ρ = -.244, p = .045). This means that rapport was significantly lower when SARA or the attendee took more time to respond. As for

Table 2: Rapport scores based on the number of recommendations delivered by SARA during the interaction.

Attendees stayed	# of attendees	Rpt_Avg	Rpt_Utopy
One reco.	12	4.13	-0.15
Two reco.	16	4.30	-0.12
Three reco.	18	4.37	-0.2
Four reco.	12	4.29	-0.05
Five reco.	9	3.87	0.05
Six reco.	2	3.54	-0.25

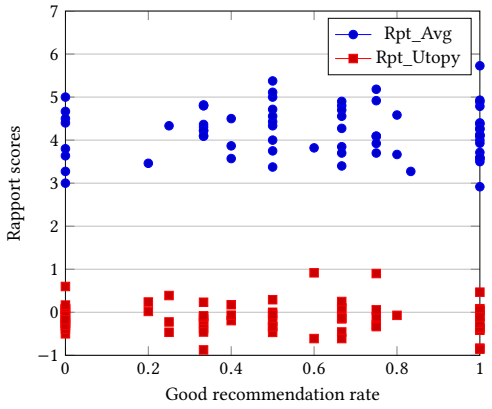


Figure 2: The circle markers represent the relation between average rapport score and good recommendations rate. The square markers represent the relation between rapport utopy and good recommendations rate.

task-performance, we found a strong, negative correlation between the balance of word count (*Balance_WordCount*) and the average rapport score (*Rpt_Avg*) (ρ = -.5, p < .001), meaning that rapport was low when SARA spoke much more than the attendee. We also expected interaction duration and/or number of turns to be positively correlated with rapport, meaning that people enjoying the interaction would stay longer. However, we did not find any significant results supporting our claim, which can also be explained by the attendees' limited time.

5.3 Investigating Rapport Vs Task

In order to investigate our hypotheses, we performed multiple Spearman correlation tests to determine the relationship between SARA's task performance and interpersonal features. More specifically, we tried to find correlations between rapport utopy and ratios of good recommendations, and between rapport utopy and bad recommendations without explanations. Unfortunately, none of these tests were significants, meaning that SARA's task performance was not influencing rapport dynamics during our field trial. One possible explanation might be that only few recommendations were considered as bad recommendations (8%). We did not find any evidence supporting our hypotheses H1.a or H1.b either, meaning that average rapport did not influence attendees' acceptance rate. This can be explained by the small standard deviation score for *Rpt_Avg*. Although we did not manage to validate our hypotheses during this field trial, we still found interesting results that will inform the design of the next generation of personal assistants. In the next section, we will discuss in more details about these results, and define more general guidelines to improve human-computer interactions.

6 DISCUSSION

Overall, our analyses revealed the role that face management plays in both task performance and rapport building. The status of the attendees and the length of the interactions did not allow SARA to break social boundaries, resulting in a lower likelihood of rapport increase, but also in a lower person recommendation acceptance rate. Our analyses also shed the lights on the influence that interactional features have on both rapport and task performance. We rely on our analyses to propose the following guidelines to specifically increase relevance and reliability of recommendation tasks, but also more generally to enhance ECAs' rapport building capabilities, both with the aim of improving user's experience.

6.1 Improving relevance and reliability of recommendation tasks

Although it would be hard for a system to reach a 100% recommendation success rate, there are still solutions that would help a personal assistant to improve the relevance of its recommendations. Eliciting more specific preferences might reduce the number of *bad* or even *relevant* recommendations. Moreover, enabling the system to explain the choice behind its recommendation may reduce the number of recommendations rejected without justification.

6.1.1 Deeper preferences elicitation to increase general task efficiency. Our results revealed that 8% of the recommendations delivered by SARA were considered as bad or irrelevant because they did not match the attendee's needs or, in cases of sessions, when there were already over. The second situation can easily be avoided by adding the actual time as an additional constraint to the recommendation database query. The first situation is due to an insufficient knowledge of the user's needs, goals or preferences, and mostly occurred when the attendee did not provide enough information during the goal_elicitation and the interest_elicitation phases. One potential solution to overcome this challenge would be for personal assistants to explicitly confirm user's interests before going to the recommendations phases. The 9% of relevant recommendations when attendees asked for a more precise or specific recommendation in one relevant domain, or asked for a person from the same domain, but from a specific region, could also be reduced in the future by the same mechanisms combined with a deeper task domain representation.

6.1.2 Explanations to increase trust. One question that is still left unanswered is why did attendees reject recommendations without giving any explanation. One potential solution to answer to this question is for personal assistants to ask for explanations whenever users refuse a recommendation without giving any information. Another solution would be to design assistants able to provide explanations to the users about their decision making process and why they suggested this particular recommendation. This explanation might encourage the user to refine his request in return, reducing system's uncertainty and risks of bad recommendations.

Moreover, as demonstrated in [28, 40], systems explaining their decisions get more trust from their users.

6.2 Enhancing ECAs' rapport building capabilities

According to [37], rapport can be influenced by three different subcomponents: mutual attentiveness, coordination, and face management. Our analyses emphasized the importance of these three subcomponents, and shown that we could still improve coordination and mutual attentiveness, respectively by adapting SARA's behavioral style to the user's one, and by reducing SARA's response time. We also propose a solution to break the politeness barrier that constitutes one of the reason why we did not notice rapport increases during our interactions.

6.2.1 Longer and repeated interactions to break politeness barrier. As explained in section 5.2, we believe that our interactions were too short to efficiently break through the politeness barrier. Although it might have been hard to keep attendees in SARA booth for a longer time, one solution to extend user's experience would have been to implement a mobile version of SARA directly within the conference online collaboration platform. This way, attendees would have been able to interact with SARA whenever they wanted, extending overall interaction time, and giving SARA the opportunity to use a wider range of conversational strategies. Indeed, conversational strategies such as referring to shared experience, or violating social norms to match interpersonal norms (e.g. teasing someone) plays an important role in rapport building [44]. However, these strategies mostly occur in the later states of the relationship, where politeness and face management are less important. More generally, designing ubiquitous ECA across multiple devices would encourage users to have repeated interactions with the agent, allowing the later to break the politeness barrier inherent to single short interactions, and build rapport in a more complex way.

6.2.2 Entrainment modeling to increase coordination. Our analyses shed light on the role that linguistic alignment might play in rapport building. Indeed, we found that imbalanced word count was negatively correlated with both rapport and task-performance, meaning that an misaligned linguistic style might lead to lower rapport and task-performance. One solution to overcome this potential issue is to design an ECA that is able to adapt its linguistic style to that of the user. This adaptation, also called entrainment, is known to increase the engagement and coordination between two participants [27], but also task efficiency [26]. To model this entrainment phenomenon, we might first need to detect user's linguistic style in real time and build a NLG module able to generate sentences accordingly. One potential solution might be to use reinforcement learning to optimize the ECA linguistic style in real time by giving positive rewards whenever rapport increase during the interaction. A similar solution has been proposed in [29].

6.2.3 Incremental architecture to increase mutual attentiveness. Through this field trial, we also highlight the importance of reducing ECA's response time, as we noticed that a long response time was possibly the cause of low rapport scores. Ward et al. [41] previously found that response delays in dialog systems frequently disrupt interactions: users might indeed become frustrated by such a latency, and could end the interaction prematurely. This particular challenge can be overcome by incrementally processing user's speech, meaning that the ECA could start planning its response before the user is even done speaking [4]. The ECA could for instance reduce delay by playing filler utterances while selecting an appropriate answer [34]. Another possible way to reduce delay, specifically during recommendation phases, would be to play a social sentence such as "I think you would love meeting this attendee", or "this session looks really interesting" while the task reasoner query the recommendation database. By the time the ECA is done saying one of these sentences, the task reasoner would have received the actual recommendation to deliver. Such an incremental solution has already been implemented and evaluated in [38].

6.2.4 Combination of task and social reasoning to increase rapport. Eventually, we could improve our interactions by combining both task and social reasoners together. Indeed, our current social reasoner only "decorates" the system task intention, changing the style of SARA's sentence rather than its content. The social state (user's conversational strategies and rapport score) does not influence the course of the interaction. However, as suggested in [43], interleaving task and social content tends to increase both user's engagement and task-performance. One first step toward this is to incorporate social factors such as rapport and social conversational strategies into SARA's task reasoner using reinforcement learning, and considering social and task reward functions [47].

7 CONCLUSION

In this paper, we reported the results of a field trial with our socially-aware robot assistant. This assistant was designed to help attendees of a conference to find interesting people to meet or sessions to attend that matched their goals. Besides her task-oriented goal, SARA was also designed to fulfill a social goal, namely building rapport with the attendee with whom she was interacting. Our initial assumption was that the rapport dynamics between SARA and its attendee would be partially driven by the former task performance. Our preliminary analyses of the interactions we recorded during the conference did not allow us to find evidence that supported our claim. Given the nature of this field trial, we were not able to evaluate our social pipeline (conversational strategy classifier, rapport estimator and social reasoner). Therefore, our next step is to run a proper evaluation of our system by comparing, for instance, our socially-aware robot assistant with a task-oriented only personal assistant. Overall, our analyses emphasized the influence of coordination, mutual attentiveness, and face management on rapport building and task performance. In this paper, we thus proposed specific solutions focusing on these three sub-components, as we believe this will increase the social awareness of the future generation of ECAs.

ACKNOWLEDGMENTS

This work was supported in part by generous funding from Microsoft, LivePerson, Google, and the IT R&D program of MSIP/IITP [2017-0-00255, Autonomous Digital Companion Development]. We also would like to thank Lauren Simmons, Fadi Botros, and Ran Zhao for their significant contributions to the system operation in the field trial.

REFERENCES

[1] Nalini Ambady and Robert Rosenthal. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. (1992).

[2] Amos Azaria and Jason Hong. 2016. Recommender Systems with Personality.. In *RecSys*. 207–210.

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. OpenFace: an open source facial behavior analysis toolkit. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*.

[4] Timo Baumann, Casey Kennington, Julian Hough, and David Schlangen. 2017. Recognising conversational speech: What an incremental asr should do for a dialogue system and how to get there. In *Dialogues with Social Robots*. Springer, 421–432.

[5] Pauline M Berry, Melinda Gervasio, Bart Peintner, and Neil Yorke-Smith. 2011. PTIME: Personalized assistance for calendaring. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2, 4 (2011), 40.

[6] Timothy W Bickmore, Laura M Pfeifer Vardoulakis, and Daniel Schulman. 2013. Tinker: a relational agent museum guide. *Autonomous agents and multi-agent systems* 27, 2 (2013), 254–276.

[7] Woody Bledsoe. 1986. I had a dream: AAAI presidential address. *AI Magazine* 7, 1 (1986), 57.

[8] Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*. Vol. 4. Cambridge university press.

[9] Angelo Cafaro, Nadine Glas, and Catherine Pelachaud. 2016. The effects of interrupting behavior on interpersonal attitude and engagement in dyadic interactions. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 911–920.

[10] Justine Cassell. 2000. *Embodied conversational agents*. The MIT Press.

[11] Justine Cassell and Timothy Bickmore. 2003. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction* 13, 1-2 (2003), 89–132.

[12] Justine Cassell, Alastair J Gill, and Paul A Tepper. 2007. Coordination in conversation and rapport. In *Proceedings of the workshop on Embodied Language Processing*. Association for Computational Linguistics, 41–50.

[13] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2004. BEAT: the behavior expression animation toolkit. In *Life-Like Characters*. Springer, 163–185.

[14] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1061–1068.

[15] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 835–838.

[16] Michael Freed, Jaime G Carbonell, Geoffrey J Gordon, Jordan Hayes, Brad A Myers, Daniel P Siewiorek, Stephen F Smith, Aaron Steinfeld, and Anthony Tomasic. 2008. RADAR: A Personal Assistant that Learns to Reduce Email Overload.. In *AAAI*. 1287–1293.

[17] Erving Goffman. 2006. The presentation of self. *Life as theater: A dramaturgical sourcebook* (2006).

[18] Mathieu Guillame-Bert and James L. Crowley. 2012. Learning Temporal Association Rules on Symbolic Time Sequences. (2012), 159–174.

[19] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2011. Virtual Rapport 2.0. In *International Workshop on Intelligent Virtual Agents*. Springer, 68–79.

[20] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *International workshop on intelligent virtual agents*. Springer, 205–217.

[21] Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. 2013. I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM Transactions on Management Information Systems (TMIS)* 4, 4 (2013), 19.

[22] Michael Madaio, Justine Cassell, and Amy Ogan. 2017. The impact of peer tutors' use of indirect feedback and instructions. Philadelphia, PA: International Society of the Learning Sciences.

[23] Pattie Maes. 1989. How to do the right thing. *Connection Science* 1, 3 (1989), 291–323.

[24] Yoichi Matsuyama, Arjun Bhardwaj, Ran Zhao, Oscar J Romero, Sushma Anand Akoju, and Justine Cassell. 2016. Socially-aware animated intelligent personal assistant agent. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 224.

[25] Karen Myers, Pauline Berry, Jim Blythe, Ken Conley, Melinda Gervasio, Deborah L McGuinness, David Morley, Avi Pfeffer, Martha Pollack, and Milind Tambe. 2007. An intelligent personal assistant for task and time management. *AI Magazine* 28, 2 (2007), 47.

[26] Ani Nenkova, Agustin Gravano, and Julia Hirschberg. 2008. High frequency word entrainment in spoken dialogue. In *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers*. Association for Computational Linguistics, 169–172.

[27] Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology* 21, 4 (2002), 337–360.

[28] Florian Nothdurft and Wolfgang Minker. 2016. Justification and transparency explanations in dialogue systems to maintain human-computer trust. In *Situated Dialog in Speech-Based Human-Computer Interaction*. Springer, 41–50.

[29] Hannes Ritschel, Tobias Baur, and Elisabeth André. 2017. Adapting a Robot's Linguistic Style Based on Socially-Aware Reinforcement Learning. (2017).

[30] Oscar J. Romero. 2011. An evolutionary behavioral model for decision making. *Adaptive Behavior* 19, 6 (2011), 451–475.

[31] Oscar J Romero, Ran Zhao, and Justine Cassell. 2017. Cognitive-Inspired Conversational-Strategy Reasoner for Socially-Aware Agents. *IJCAI* (2017), 3807–3813.

[32] Tanmay Sinha. [n. d.]. Cognitive Correlates of Rapport Dynamics in Longitudinal Peer Tutoring. ([n. d.]).

[33] Tanmay Sinha and Justine Cassell. 2015. We click, we align, we learn: Impact of influence and convergence processes on student learning and rapport building. In *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence*. ACM, 13–20.

[34] Gabriel Skantze and Anna Hjalmarsson. 2013. Towards incremental speech generation in conversational systems. *Computer Speech & Language* 27, 1 (2013), 243–262.

[35] Helen Spencer-Oatey. 2005. (Im) politeness, face and perceptions of rapport: unpackaging their bases and interrelationships. (2005).

[36] Marcus Thiebaux, Stacy Marsella, Andrew N Marshall, and Marcelo Kallmann. 2008. Smartbody: Behavior realization for embodied conversational agents. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 151–158.

[37] Linda Tickle-Degnen and Robert Rosenthal. 1990. The nature of rapport and its nonverbal correlates. *Psychological inquiry* 1, 4 (1990), 285–293.

[38] Vivian Tsai, Timo Baumann, Florian Pecune, and Justine Casell. 2018. Faster Responses are Better Responses: Introducing Incrementality into Sociable Virtual Personal Assistants. In *Proceedings of the 2018 International Workshop on Spoken Dialog System Technology*.

[39] Gokhan Tur, Andreas Stolcke, Lynn Voss, Stanley Peters, Dilek Hakkani-Tur, John Dowding, Benoit Favre, Raquel Fernández, Matthew Frampton, Mike Frandsen, et al. 2010. The CALO meeting assistant system. *IEEE Transactions on Audio, Speech, and Language Processing* 18, 6 (2010), 1601–1611.

[40] 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 *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 997–1005.

[41] Nigel G Ward, Anais G Rivera, Karen Ward, and David G Novick. 2005. Root causes of lost time and user stress in a simple dialog system. (2005).

[42] Yorick Wilks. 2010. *Close engagements with artificial companions: key social, psychological, ethical and design issues*. Vol. 8. John Benjamins Publishing.

[43] Zhou Yu, Alexander I. Rudnicky, and Alan W. Black. 2017. Learning Conversational Systems that Interleave Task and Non-Task Content. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*. 4214–4220.

[44] Ran Zhao, Alexandros Papangelis, and Justine Cassell. 2014. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *International Conference on Intelligent Virtual Agents*. Springer, 514–527.

[45] Ran Zhao, Tanmay Sinha, Alan W Black, and Justine Cassell. 2016. Automatic recognition of conversational strategies in the service of a socially-aware dialog system. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 381.

[46] Ran Zhao, Tanmay Sinha, Alan W Black, and Justine Cassell. 2016. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *International Conference on Intelligent Virtual Agents*. Springer, 218–233.

[47] Zian Zhao, Michael Madaio, Florian Pecune, Yoichi Matsuyama, and Justine Casell. 2018. Socially-Conditioned Task Reasoning for a Virtual Tutoring Agent. In *Proceedings of the 2018 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems.